CONTEXT-BASED METRICS FOR EVALUATING CHANGES TO WEB PAGES

A Thesis

by

SUVENDU KUMAR DASH

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2003

Major Subject: Computer Science

CONTEXT-BASED METRICS FOR EVALUATING CHANGES TO WEB PAGES

A Thesis

by

SUVENDU KUMAR DASH

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Approved as to style and content by:

Frank Shipman (Chair of Committee)	Marietta Tretter (Member)
Richard Furuta (Member)	Valerie E. Taylor (Head of Department)

December 2003

Major Subject: Computer Science

ABSTRACT

Context-Based Metrics for Evaluating Changes to Web Pages. (December 2003)
Suvendu Kumar Dash, B.E., Sambalpur University
Chair of Advisory Committee: Dr. Frank Shipman

The web provides a lot of fluid information but this information changes, moves, and even disappears over time. Bookmark lists, portals, and paths are collections where the building blocks are web pages, which are susceptible to these changes. A lot of research, both in industry and in academia, focuses on organizing this vast amount of data. In this thesis, I present context-based algorithms for measuring changes to a document. The methods proposed use other documents in a collection as the context for evaluating changes in the web pages. These metrics will be used in maintaining paths as the individual pages in paths change. This approach will enhance the evaluations of change made by the currently existing Path Manager, in the Walden's Paths project that is being developed in the Center for the Study of Digital Libraries at Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

TABLE Page

TABLE Page

1. INTRODUCTION

The web provides a lot of fluid information but this information changes, moves, and even disappears over time [Brewington, Cybenko 2000]. Bookmark lists, portals, and paths are collections where the building blocks are web pages, which are susceptible to these changes. A lot of research in both industry and in academia is going into organizing this vast amount of data [Ashman 2000; Chakrabarti et al. 1999]. While all forms of document change over time, the materials present on the web are not as fixed - that is they change more frequently, than documents written on paper [Levy, Marshall 1994]. If we are maintaining a digital library of web-based documents, we need to constantly 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 directory, are examples of page collections that can become out-of-date as changes are made to their components [Brewington, Cybenko 2000]. 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 to identify the degree of changes to these web pages and decide 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.

Research to automatically maintain these web page collections has resulted in a variety of algorithms and tools that compute content-based metrics of change to a page, but they do not take into account the use of the page (its context) [Salton 1989; Revilla et al. 2001]. 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 a geographic region.

This thesis follows the style and format of *ACM Transactions on Information Systems*.

Now consider a page about a visiting Monet exhibit at a museum. 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. In this thesis, I present context-based algorithms for measuring changes to a document. The methods that we are currently proposing use the other documents in a collection as the context for evaluating changes in the web pages. These metrics will be used in maintaining paths [Walden's Paths] as the individual pages in paths change. This approach will enhance the evaluations of change made by the currently existing Path Manager [Revilla et al. 2001], in the Walden's Paths project that is being developed in the Center for the Study of Digital Libraries at Texas A&M University.

2. RELATED WORK

A variety of research is related to the topic of evaluating changes to web pages based on the context of their use. This work includes systems supporting the maintenance of a collection of web pages, algorithms for detecting similar web pages, profiling the user's context of browser usage, etc. In addition, since the calculation of context in this project is represented as a term vector space model, research using term vector spaces to represent collections of web pages is also related.

2.1 DETECTING CONTENT-BASED CHANGES IN WEB PAGES

Various efforts have been made in the direction of identifying content-based changes to web pages. WebBase [Hirai et al. 1999] is a project being developed to store and maintain a large shared repository of web pages at the Stanford Digital Libraries Group. It is investigating various issues in crawling, storing, indexing, and querying of large collections of Web pages. Since the collection consists of web pages that change over time, one of the challenges that they are working on is continuously updating and maintaining a large collection of web pages. Tools have been developed to detect simple content-based changes in the web pages. Whenever there is a change in the content of the web page these tools automatically detect these changes. These are similar to the tools currently widely used in version management systems. HtmlDiff [Douglis, Ball 1996] written by Tom Ball at AT&T Bell Labs calculates the difference between two HTML pages. This tool was later used in AT&T Internet Difference Engine AIDE [Douglis et al. 1998] to develop a tool for tracking and viewing modifications to World-Wide-Web pages, which has been extended to support recursive tracking of pages. Ciao [Chen et al. 1995] is a graphical navigator that allows users to query and browse structural connections embedded in a document repository. Over time, both these two tools were combined to develop a tool called the WebGUIDE [Douglis 2002].

2.2 CONTEXT-BASED CRAWLING AND DETECTION OF SIMILAR PAGES

Some systems have been developed that take collections of web pages as input and produce similar web pages in the World-Wide-Web as output. They use focused crawlers to implement this functionality. The NEURODOC project [Lelu, Francois 1992] was a "novel prototype of neural browser for information retrieval. It relies on two unsupervised neural models for extracting relevant fuzzy clusters from the data, in other words for identifying data poles and their related environment" [Lelu, Francois 1992]. The Context Focused Crawler (CFC) [Diligenti et al. 2000] presents a focused crawling algorithm that tries to calculate the context of the page and uses this context to locate similar pages on the web. Focused crawlers usually cover a subset of the web depending on the specific topic to be crawled. The system tries to fit pages into previously defined contexts by calculating the context of each test page within a collection. They build a context graph and then try to fit the test document into somewhere in the graph.

In addition, context of a web document can be drawn from the history of the pages that the user has browsed. Automatic Link Generator [Wilkinson, Smeaton 1999] calculates a user's local context from the history of pages that the user was browsing. The history of the browser usage acts as a local context for the particular user. The system generates "links that relate to a structure associated with the information space such as an overall hierarchy, or links that refer to the semantic closeness of pieces of information, or reference links that point to related though not semantically close pieces of information" [Wilkinson, Smeaton 1999]. In addition, the system suggests the user similar pages that the user can use to create links at the time of use, which need more sophisticated user models "that take into account not only location, explicit needs, but user history, and long standing user preferences"[Wilkinson, Smeaton 1999].

2.3 GETTING CONTEXT IN THE VECTOR SPACE MODEL

A lot of research in the information retrieval field attempts to model text documents as vectors and gather mathematical data about those documents. This kind of representation is known as the Vector Space Model by the research community. This method is easy to

implement but has the drawback of losing the local information presented by the sentences. This model can be refined by using sentences as an intermediate model that takes the context of the words into account instead of getting the words into the model [Salton 1989; Pullwitt, Der 2001], so that the loss of local information would be reduced. Vector Space Models are used a lot in the field of text classification. Many tools have been developed in the text classification field of research to explore large, heterogeneous collections of web content. Text classification attempts to determine the category/topic of the page, often within a hierarchy of topics. An example is Dumais' Hierarchical classifier [Dumais, Chen 2000], which uses Support vector machine (SVM) [Dumais et al. 1998] classifiers, to help users create structured knowledge hierarchies. The tool was used to support classification of search results. Currently the Context-Based module in the Path Manager is using words in the web document to calculate the context of the document. In the future work it might include a set of words in the term-vector instead of single words.

3. WALDEN'S PATHS

3.1 WALDEN'S PATHS

Walden's Paths [Walden's Paths] is an application that can be used by the K-12 educators to organize World-Wide Web material for their students use. Figure 1 shows how the user browses through the paths. It allows the creation of trails [Bush 1945], paths using pages on the web that have been created by others [Furuta et al. 1997]. 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.

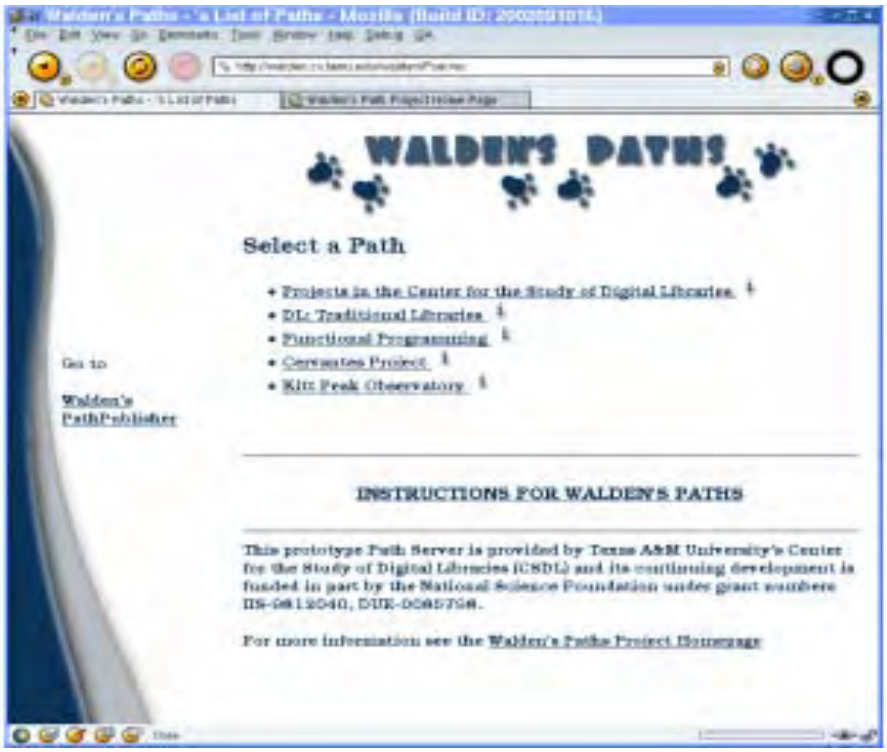

Figure 1. Walden's Paths

As the paths are built upon web pages, Walden's Paths 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 [Revilla et al. 2001], a tool

that is meant to focus the collection maintainer's attention on the pages that have changed the most.

3.2 PATH MANAGER

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, the Johnson's algorithm and the Proportional algorithm, use signatures that abstract characteristics of a page including headers, links and text.

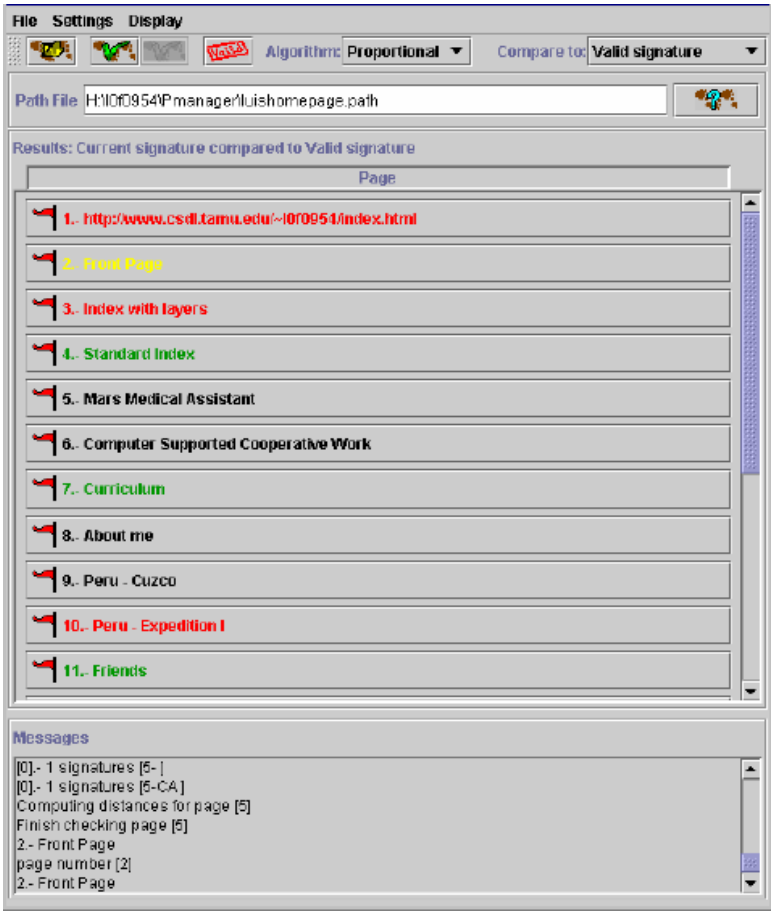

Figure 2. Path Manager

4. IMPLEMENTATION

Up to now, the Path Manager has focused on content-based methods for evaluating changes to 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. My research investigates the use of context-based metrics of change 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.

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 punctuations, stop words including 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 page. The term weight is calculated using many heuristics. The weight of a term is given as log (term frequency + 5). The terms that are nouns were detected from a list of nouns and were given more weight, by multiplying a weight factor of 10 to the weight as calculated above.
- Save the Page Term Vectors and Weight Vectors for all pages in the signature File of 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 Results section, initial values for representing changes to the user are:

- A two-degree or greater move by the page vector towards the path vector should be indicated as a move towards the path
- A move between positive two and negative two degrees should be indicated as being similar to the prior situation, and
- A two-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.

These values are only initial recommendations and need to be evaluated in practice. As the results in the Results section indicate, this approach at noticing changes in context produces a range of angles and no single cutoff would always generate the desired results.

5. EVALUATION

To evaluate the described algorithm's performance in determining whether changes result in a move out of the context of a collection, I have compared the algorithm with human catalogers. 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 from 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 page selected 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 page was about elephants and the second page was the CNN Financials page. 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 be away from the path vector.

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

6. RESULTS

This section presents the results for each of the 20 paths. Each path is presented on a separate page with the topic of the path at the top of the page followed by a table showing the results of three measures of change. The first measure is the angle between the original page's term vector and that of its replacement. The second measure is the "high", "medium", or "low" rating of the change by the Walden's Paths Path Manager using the Proportional Algorithm. The third measure is the 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.

In practice, 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.

Before presenting the results from the individual paths, I present a summary of these results showing the averages, ranges and standard deviations for the results from the 20 paths.

Table 1 shows the averages, ranges and the standard deviations of the values in the following 20 tables.

Table 1. Averages of the Twenty Paths

		The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a Similar Page
Angle of the changed page to the original page (in degrees)	Average	78.0996	81.85915	75.13672
	Ranges	30.77 to 88.15	77.025 to 87.726	35.091 to 84.5424
	Standard Deviation	15.65433	2.890859	10.75518
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)		High Level	High Level	High Level
Angle of the page to the path (in degrees)	Average	-7.81489	-9.07515	1.94304
	Ranges	-23.19 to 1.639	-45.03 to 0.876	-15.18 to 14.3
	Standard Deviation	6.946088	10.57454	6.801181

From the values in the last row of the table above, 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. Here in the experiment we have chosen the paths so that the path doesn't contain any similarity with the Elephants and the CNN Financials page. These results indicate that, on average, the algorithm does differentiate between replacements by pages with similar content and replacements with unrelated pages. It should also be noted that the top row, showing the angle between the original and replacement page, does not provide much difference between these cases.

The Results for the context-based metrics is shown here in the tables:

6.1 PATH ABOUT SEARCH ENGINES

Table 2. Path about Search Engines

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page talking about Internet (Similar Page)
Angle of the changed page to the original page (in degrees)	85.746	78.537	73.729
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-18.737	-7.425	+0.932

The results in Table 2 show that, the Elephants page moved 18.7 degrees away from the path whereas the CNN Financials page moved 7.4 degrees away from the path. The Internet page (Similar Page) moved 0.9 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.2 PATH ABOUT MUSIC

Table 3. Path about Music

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page talking about Sheet Music (Similar Page)
Angle of the changed page to the original page (in degrees)	86.76	77.209	71.727
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-16.868	-11.00	-3.868

The results in Table 3 show that, the Elephants page moved 16.8 degrees away from the path whereas the CNN Financials page moved 11.0 degrees away from the path. The Sheet Music page (Similar Page) moved 3.8 degrees away from the path. This represents that the similar page moved away from the path but moved less away from the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.3 PATH ABOUT MOVIES

Table 4. Path about Movies

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page of NY Times Movies Page (Similar Page)
Angle of the changed page to the original page (in degrees)	87.025	83.065	80.023
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-15.622	-5.6759	2.8108

The results in Table 4 show that, the Elephants page moved 15.6 degrees away from the path whereas the CNN Financials page moved 5.6 degrees away from the path. The NY Times movie page (Similar Page) moved 2.8 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.4 PATH ABOUT INDIAN HISTORY

Table 5. Path about Indian History

	The whole page changed to a page talking about Giraffes	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Indian History (Similar Page)
Angle of the changed page to the original page (in degrees)	87.916	30.77	87.726	84.5424
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-18.821	1.639	-18.371	-3.012

The results in Table 5 show that, the Elephants page moved 1.6 degrees towards the path whereas the CNN Financials page moved 18.3 degrees away from the path. The Indian History page (Similar Page) moved 3.0 degrees away from the path. This represents that the similar page moved away from the path but moved less away from the path as compared to the elephant and the CNN Financials page. The reason behind the elephants page moving closer to the path is that many of the Indian History pages have discussions of elephants used in the wars.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.5 PATH ABOUT ZOOS

Table 6. Path about Zoos

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about a National Park in Thailand (Similar Page)
Angle of the changed page to the original page (in degrees)	35.859	84.531	35.091
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	1.296	-45.030	1.930

The results in Table 6 show that, the Elephants page moved 1.2 degrees towards the path whereas the CNN Financials page moved 45.0 degrees away from the path. The Internet page (Similar Page) moved 1.9 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page. The reason the Elephants page moves towards the path is that the elephants page includes discussions of elephants similar to those found in zoo pages.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.6 PATH ABOUT NETWORKING BASICS

Table 7. Path about Networking Basics

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Basic Concepts in Networking (Similar Page)
Angle of the changed page to the original page (in degrees)	87.313	80.138	82.763
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-8.501	0.876	4.011

The results in Table 7 show that, the Elephants page moved 8.5 degrees away from the path whereas the CNN Financials page moved 0.8 degrees towards the path. The Internet page (Similar Page) moved 4.0 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant page. The emphasis on technology companies in the stock market might cause the CNN Financials page's fit in this context.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.7 PATH ABOUT EMAIL

Table 8. Path about Email

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about How to stop Spam (Similar Page)
Angle of the changed page to the original page (in degrees)	85.833	77.025	77.999
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-11.806	-0.831	-0.667

The results in Table 8 show that, the Elephants page moved 11.8 degrees away from the path whereas the CNN Financials page moved 0.8 degrees away from the path. Again, the technology focus of the stock market may cause this similarity. The Internet page (Similar Page) moved 0.6 degrees away from the path. This represents that the similar page moved away from the path but moved less away from the path as compared to the elephant page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.8 PATH ABOUT SOFTWARE HISTORY

Table 9. Path about Software History

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Software History (Similar Page)
Angle of the changed page to the original page (in degrees)	88.150	86.369	81.667
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-8.325	-2.415	8.626

The results in Table 9 show that, the Elephants page moved 8.3 degrees away from the path whereas the CNN Financials page moved 2.4 degrees away from the path. The Internet page (Similar Page) moved 8.6 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.9 PATH ABOUT ALLERGIES

Table 10. Path about Allergies

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Allergies (Similar Page)
Angle of the changed page to the original page (in degrees)	79.99	78.73	76.86
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-1.71	-4.69	14.18

The results in Table 10 show that, the Elephants page moved 1.71 degrees away from the path whereas the CNN Financials page moved 4.69 degrees away from the path. The Allergies page (Similar Page) moved 14.18 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.10 PATH ABOUT TEXAS HISTORY

Table 11. Path about Texas History

	The whole page changed to a page about Mexican History	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Texas History (Similar Page)
Angle of the changed page to the original page (in degrees)	82.52	85.51	80.92	77.374
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level	High Level
Angle of the page to the path (in degrees)	1.38	-6.66	-3.19	5.35

The results in Table 11 show that, the Elephants page moved 6.66 degrees away from the path whereas the CNN Financials page moved 3.19 degrees away from the path. The Texas History page (Similar Page) moved 5.35 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page. Also, from above we can see that the Mexican History page has moved 1.38 degrees towards the path. The reason behind this is that Texas History and Mexican History overlap and the terms used in those documents are similar. This is one of the cases where the algorithm will give false positives.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.11 PATH ABOUT CANCER

Table 12. Path about Cancer

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Cancer (Similar Page)
Angle of the changed page to the original page (in degrees)	83.68	85.71	80.47
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-5.86	-8.47	-3.83

The results in Table 12 show that, the Elephants page moved 5.86 degrees away from the path whereas the CNN Financials page moved 8.47 degrees away from the path. The Cancer page (Similar Page) moved 3.83 degrees away from the path. This represents that the similar page moved away from the path but moved less away from the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.12 PATH ABOUT CLIMATE CHANGE

Table 13. Path about Climate Change

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Climate Change (Similar Page)
Angle of the changed page to the original page (in degrees)	80.58	82.13	82.78
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-3.069	-3.34	3.52

The results in Table 13 show that, the Elephants page moved 3.06 degrees away from the path whereas the CNN Financials page moved 3.34 degrees away from the path. The Climate Change page (Similar Page) moved 3.52 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.13 PATH ABOUT OZONE DEPLETION

Table 14. Path about Ozone Depletion

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Ozone Depletion (Similar Page)
Angle of the changed page to the original page (in degrees)	82.56	81.56	70.67
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-23.19	-25.53	-15.18

The results in Table 14 show that, the Elephants page moved 23.19 degrees away from the path whereas the CNN Financials page moved 25.53 degrees away from the path. The Ozone Depletion page (Similar Page) moved 15.18 degrees away from the path. This represents that the similar page moved away from the path less than the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.14 PATH ABOUT WEDDING

Table 15. Path about Wedding

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Wedding (Similar Page)
Angle of the changed page to the original page (in degrees)	78.59	79.17	76.11
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-12.79	-10.84	-4.13

The results in Table 15 show that, the Elephants page moved 12.79 degrees away from the path whereas the CNN Financials page moved 10.84 degrees away from the path. The Wedding page (Similar Page) moved 4.13 degrees towards the path. This represents that the similar page moved away from the path less than the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.15 PATH ABOUT SOCIOLOGY

Table 16. Path about Sociology

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Sociology (Similar Page)
Angle of the changed page to the original page (in degrees)	79.95	82.50	63.19
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-2.549	-5.81	3.42

The results in Table 16 show that, the Elephants page moved 2.54 degrees away from the path whereas the CNN Financials page moved 5.81 degrees away from the path. The Sociology page (Similar Page) moved 3.42 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.16 PATH ABOUT SOFTWARE ENGINEERING

Table 17. Path about Software Engineering

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Software Engineering (Similar Page)
Angle of the changed page to the original page (in degrees)	82.76	82.32	74.34
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-8.72	-8.34	7.72

The results in Table 17 show that, the Elephants page moved 8.72 degrees away from the path whereas the CNN Financials page moved 8.34 degrees away from the path. The Software Engineering page (Similar Page) moved 7.72 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.17 PATH ABOUT DISTANCE LEARNING

Table 18. Path about Distance Learning

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Distance Learning (Similar Page)
Angle of the changed page to the original page (in degrees)	79.12	82.54	73.64
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-8.97	-11.77	-4.18

The results in Table 18 show that, the Elephants page moved 8.97 degrees away from the path whereas the CNN Financials page moved 11.77 degrees away from the path. The Internet page (Similar Page) moved 4.18 degrees away from the path. This represents that the similar page moved away from the path less than the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.18 PATH ABOUT SCIENTISTS

Table 19. Path about Scientists

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Scientists (Similar Page)
Angle of the changed page to the original page (in degrees)	82.96	83.32	75.52
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-2.36	-7.91	14.30

The results in Table 19 show that, the Elephants page moved 2.36 degrees away from the path whereas the CNN Financials page moved 7.91 degrees away from the path. The Internet page (Similar Page) moved 14.3 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.19 PATH ABOUT NANOTECHNOLOGY

Table 20. Path about Nanotechnology

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about nanotechnology (Similar Page)
Angle of the changed page to the original page (in degrees)	80.78	81.44	82.43
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	0.00117	-1.44	1.82

The results in Table 20 show that, the Elephants page moved 0 degrees towards from the path whereas the CNN Financials page moved 1.44 degrees away from the path. The nanotechnology page (Similar Page) moved 1.82 degrees towards the path. This represents that the similar page moved towards the path as compared to the CNN Financials page. The Elephants page stays as a similar page because the other pages in the path are small and so some of the terms are similar to the terms present in the path.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

6.20 PATH ABOUT EDUCATION

Table 21. Path about Education

	The whole page changed to a page talking about Elephants	The whole page changed to a page on CNN Financials	The whole page changed to a page about Education (Similar Page)
Angle of the changed page to the original page (in degrees)	78.056	82.243	81.809
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	-3.497	-0.301	5.108

The results in Table 21 show that, the Elephants page moved 3.49 degrees away from the path whereas the CNN Financials page moved 0.30 degrees away from the path. The US News page (Similar Page) moved 5.10 degrees towards the path. This represents that the similar page moved towards the path as compared to the elephant and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

7. FUTURE WORK

The results from the evaluation were in accordance to the changes that a human would perceive as per the context of the document/web page that have been written. Future work for this research is to weight terms based on their use in the document. The module that calculates the context based metrics uses a parser that was written earlier in the Path Manager. The parser scans the web page and gets the terms in the document, the hyperlinks etc. However, the parser doesn't differentiate between the terms that have been given more importance by the author like the terms that are in headings, written in bold, or having a larger font size. The context-based module was therefore not able to consider this information in calculating the context-based metrics.

To check whether this information is helpful in finding out the context-based metrics, one collection was tested in the context-based module separately. The results are shown in the following table. For this collection, the headings of the pages were given more weight and then the context-based metrics were calculated. The table on the next page shows the angles of change using both the un-weighted and weighted terms. The results show that the change metrics went up to almost 55 degrees for the non-similar pages. These larger angles are an improvement since they provide greater range in which to select cutoff values for potentially problematic page replacements. If a parser can identify headings consistently, then the context-based module can use this information to potentially improve the results. Issues remaining include how to weight headings to balance false positives with improved range. This is an empirical question, that requires comparing a variety of pages and collections to determine.

Table 22. Path about Elephants (with Headings given more weight)

	Page with one paragraph changed	The whole page changed to a page talking about Giraffe	The whole page changed to a page on CNN Financials	The whole page changed to a page talking about Elephants (Similar Page)
Angle of the changed page to the original page (in degrees)	20.59	80.668	81.041	75.5819
Angle of the changed page to the original page (in degrees) with more weights given to headings	8.366	86.37	86.5	24.32
Proportional Algorithm Degrees of change (High Level, Medium and Lowest)	Medium Level (Green)	High Level	High Level	High Level
Angle of the page to the path (in degrees)	0.913	-4.658	-7.202	-1.837
Angle of the page to the path (in degrees) with more weights given to headings	-0.254	-52.707	-54.179	+0.756

The results in Table 22 show that, the Giraffe page moved 52.7 degrees away from the path whereas the CNN Financials page moved 54.1 degrees away from the path. The Internet page (Similar Page) moved 0.75 degrees towards the path. This represents that the similar page moved towards the path as compared to the Giraffe and the CNN Financials page.

Note:

-ve degrees represent that the page has moved away from the path

+ve degrees represent that the page has moved towards the path

8. CONCLUSION

The results for the context-based metrics of change were in accordance to the human perspective of changes to the web pages. These metrics can be used in addition to the content-based metrics of change to help the user/maintainer of Walden's paths to maintain collections of web pages. The content-based change metrics aid users in evaluating changes to a web page but not in evaluating page replacements. The context-based metric will help users determine replaced pages are topically appropriate for the path.

Future enhancements to this work include the weighting of terms based on their use in the documents. This metric of appropriateness of documents for a path/collection might also be of value in selecting replacements for pages no longer available.

REFERENCES

Ashman, H. (2000). “Electronic Document Addressing – Dealing with Change,” *ACM Computing Surveys 32*, pp. 201-212.

Brewington, B. & Cybenko, G. (2000). “How Dynamic is the Web,” *Proceedings of WWW9 –9th International World Wide Web Conference (IW3C2),* pp. 264-296.

Bush, V. (1945). “As We May Think,” *The Atlantic Monthly, (August 1945)*, pp. 101-108.

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.

Chen, Y., Glenn, S. F., Koutsofios, E., & Wallach, R. S. (1995). “Ciao: A Graphical Navigator for Software and Document Repositories,” *International Conference on Software Maintenance*, pp. 66-75.

Diligenti, M., Coetzee, F., Lawrence, S., Giles, C., & Gori, M. (2000). “Focused Crawling using Context Graphs,” *26th International Conference on Very Large Databases, VLDB 2000*, pp. 527-534.

Douglis, F. (2002). “WebGUIDE: Querying and Navigating Changes in Web Repositories,” Available on-line: http://www.research.att.com/sw/tools/aide/www5 (accessed on September 26, 2003).

Douglis, F. & Ball, T. (1996). “Tracking and Viewing Changes on the Web,” *Proceedings of 1996 USENIX Technical Conference*, pp. 165-176.

Douglis, F., Ball, T., Chen, Y., & Koutsofios, E. (1998). "The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web," *Proceedings of World Wide Web*, pp. 27-44. (Also appears as *AT&T Labs--Research TR 97.23.1*).

Dumais, S., Platt, J., Heckerman, D. & Sahami, M. (1998). "Inductive Learning Algorithms and Representations for Text Categorization," *Proceedings of ACM-CIKM98*, pp. 148-155.

Dumais, S. & Chen, H. (2000). "Hierarchical Classification of Web Content," *Proceedings of SIGIR*, pp. 256-263.

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.

Hirai, J., Raghavan, S., Garcia-Molina, H., & Paepcke, A. (1999). *WebBase: A Repository of Pages*, Stanford Digital Libraries Project Technical Report SIDL-WP-1999-0124, Computer Science Dept, Stanford University, CA.

Levy, D. M. & Marshall, C. (1994). "Washington's White Horse? A Look at Assumptions Underlying Digital Libraries," *Proceedings of Digital Libraries 1994*, pp. 163-169.

Lelu, A. & Francois, C. (1992). "Hypertext Paradigm in the Field of Information Retrieval: A Neural Approach," *Proceedings of the Fourth ACM Conference on Hypertext, Information Retrieval*, pp. 112-121.

Pullwitt, D., & Der, R. (2001). "Integrating Contextual Information into Text Document Clustering with Self-Organizing Maps," *Proceedings of WSOM'01, Workshop on Self-Organizing Maps,* pp. 54-60.

Revilla, L. F., 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.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA 1989.

Stata, R., Bharat, K., & Maghout, F. (2000). “The Term Vector Database: Fast Access to Indexing Terms for Web Pages,” *Proceedings of WWW9*, pp. 247-256.

“Walden’s Paths”, http://www.csdl.tamu.edu/walden/ (accessed on September 26, 2003).

Wilkinson, R. & Smeaton, A. F. (1999). “Automatic Link Generation,” *ACM Computing Surveys*, p. 27.

VITA

Name:	Suvendu Kumar Dash
Permanent Address:	A/L 21 V.S.S Nagar Bhubaneswar-751004. Orissa, India.
Education:	M.S. Computer Science, Texas A&M University, 2003 B.E. Electrical Engineering, University College of Engineering, Burla, Sambalpur University, Orissa, India, 1998