ABSTRACT

In this paper, I present an analysis of the continual and sometimes sudden changes of web site home page content. I study the evolutions of 2,309 home pages from 27 web sites from 1999 to 2007. These 27 sites are broken into 4 categories (school, news, online shopping, and software) in order to determine whether change occurs more frequently (both in content and structure) in one category over another. The analysis shows that not only do home pages change over time (generally in an increasing use of content and tags), but they draw more content, functionality, and formatting from external sources. This leads to the conclusion that in order to fully understand web site evolution, more work needs to be focused on resources connected to, but not fully present in source data. Furthermore, with the advancements of web programming tools (e.g. Dreamweaver) it has become increasingly easy to add, edit, and remove elements (both simple and complex) from a page; whether by an expert or novice. Therefore, it is my contention, supported by the finding in this work, that these trends will continue to grow along with complexity, thus bringing about a shift in the way webpages should be processed.

Categories and Subject Descriptors
H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia; H.3.m [Information Storage and Retrieval]: Miscellaneous

General Terms
Measurement, Experimentation

Keywords
Webpage evolution, data extraction, content analysis, HTML tag analysis, cosine similarity

1. INTRODUCTION

The web is a constantly changing and expanding giant. Professional, commercial, and individual sites go up, come down, and change continuously. Understanding how webpages evolve can help researchers build better models and algorithms for predicting and handling these events and web search providers/aggregators in determining when a page (or cluster of similarly evolving pages) should be refreshed.

Using 2,309 home pages from 27 different sites collected between 1999 and 2007, the goal of this paper is to quantify, visualize, and understand the kinds of content, structural, and use of external resource changes. This is not meant to be a definitive piece, only one that shows webpages indeed transform over time. For the context of this work, a home page is defined as the resulting web page from a given link in the full archive link page (see section 2 for further details).

The remaining sections of the paper are as follows. Section 1.1 has a brief literature review which is used to verify the measures and practices used in the paper and to highlight some of the interesting concepts in webpage evolution. Section 2 will cover the experiments setup along with tools used, data collections procedures, data processing techniques, and measures. Section 3 goes over the results of the 24 unique experiments. Section 4 is general discussions. Section 5 has the conclusions and future work ideas. And finally, section 6 is the appendix which includes graphs that are relevant to the analysis and conclusions, but were not explicitly mentioned.

1.1. Related Word

In [2], Bar-Ilam et al analyze web page changes in the informatics domain over a period of 5 years from June 1998 to June 2003. They determined web page change by running queries through a search engine and retrieving the links. They then analyzed the growth, modification, and disappearance of URLs from those lists. Their main contention was that not only should growth of web pages be considered in determining web page evolution, but pages change (50%) and disappearance (40%) as well.

In [3], Koehler randomly generated web pages (361) and crawled them weekly between December 1996 and February 2001. Changes in links and size in kilobytes (which he calls content) of the pages were monitored along with the actual status of the page (i.e. was it still there? What kind of error?...). The byte-weight of the page is computed only to show that there was an overall
increase in the number of objects; that’s all. He tracks each individual link from week to week to see if they remain the same, are modified, or removed.

Ivory and Megraw in [4] study 22,000 pages from over 1,500 sites from 2000 to 2003. They developed a system called WebTango that uses 157 page and site level measures to compare one site to another. The purpose of WebTango is to allow web site developers to compare their sites with those that are well-designed and see how they are similar and different. This by no means determines which is better, only how they are different. Of those 157 measures, those relevant to this paper include: text elements (amount of text), link elements (number of links), graphic elements (number of images), text formatting (includes tags such as \texttt{<p>}, \texttt{<b>}, \texttt{<i>}), and page formatting.

Dontcheva et al in [5], studied structural changes in 12,000 webpages from 20 different sites over five months. They broke the source code down into its Document Object Model (DOM) and then compared the structures from one period to the next. Their analysis used two strategies. The first was a strict strategy which stated that at the first sign of difference in the structure, the pages were considered different. The second was the flexible strategy which compared sections of a document together and the number of changes was recorded. Also, tag depth was analyzed. They concluded that most changes occur closer to the leaf nodes than root nodes and are therefore minor alterations.

Fetterly et al in [6], crawled 151 million pages a week for 11 weeks. For 0.1% of the pages, the source was stored and for the others, only a few attributes such as a feature vector and HTTP status codes were saved. Their goal was to determine which changes represented a greater intensity of change. Like Dontcheva et al, Fetterly et al find that the majority of changes in a webpage are small and usually insignificant. However, they find a strong relationship between the top-level domain and the frequency of change. The .com and .net domains changed more frequently than the .edu and .gov domains.

Risvik and Michelsen in [7], focus on how the large-scale of the web pose significant challenges for web crawlers. They talk about how the web grows at an exponential rate and one problem associated with that is the refresh intervals for web pages. They suggest that knowing how a web page changes and setting minimum and maximum intervals can significantly increase the efficiency of refreshing a site about the same time, but after, it was altered.

Candan et al in [8], examine how to map two tree structures that you want to compare onto the same \( k \)-dimensional space. Using k-means clustering, they can determine which notes from one tree are similar to those in another. Using web page source code collected from CNN over a period of time, they showed that their mapping system, even with trees that are only 35% similar, can obtain a precision of 70% (at 95% similar, they get virtually 100% precision).

In [9], Ryan et al gather screen shots from the 50 states home pages between 1997 and 2002 using Internet Archive. Initially, they wanted to use the actual sites, but with the slow response times from Internet Archive, they decided to go with screen shots. Their goal was to see how people viewed the importance of web site elements. The participants in the study were 180 students who grouped images of home pages together by visual similarity. They found that three primary dimensions can explain perceived variations in home pages: page layout, its navigation support, and its information density.

2. EXPERIMENTAL SETUP

In order to compare webpages from different points in time, I first needed to locate a web archival site. Internet Archive\(^1\) was such a site. It claims to have over 85 billion pages in its repository. Not only does it crawl and store full web sites, but moving images, live music, audio, and texts (printed text) as well [1]. There are several ways to view an archive for a particular web site. The first is by an individual year, categorized by month, and the second is of the full archive by year (see Figure 1). Internet Archive actually store data back to 1996, but since the number of stored pages for 1996, 1997, and 1998 was very small (only a few for the entire year), those years were discarded.

The next logical step was to select the categories and sites to crawl. The categories I settled on were school, news, online shopping, and software. I chose each category for a specific reason. My initial assumptions are as follows. School was chosen because the sites are generally more static in content and (remember, I’m only dealing with home pages) would probably see more changes in structure than content. For news, I expected a change in structure, but more importantly, content changes. For online shopping, I expected content to change somewhat, but what I was really looking for was an increased use in external objects. Finally, from software, I planned to see a mild change in both areas.

Now that I had the categories, it was time to choose the sites. I chose each site for one main reason, it was a well known site that I believed would continue to grow with new trends in order to appease an increasingly

\(^1\) http://www.archive.org
technologically savvy user base [10]. Table 1 shows the original list of 36 web sites on Internet Archive crawled by category (these 36 were all archived by Internet Archive).

For each site, the first link listed for each month was retrieved to reduce the overall set of pages that needed to be crawled. This was mainly due to the slow response time of Internet Archives, averaging between 10 and 20 seconds per request (taking several minutes to fully load the entire page) per page to retrieve (also seen in [9]). For a given site, the links page (Figure 1) had around 1,200 links; the crawler would have finished in about 7.5 days assuming 15 seconds on average per retrieval (without waiting for images to load). By taking only one per month, the time was limited to around 18 hours.

Once the pages were crawled, I had to remove 9 of the 36 original sites since more than 30% of the pages crawled for those sites returned error codes (bad links to the stored pages, see started sites in Table 1). The 30% threshold was chosen because that is equivalent to two and a half years missing data. This is quite a bit of missing data, but since I am interested in aggregating the information by category and year not by individual site, the effect of missing data should be minimal. But this is something to address in future experiments; completeness of data.

Now that all of the data had been collected, it had to be processed and measured. The pages were converted into their Document Object Model (DOM) in order to determine the average depth of a node. Also, the data was stop-worded and stemmed to calculate the cosine similarity between pages linearly across time. Finally, the original source code was parsed for counts of specific tags. These processes and more will be covered in sections 2.3 and 2.4.

<table>
<thead>
<tr>
<th>School</th>
<th>News</th>
<th>Online Shopping</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Iowa</td>
<td>CNET</td>
<td>Amazon</td>
<td>Microsoft</td>
</tr>
<tr>
<td>University of California Sand Diego</td>
<td>ABC News</td>
<td>eBay</td>
<td>FileMaker</td>
</tr>
<tr>
<td>Harvard</td>
<td>CBS News</td>
<td>Buy.com*</td>
<td>Microsoft SQL Server*</td>
</tr>
<tr>
<td>MIT</td>
<td>CNN</td>
<td>Overstock</td>
<td>My SQL</td>
</tr>
<tr>
<td>California State University San Marcos*</td>
<td>MSNBC News*</td>
<td>Walmart</td>
<td>Mac</td>
</tr>
<tr>
<td>San Diego State University</td>
<td>New York Times*</td>
<td>Target</td>
<td>Red Hat*</td>
</tr>
<tr>
<td>University of California Los Angeles</td>
<td>Reuters*</td>
<td>JCPenny*</td>
<td></td>
</tr>
<tr>
<td>University of Oregon</td>
<td></td>
<td>Macy’s</td>
<td></td>
</tr>
<tr>
<td>Portland State University</td>
<td></td>
<td>Kohls</td>
<td></td>
</tr>
<tr>
<td>Washington State University</td>
<td></td>
<td>Dillards</td>
<td></td>
</tr>
<tr>
<td>University of Washington</td>
<td></td>
<td>Staples</td>
<td>Office Depot*</td>
</tr>
</tbody>
</table>

Table 1) Original 36 sites crawled; listed by category, starred are those that were removed
2.1. Tools used

For this project, four different tools were required for processing and retrieving, storing, and charting the data. For charting the data, Microsoft Excel 2007 was used. For data storage, FileMaker Server 9 Advanced (FMSA 9) was used. FileMaker 9 Advanced (FMA 9) was used to create the database (which was then hosted by FMSA 9 after switching to Java), perform the initial crawling and link parsing, HTML tag analysis, and storage in the beginning. And finally, for the bulk of data processing and retrieval, I used Java. A simple JDBC connection was established between FMSA 9 and Java. FMA 9 was then used to export the data directly into an xls file (built in feature) which was then used to generate the graphs.

2.2. Collecting the data

In the beginning, I had 36 sites to crawl. The first thing I had to do was create 36 seed URLs. These initial seeds would point to the full archive page of links from which the individual links would be extracted and then added to the frontier for crawling in a FIFO manner (see Figure 1). The format for the URL generated is as follows: http://web.archive.org/web/*/<web address>/. For example, the University of Iowa would be http://web.archive.org/web/*//uiowa.edu.

In the beginning, I decided to use FMA 9 for all of the crawling, processing, and analysis because I am very familiar with the software and since version 8.5, a web interface was added. This interface is nothing more than a browser window without the tool bars. Using scripts, I can pass variables to the web viewer (as it is called) and extract many different items; the most relevant to this project is the source code of a webpage. Once the source had been extracted, the pages were parsed for relevant anchor tags and one link for each month and year was obtained and added to the frontier. By relevant anchor tags, I mean those whose href consisted of the domain for the archival portion of Internet Archive; web.archive.org. Using the example site from above (http://uiowa.edu), the links to archived pages look something like this: href=http://web.archive.org/web/19970606103026/http://www.uiowa.edu/. As you can see, web.archive.org is present in the link. Also, I checked to see if the character ‘*’ exists. This is a wild card character and is used to obtain sets of pages that fall into different ranges such as month or year. I removed these because they are used to constrain the current found set as a search option and therefore have no value to this project. All links that pass these tests are added to the frontier.

I began the crawl of the 36 archived pages and 12 hours later, it had not even finished the initial crawl let alone started on the new links. What I found was that the web viewer had to load the entire page including images before the source could be extracted. As mentioned earlier, Internet Archive is a very slow responding site. In order to use FMA 9, it took on average 10 minutes to fully load a webpage. I knew this would not be feasible for thousands of pages. So, while FMA 9 was crawling, I began to work on a Java crawler.

I stopped FMA 9 once it had crawled the original 36 seed sites (about 24 hours). Then, I uploaded the FMA 9 database into FMSA 9 and begin the Java crawl. The process was simple. Since I was only interested in the webpage that first came up when the link was followed, all the Java program needed to do was connect to FMSA 9 and while iterating through the records, retrieve the link to crawl, crawl it, and return the webpage source to FMSA 9. At this point in time, I had not figured out what measures I wanted to use so all I needed to do was store the source for later analysis. After this process I removed 9 of the sites due to a lack of successful crawls (most of which were missing 50-75% of the data). From the original 36 seeds, 2,854 pages were found. After removing those 9, 2,309 pages remained; all further analysis will be done on these 2,309 home pages.

2.3. Processing the data

Before any of the processing began, I created a simple Java file creation script to store the webpage source code for each record in FMSA 9 as an htm file. The name of the file was the primary key for that tuple (e.g. 374.htm). This course of action was necessary in order to process the source using the HTMLEditorKit class in Java.

To analyze structural changes, a DOM was generated for each webpage. Using the HTMLEditorKit in Java, the source was parsed for start, end, and simple tags (simple tags being those that are both a start and a stop e.g. <meta ....../>). Those tags are then used as a node in the DOM hierarchy. Each webpage was then stop-worded 2 and stemmed 3 to prepare them for measuring similarity (more in section 2.4). Finally, each webpage was processed for total and unique links using a FMA 9 script.

2.4. Measuring the data

There are three types of measures used in this paper. The first is tag counts. In FMA 9, counting a tag is quite

\[\text{List of stop-words from:}\]
http://www.lextek.com/manuals/onix/stopwords2.html

\[\text{Porter Java Stemmer:}\]
http://tartarus.org/~martin/PorterStemmer/java.txt
simple. The function is patternCount("<attribute name>",
<pattern to count>"). In order to count anchor tags, for
example, the function is patternCount(source; "<a "
where source is the name of the attribute containing the
webpage source data.

The second type is DOM depth. Continuing from
where we left off in Section 2.3, while the tree is being
generated, the depth of each tag is being recorded. Once
the file is fully processed, the average node depth is then
computed by taking the depth from each node and
dividing it by the total number of those nodes.

Finally, the third type is cosine similarity. Two sets
of calculations were derived from the data. The first was
site specific. Before this value is calculated, the data is
sorted by domain name and then year. Only pages from
the same site were compared so the first page for each site
did not have a value since there was nothing to compare it
to (thus the range of 2000 to 2007). The second was a
categorical average using the previous values. This was
used to check whether or not a particular class changed
more or less than the others. Cosine similarity is
calculated using the following formula.

\[
\text{sim}(v_p, v_{p-1}) = \frac{v_p \cdot v_{p-1}}{\|v_p\| \cdot \|v_{p-1}\|}
\]

(1)

Where \( p \) is the current page from the current site and
\( p - 1 \) is the previous page. \( v_p \) is the vector of stop-
worded and stemmed words for the current page along
with their frequencies. \( v_{p-1} \) is the vector of terms and
frequencies for the previous page.

Taking you back to the beginning of this paper, due
to the lack of working links, I decided the minimum
number of required pages per year was 4 (one per
quarter). Therefore, once all of the individual measures
were calculated, they were aggregated by category and
quarter. After charting most of these measures by year, I
noticed that the change per quarter was negligible at best.
So I decided to further the aggregation by category and
year. This provided much better results and is the
aggregation method depicted in the following graphs.
Also, for comparison, the overall averages were
calculated by year. These graphs can be seen in the
Appendix A (section 6).

All together, there are 24 unique measures by
category and year: average links, average unique links,
percent change in links [1-(average unique links/average
links)], average bytes (the size of the page in bytes),
average images (images on a page), average meta (meta
data tags), average style sheets (linked style sheets),
average scripts (scripts in the document), average div tags
(divider tags), average Embed (embedded objects),
average tables, average paragraphs, average H1, average
H2, average H3, average H4, average Li (list items),
average B (bold tags), average I (italics tags), average
Strong (same as bold), average Em (same as italics),
average Style (internal style sheets), average depth (this is
DOM depth), and average cosine similarity. I also tested
for tags \(<s>, <u>, <center>, <h5>, <h6>, and <font>, but
the vast majority of the values were 0 so they were
omitted from the results. Most of the tags covered in this
paper directly correlate to Dontcheva et al [5]
experiments. In addition to the average cosine similarity,
which is computed using the averaged for each category, I
calculated the 27 individual cosine similarities. I will
cover several of the more interesting ones in section 3.6
and the rest are in Appendix B.

3. RESULTS

First, the overall size in bytes is covered to show how
the pages grow internally in general (Figure 2). Then,
link analysis is performed. The average number of links
class per category and year is computed, then the average
number of unique links is derived from those, and finally,
the percent fewer links is found using the previous two
(Figure 3). Next, the change in header information and
scripts is computed in order to see how the use of external
style sheets, internal and external scripts, and the use of
meta data to describe the site had changed over time.
Then text formatting comes (Figure 4). This is where the
changes in text emphasis is covered (Figure 5). After
that, body structure is examined in much the same manner
as text formatting (Figure 6). Finally, DOM depth and
cosine similarity are examined (Figure 7).

Almost every graph increases over time indicating an
increase in the number of general HTML tag usage. In
general, news sites lead the way in the increased use of
tags. This paper does not determine whether or not any of
these increases or decreases are positive or negative;
simply that a change exists.

For a more thorough analysis, the six sections
discussed in the opening of this section are covered in
greater detail.

3.1. Size in Bytes

For an overall look at webpage growth in size, I
computed the number of bytes per category and year. As
you can see from Figure 2 below, the general pattern is
upward with only software indicating a significant
decrease in year 2005. There could be several
explanations for this, but as you will see later on, this
year is a significant one for software.
3.2. Links

Figure 3(a) shows the average number of links per page by category and year. As expected, the number of links increases. I say as expected because as more and more computer savvy users view these pages, they can handle an increasing number of options and information on a given page. Also, with the advancements of web programming tools (such as Dreamweaver, Visual Web Developer, and FrontPage) it has become increasingly easy to add, edit, and remove elements (both simple and complex) from a site. Furthermore, the number of links for news sites far exceeds those for the other three categories. This is also expected since news sites generally have numerous links to both internal pages and external pages containing both current and past stories.

Looking at Figure 3(b), it appears that the number of unique links follows the same course as overall links only on a reduced scale. But as you can see from Figure 3(c), there really are not any clear conclusions that can be drawn from the oscillations. Basically, from one year to the next, it is nearly impossible to figure how the distribution of unique to overall links will be. Generally speaking though, it follows an increasing pattern.

3.3. Header Info and Scripts

In most cases, style sheets, meta data, and scripts (to a lesser extent) are found in the header section of a web page. Scripts can be found in the body section as well. What brings these three together is the notion of other than HTML source and/or use. For example, style sheets call external code stored in .css files. These style sheets can either change the look and feel of a page slightly or significantly. Either way, the code is not present in the HTML document. For scripts, the code ran is other than HTML code (such as JavaScript and VBScript). Finally, meta data has no affect on the web page, but on crawlers (for example).

As you can see from Figure 4(a), the general use of style sheets is on the rise with the exception of news sites. Those sites seem to have suddenly shot up in the use of style sheets from 2000 to 2001 and hovered for about
three years and then scaled back 50%. A reason for this could be that news sites wanted a way to increase the overall look and feel of their sites so they decided just add style sheets instead of making changes to existing ones. Then they either combined the files or started over.

Figure 4(b) is less exciting depicting only a slight increase in the number of meta data statements. In the future, I would like to go back to this portion of the data and see how the content has changed over time (i.e. the number of key works in each tag compared to previous years). The reason being, the number of tags does not necessarily need to increase for the meaning to change; because as you add content to your site, you need only update the tag data.

Figure 4(c) has a similar pattern as meta data; a slight increase. News, again, is increasing at a pace greater than the other categories. This could indicate a boost in the use of dynamic data or HTML code, fancy features such as mouse over menus (Dreamweaver has an interface to add navigation bars with multiple images which does all of its work using JavaScript – and a lot of it), image pre-loaders, and embedded objects (as seen in Figure 6(d)).

3.4. Text Formatting

Text formatting is used to emphasize a word, phrase, sentence, or passage of importance. Here I measure the use of 8 distinct values covered by 10 tags. First, I will talk about header tags 1-4 (as mentioned before, 5 and 6 are virtually nonexistent so they were omitted from the results).

One would expect to see headers used across all sites. They are simple to use and produce the desired effect. Not surprisingly, H1 tags (the largest of the 4 seen in Figure 5(a)) is used sparingly on home pages. What was surprising was how these were virtually nonexistent in code before 2005 (with the exception of the spike in 2003 for software sites). H2 tags (Figure 5(b)) are still rather large and I expected to see similar results for H1. The numbers were relatively the same, but the use of these tags actually began its rise in 2003. For H3 tags (Figure 5(c)), one would expect to see more of those than H1 and H2 because they are smaller and can be used to emphasize delineations between segments of text as opposed to general page headers. The data seems to support this, but again, there is not an increase in its use until around 2005. Finally, H4 (Figure 5(d)) is a tag that is not used as much because of its size. Usually, H4 is not much larger than the text itself so bold or strong might be another option for developers. As expected, the numbers are low, but surprisingly, the graph indicates that it is used with similar frequency to H1 and H2.

The text formatting tag covered is list item (Figure 5(e)). This tag is used to either bullet or number data depending on the parent node (<ol> of ordered and <ul> for unordered). I was not sure what to expect from this element other than an increasing use. What was found
was this tag was not very popular in the past, but is used upwards of 80 times a page; the highest in 2007 being online shopping. This can be attributed to lists of products, details, or maybe sales.

Bold and strong tags (Figures 5(f)(j)) are also used to emphasize text. This is a very common feature in web sites. The reason I used both bold and strong is for older pages, strong was used more as opposed to today where bold is more prevalent. What I found was what was expected. For bold tags, news sites were far and away leaders in their use. This is because news sites generally have many stores on their home pages and can differentiate between them using bolded titles. Strong, on the other hand, seemed to be dead around the time this data was collected. The only real use was in 2003 and that could be attributed to one page in the set using it.

Italics and em tags (Figures 5(g)(h)) are also used to emphasize text. And as with strong, em is an older tag that is also not used very often as can be seen with a high of about 2.5 in 2006-7. Again, this could be caused by a single page. Italics seem to be just as important as bold especially in news sites which run away again from the rest of the other categories. It also appears that italics is more popular (while only slightly) with the other three categories than bold. Just as with bold, italic text can also be used as a separator between sections. Furthermore, the similarity of these results might also indicate that bold and italics are used in conjunction.

One would expect that with an increase in the number of external style sheet links that the use of internal style sheets and text styling would decrease. However, the data suggests that the use of internal style sheets and internal text styling is on the rise and again, news sites are leading the way (Figure 5(j)). To most website developers, this just leads to a larger amount of work, but if the goal is to make changes to just one page instead of a set, then this could be expected. Also, the number of styles per page is only around 3 so it is reasonable to say that these are page specific stylistic changes and that the bulk of styling information is still coming from external style sheets.
Figure 5) (a)(b)(c)(d) are the average number of header tags used.  (e) is the average number list items.  (f) is the average number of bold tags.  (g) is the average number of italics tags.  (h) is the average number of emphasis tags (same as italics).  (i) is the average number of strong tags (same as bold).  (j) is the average number of internal style sheets used.

3.5. Body Objects

Body objects just means elements that are used to form the way data is sectioned and how media is presented. The first object discussed is the paragraph.

Since I am only dealing with home pages, I expected to see a relatively static number of paragraphs. This is because the home page is usually used to give the viewer an overview of what the site is about and some pertinent information. It is not used as an information dispensing tool. Looking at Figure 6(a) shows exactly this with a very odd spike in the number of paragraphs in software from 2006-7. This could be attributed to an increase in the amount of information placed on a home page. Microsoft, for example, now has a tabbed feature which allows a lot more information on its home page without increasing the overall length of its page.

The next elements covered are images (Figure 6(b)). From the graph, there are three interesting things happening. First, the number of images for software and school in 2007 is lower than in 2003. This might suggest an increase in other media or images/graphics being placed by style sheets. The second is although online shopping has almost doubled since 1999, it is now half of what it was in 2004. This might also lead to the same conclusions as before. The third is of no surprise, news sites again lead the way. This is because a picture goes hand-in-hand with a story. News sites often display a title and truncated description of a story and an image. Surprisingly, it too is starting to oscillate and taper off, but it might simply mean that at around 150 images a page, they have reached their max capacity.

One thing not mentioned above is the use of images for buttons and navigation. This too will increase the overall numbers, but will not be seen by most viewers as an image, but more of a graphic. Right now, unless they are set by the style sheet, I have no way of determining whether or not an image is a picture or a graphic.

There are many ways to embed an object on a web page, but the only one I measured for was the most frequently used embed tag. Object embedding is used to add multimedia content such as Flash and QuickTime movies to a site. I have to say that after looking at Figure 6(d), I was rather surprised. I fully expected to see a jump in its use somewhere around 2003-4. But the data suggests there are on average less than 1 per home page. This could be attributed to my site selections. If I would have picked television sites such as ABC, CBS, and NBC, I would have seen the increase I was expecting since these sites use a lot of Flash.
That is, changing at the same rate. This is unexpected because the news changes continuously and the home page is the first place users stop before navigating to the top stories. This might simply mean that instead of text, news sites are moving more towards video, images, and headlines which gains support from the results in 6(b).

3.6. Tag Depth and Cosine Similarity

Both of the following results were a little surprising. Figure 7(a) shows the average tag depth by category and year. As you can see, depth oscillates between 1.65 and 2.25 and ultimately ends down. Up until now, virtually every measure has seen some sort of incline so it was quite interesting to see these results. Also, I thought that as bytes increased (which only measures the size of the text in the file), so would, to some extent, the depth of tags. That is, I expected more information to be added on to existing data as opposed to simply more data.

Figure 7(b) shows the average cosine for all 27 sites in their respective categories. In general, the average content remained steady at around 70% similarity from year to year. Not surprisingly, news had the lowest score, but it made its way back up to the rest of the pack by around 2005. What this suggests is the content of news sites is becoming more in tune with the other categories. Figure 7) (a) is the average depth of the DOM tree. (b) is the average cosine similarity.

Below are four cosine similarity graphs for individual sites. These were the most interesting of the 27 which is why they are shown here. For the entire set, see Appendix B.

For CNET (Figure 8(a)), if years 2000 and 2003 were removed, this site would have seen little change at all. As seen above with the average cosine values (Figure 7(b)), most news sites were on average 20% lower in the amount of similar content. This might be because CNET’s home page is covered with advertising, lists of
blogs, categories of reviews, and very little news content. But if the extra content was removed from the calculations, then the pattern below might follow more traditional story based content home pages as seen with the other sites in this category. This might actually be an interesting research topic in the future. Can you build a wrapper that can be used to determine what is and is not relevant in ascertaining the evolution of a given site?

Figure 8(b), Harvard, definitely stands out from the rest of the school sites in terms of how often its content changes. Most of the school home pages are relatively flat and between 70-80% similar to the previous years, but as you can see, in 2003, Harvard began making massive changes to the content on its home page and continued doing so through the end of the data set. I visually inspected some of these years and in 2003, Harvard completely changed its home page. This can explain the sharp decline, but not the continual low results. The constant state of change can be attributed to the fact that most of the textual content on the home page are current issues related. That is, the home page is used for navigation as opposed to information dissemination.

Kohl’s (Figure 8(c)) shows how volatile information on online shopping home pages can be. For this category, the average was quite misleading. Some of the sites saw little or no change over the course of years which combined with those with varying change, forced the average up.

Finally, MAC (Figure 8(d)), is the only software site to drop below 60% similarity. This can be attributed to MAC’s unique use (for its category) of content space on the home page. Instead of general information and links, MAC highlights its new products, product features, and other information as a news site would; in the form of headlines and brief descriptions. Therefore, as the company progresses, so does the content on their home page.

Figure 8) cosine similarity for (a) CNET, (b) Harvard, (c) Kohl’s, and (d) MAC.
4. DISCUSSION

As you have seen, website home pages have continuously changed over time; some more than others. With the increased use of external references such as style sheets and scripts (both internal and external) it is imperative to take both the functionality and stylistic changes these add to the structure, content, and text formatting when performing web site analysis. Just searching for tags might no longer be enough to determine text emphasis and structure in an increasingly dynamic environment. Also, different categories of sites follow the same general pattern (up), but at varying degrees of severity. This could be used to determine how often a web page needs to be refreshed. As Risvik and Michelsen mention in [7], knowing how a web page changes and setting minimum and maximum intervals can significantly increase the efficiency of refreshing a site shortly after it was altered.

Now, I want to go back to the assumptions I made about the categories at the beginning of the paper. First, I expected school home pages to have more structural changes than content because the bulk of the information on these pages were informational and used for navigation as opposed to current events based. I believe this was supported by the experimental results. The structure (tags and use of external resources) steadily increase over time for the majority of the measures. As for the content, every school except Harvard had a sustained high cosine similarity score.

Next I spoke of news and how both structural and content changes (especially content). Again, the data supports the structural changes as the use of almost every tag increased. As for content, between 2000 and 2005, the similarity score was lower than the average, but then, it jumped back up. I mentioned earlier that this could be due to an increased use of multimedia as a textual replacement. Therefore, I believe my assumption seemed intuitively correct at the onset of the experiment, but was not upheld entirely by the results.

For online shopping, I expected content to change somewhat, but what I was really looking for was an increased use in external objects. The reason being, online shopping generally uses more dynamic features such as session variables and user specific content, it track shopping cart values, and are usually aesthetically pleasing (images, graphics, multimedia, use style sheets,…). What the results show is that is just not the case when compared to other categories. The use of external resources is on the rise, but that is shown for all.

Finally, from software, I planned to see a mild change in both areas. I expected to see a mix because products for these companies are released only a few times a year at the most and they want to keep their sites up to date so the structure ought to mimic the current trends. For the most part this is supported. If anything, the results suggest that the structure might actually change at one of the slower rates of the four. But for the content, it is very stagnant at around 70-80% which would be much higher on average if MAC was taken out.

5. CONCLUSIONS AND FUTURE WORK

The goal of this paper was to show that web pages evolve over time and in an expanding manner. I believe that I have shown enough evidence to support this claim. The tests performed need to be extended and the number of sites increased in all categories to say for certain how a home page changes over time, but I think this little bit is enough to warrant further research. One particular area that deserves an increased focus is the manner in which external code from scripts and style sheets affect webpage evolution. As we saw in Figure 4(a), however slightly, external style sheets are on the rise and from Figure 4(c) it is obvious that external functions (scripts) are becoming extremely popular.

In the future, I would like to expand the number of sites and categories. Also, instead of just settling for one home page every quarter, I will set aside more time to get multiple pages per month. This should increase the accuracy of how often a single home page changes over time. That figure could then be used in more general terms to determine how often similar pages would change and thus, how often to re-crawl. Also, I will focus more on data completeness and collect data from more sites than I truly need since Internet Archives (although a great source for archived web site data) seems to have link problems and does not store a comprehensive archive of data (gaps in the crawled sets of sites). Or I could even find an alternative archival source. Furthermore, instead of just using home pages, I would like to expand to content pages as well. As mentioned earlier in the paper, I would also like to expand the content search of meta data tags to see how the key words and phrases are changing over time. This information can also be used in conjunction with the page information above to help determine when a site was updated and how the site developers feel the site should be placed by search engines. Finally, I would like to compare the static categories to clusters derived from a classifier. As seen in section 3.6, there are sites in categories that do not follow the trends of those whose purpose is the same. Likewise, there are sites whose patterns are more similar to those of another group.
6. APPENDIX

Appendix A and B have been removed from this paper to save room. To view these graphs, go to:
http://instruct.biz.uiowa.edu/courses/6k070AAA/rhylock/papers/WebpageEvolutionOverTimeAndCategory.pdf

7. REFERENCES

Online annotated bibliography:
http://instruct.biz.uiowa.edu/courses/6K070AAA/rhylock/papers/WebpageEvolutionOverTimAndCategory.htm


