

# Using the Internet Archive for measuring web site change: A methodology

Andrew King<sup>1</sup>, Florian Stahl<sup>2</sup> and Savvas Papagiannidis<sup>3</sup>

*As Internet technologies advance and design trends evolve, so does web site design. With the increasing significance that web sites have for all organisations defining and measuring change over time can have important practical implications for marketing purposes. The aim of this paper was to analyse the development of web sites over time, measuring which elements change and how much. The element list was compiled based on the human computer interaction literature. A system was then built to collect and analyse web sites using the design elements list. When it came to the data collection, we used the Wayback Machine available as part of the Internet Archive. Our longitudinal analysis spanned a decade of web site development. The paper discusses the methodology adopted and the early findings of our research.*

Keywords: web site change, human computer interaction, emarketing, Internet Archive

## 1. Introduction

With the size and importance of the World Wide Web (WWW) exponentially increasing and with a true understanding of its dynamic nature remaining elusive, web site change is an interesting phenomenon, warranting studies at various levels and scales. For example, there have been studies ranging from in-depth focus on how a single web site evolves over time (Calzarossa and Tessera, 2008), to pervasive studies that investigate change of the web as a whole (Fetterly *et al* 2004, Cho and Garcia-Molina 2000). The benefits and challenges of the qualitative and quantitative approaches methods are accentuated, due to the sheer size of the web combined with the inherently subjective nature of web site evaluations. Although by definition, varying a web site design results in change, agreeing as to the extent and significance of a particular change may not be as straight forward. Diverging opinions promote the usage of qualitative methods for studying change, as they enable capturing the richness and essence of change. However, the size of the WWW, combined with the ability for automated programs to analyse change, means that quantitative methods are equally important. This is evident, for example, in Fetterly *et al*'s (2004) investigation, during which over 150 million web pages were crawled weekly for 11 weeks.

---

<sup>1</sup> Andrew King, Business School, Newcastle University, UK, Email: a.l.king@newcastle.ac.uk

<sup>2</sup> Florian Stahl, Business School, Newcastle University, UK, Email: florian.stahl@newcastle.ac.uk

<sup>3</sup> Dr Savvas Papagiannidis, Business School, Newcastle University, UK, Email: savvas.papagiannidis@ncl.ac.uk

Web sites evolve over time either because their owners want to add more features, improving the quality of service or because advances in technology offer new capabilities. Of course, change for the sake of it, may not have the expected results. For example, considering a user cognitive lock-in (Casalo *et al* 2008 p.329) and that a familiar web site is more likely to entice consumers to return, one could argue that large scale redevelopment could even have potentially a negative impact. An example of this clash can be seen in the case of Facebook that overhauled its interface, in order to improve its offering, but managed to alienate many of its users in the process, with over 500 groups dedicated to bringing the old design back.

This paper does not aim to investigate the change management process itself or even what constitutes good change for individual cases. Instead, by focusing on relatively large changes within web sites, this investigation aspires to create an automated evaluation framework for measuring longitudinal change within a cluster of companies or for a specific technology. After outlining the relevant literature on which our evaluation framework was based, we will present the algorithm and technology used for our software. We will then present a few early findings while testing it and offer ways that it could be potentially extended.

## 2. Literature Review

The web site design literature was reviewed with the aim of identifying an extensive list of web site design features, which could be automatically retrieved, and subsequently analysed over time. However, due to the subjectivity inherent in any evaluation of a web site, there is little agreement on what specific features one should use. Taylor *et al* (2002) interviewed IT staff from 25 organisations and concluded that web design was dominated by *ad hoc* methods. Still, there exist lists of web site design features in the literature, although due to the subjective nature of web design, they tend to focus on qualitative investigations. Such lists may contain features such as: use of humour, eye catching images, accurate information/screen layout and a logical structuring of information (Zhang and von Dran, 2002). Consequently, their adaptability for a quantitative approach is limited. The web site design feature list produced by Zang and von Dran (2002) did highlight the use of a loading bar to indicate how long one would have to wait, the presence of multimedia, and the use of a mark as to one's location within a web site. Taylor *et al* (2002) mentioned pluralistic and multipurpose designs. For example, do web

sites have different navigational pathways for different users? Cyr *et al* (2009) assessed the extent of colour usage across web pages using a percentage of page covered mechanism. Teo *et al* (2003) highlighted the importance of the search function, and came up with an interactivity scale based on the various features of a web page, such as the ability to move objects around the screen. Xie and Barnes (2009) utilised the Wayback machine to conduct a longitudinal study of airline company web sites, in Britain, over a seven year period. However, the focus of their investigation was opinion-based, which is a challenge when it comes to adopting a quantitative approach.

While studies like those mentioned above show that the literature did provide some interesting design features, which potentially, could have been incorporated into the evaluation framework, this proved to be more complex than initially envisaged. This problem was faced previously by Bauer and Scharl (2000) who pointed out that while many design features may initially look promising, too many of them raise issues when attempting to compute them. Difficulty in adopting aspects of qualitative investigations, for those with a quantitative focus, was to be expected. However, even the quantitative investigations did not significantly build upon the evaluation features originally considered (Bauer and Scharl 2000). One exception was an extensive list of criteria for the quantitative evaluation of web sites by Olsina and Rossi (2002). This list though was domain specific (a case study of an online book store was used as an example) and as such not everything can be generically used.

Instead of analysing web site change through a list of design features and assessing how the usage of these features varied with time, we decided to analyse web site change per se. This raises the question of what 'change' actually constitutes. For Cho and Garcia-Molina (2002) change occurred when a page was modified. They crawled 720,000 pages on a daily basis, from 270 sites over a four month period. Although the investigation was very large, there was no information about the degree of the change that occurred. More recently, Artail and Fawaz (2008) note the growing usage of RSS (really simple syndication) feeds, which introduce regularly changing content to a web site. This raises the question of whether or not something can be considered as a change, if it is in fact made to change. This is something that the framework implemented by Cho and Garcia-Molina (2002) would not be able to resolve. Koehler (2002) used a framework of change that was similar to Cho and Garcia-Molina. They defined changed as a binary construct depending on whether or not the byte weighting on an individual page had changed. Whilst

this has similar limitations to checking if the page has been modified or not, Koehler (2002) went a step further by creating an index for web site change, by calculating the percentage of a web sites web pages' that had changed. This method is not suitable for this investigation, as our aim is to investigate large changes, and a binary change in byte weighting provides no insights into the significance of the change. Fetterly *et al* (2003 p.217) analysed the words and their sequence within a document, and compared them to the version previously saved. This method does allow for the degree of change to be calculated, but only with regards to how the words change.

As it has become clear while reviewing the literature, compiling a framework for an automated quantitative approach is not a trivial task. Following the literature and our own ideas while reviewing web technologies specifications, we compiled the list of features summarised in Table 1. These were used as part of the methodology presented in the section following.

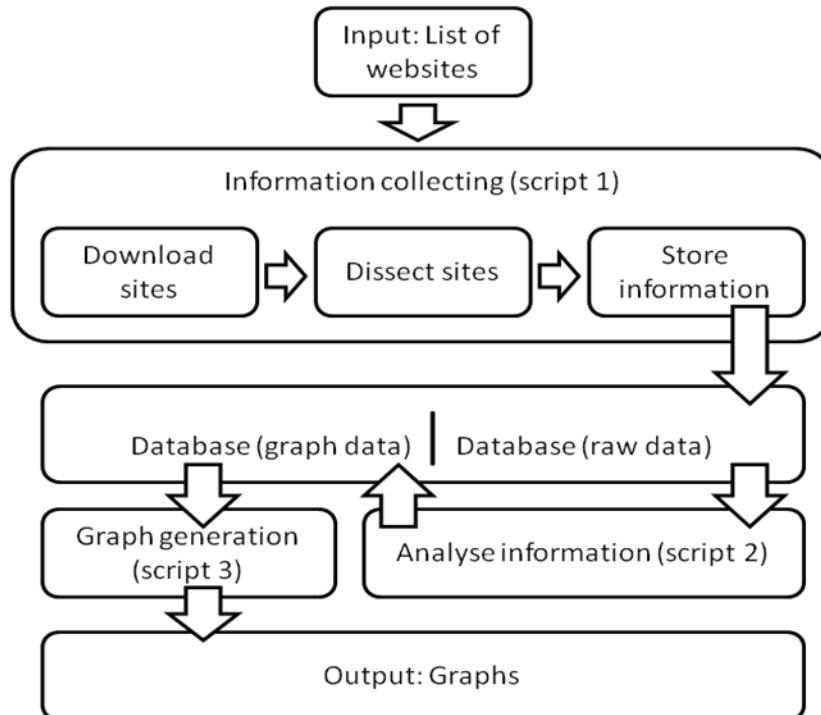
**Table 1: Web site features used in our investigation**

<b>Criteria</b>	<b>Description</b>
Colours	The amount used, the number of different shades.
Images	The amount used and their file size.
File size	Total file size of the web site.
Structure/Complexity	Whether tables or divisions are used, and how many.
Links	Do they link internally or externally, is https used.
Contact details	If these are included on the homepage or not, evidence of trying to build trust
File extensions	HTML or PHP or other as an indicator of how dynamic the web site is.
Optimal browser/resolution	Do companies recommend a browser or a resolution.
Technology	What technology (or combination) is used to construct the homepage: CSS, Flash, Java (applets) and JavaScript .

### 3. Methodology

This section presents the methodology for measuring web site change and the software used to automate the process. The overall process is illustrated in Figure 1.

Figure 1: The data collection and analysis process



#### 3.1 Software

The software was built using PHP and MySQL, using 2 free libraries for retrieving web pages and drawing analysis graphs. Application development revolved around three scripts that undertook the key tasks required. Additionally, another eight scripts were built, in order to administrate and serve the core functionality.

#### 3.1 Data collection

The main script is the data collector script which retrieves web pages from the Internet Archive and more specifically the Wayback Machine, which allows browsing through over 150 billion web pages (<http://www.archive.org/web/web.php>). The script retrieves one page per month over the period from 1996 to 2008 (2008 is the latest year for which data is currently available in the Archive). Over the available 13-year period, this results in 156 records per web site analysed.

The Archive provides a URL-based interface to automatically return an archived version of the page as close as possible to the date specified. Consequently, it was not necessary to input specific dates to retrieve pages. A specification in the format of YYYYMM was sufficient to retrieve the desired home pages (analysis is limited to home page). As not all sites are available on a monthly basis, this implies that the archive may not return a record for the requested date. Our checks indicated that when it came to popular corporate sites the vast majority of them were available. Other minor issues that the software had to deal with included computation of the URL leading to another date than specified (sometimes to live versions of a web page) and incomplete data.

Once the data collection script completes downloading a web page it starts dissecting it using regular expressions and other string pattern matching functions (a full text copy of the web page is stored for each evaluated web page, in order to enable subsequent analyses of the data gathered). An advantage of this is that a part of the analysis is done before the data is stored in the database, reducing the computations necessary to calculate the changes. Factors examined at this stage are the use of technologies such as Cascading Style Sheets (CSS), Adobe Flash, Java and JavaScript. In addition, the length of code and text, as well as the numbers of images, links, colours (only Hexadecimal definitions are counted) and the “under construction status” are also examined. Colours can only be counted if they are defined in valid Hexadecimal format, in order to avoid potential confusion. It is worth noting that as the script uses pattern matching across the entire available web page code, all instances matched are counted. For example, in one case, functionality was coded and then placed in the page’s comments, where it did not actually work. As comments typically do not account for a significant proportion of a page this was not deemed to have a significant impact on our analysis. However, future versions can potentially address this issue. Among the factors that we also considered, but have not yet included in the analysis are: input fields, VeriSign links, links to social networking sites (apparently no links until after 2008), recommended resolution, the address of the company, site maps, forums and frequently asked questions sections (FAQ). A few of these factors are promising for future investigations.

The speed of retrieving pages and analysing them depends on the available hardware resources and Internet connection speed. Data storage per web site evaluated is about 5MB of data.

### 3.2 Calculating change

The second script calculates change based on the data. There are three methods employed to determine whether a change has occurred.

- A) “first time changes” measure the introduction of technologies, such as CSS, and are only triggered when the first change occurs.
- B) “absolute changes” such as the file extension or the usage of frames and finally
- C) “relative changes” which are triggered when the number of text, links etc have changed more than a specified percentage compared to the previous change. This is important as this type is not triggered based on the previous date, but based on the last change, which allows for more realistic calculations.

Changes of type A and B can be classified as binary, whereas changes of type C as non-binary. For performance reasons, the results of when changes occur are written back into the database, enabling outputting graphs faster. While undertaking preliminary checks to test the reliability of the software, it turned out that no major changes were omitted. To the contrary the software appeared to be over sensitive. This led us to define a large change as a change during which three or more changes happened at once. Further tests showed that only once did the software return a homepage that scored as having three changes or more, but with no noticeable change evident to the human eye. Consequently the 3-change rule was selected as it offered a good balance between validity and reliability. It is worth noting that all changes are considered as equal, although a weighting system has been implemented into the software. Experimenting with the weights of each change may help offer a more realistic measurement of change. Still, one has to keep in mind that automated evaluation is not exact science. Neither is of course a manual one. Between the two though, one can expect the automated one to apply the rules of evaluation consistently without diverting from them, as a human could have potentially done over time. Consequently, for the same subjective assumptions the validity and accuracy of analysis performed by the software should be higher.

### 3.3 Visualising and managing the evaluations

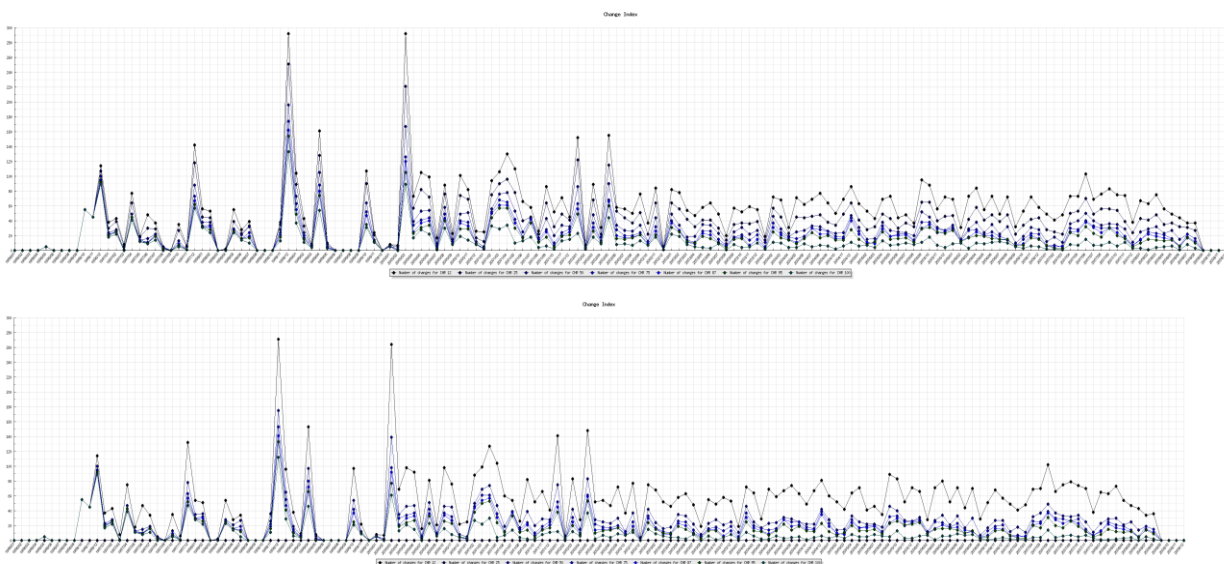
The third script outputs graphs visualising change over time for a given sample. Since this is a relatively computing intensive process, at least if the aim is to produce the graphs dynamically in real time, it was decided to separate the computation of changes and the drawing of graphs. In addition, a number of other scripts enabled the user to perform other

tasks, such as importing web sites into the system or monitoring or stopping a script's execution.

#### 4. Findings: Applying the methodology

Referring back to the software development stage of the project, the graphs in Figure 2 show how the change percentage impacted on what constituted change. Depicted is the sum of change indicators for each point in time for all analysed companies in the test sample. Starting at the top, the lines represent a change rate of 12.5, 25, 50, 75, 87.5, 95 and 100 per cent respectively. The first graph shows that the data trend does not appear to change significantly with the various change percentages. To examine what percentage was most appropriate for the investigation, the lower graph was plotted without the binary change indicators such as file extension, in case the binary changes, unaffected by the change percentage, may have potentially biased the plot. As the project is geared towards investigating larger changes, 12.5 as a percentage change indicator returned too many changes, while the 25 percent change did not result in a satisfactory improvement. The change rates of 50 percent and above converge, for this reason a change indicator of +/- 50% was chosen, as it reliably detects changes and maintains the overall trend while minimising any risk of not detecting large changes.

**Figure 2: Determine the most appropriate change rate**



For clarity sake, the three companies chosen to illustrate the software's analyses are the 3 companies (Time Warner, Walt Disney, News Corp.) of the entertainment industry in the top 100 Fortune 500 list of 2008. Figure 3 shows all the data collected from the Wayback

machine for the companies chosen over the time period that was available. The x axis has the months and corresponding year for the data. Each point on the y axis represents an individual company. As one can see, the data retrieval was highly successful, particularly for the more recent years. The gaps in the graph correspond to periods for which data was not available. As it can be seen this counts for the whole data set up to 2000/03 with only one block of subsequent data from 1998/12 to 1999/05. The data collection is fairly reliable for the rest of the set, with the exception of the period from 2000/10 to 2002/01 for Time Warner and 2002/03 to 2004/12 for News Corp.

**Figure 3: Data available**

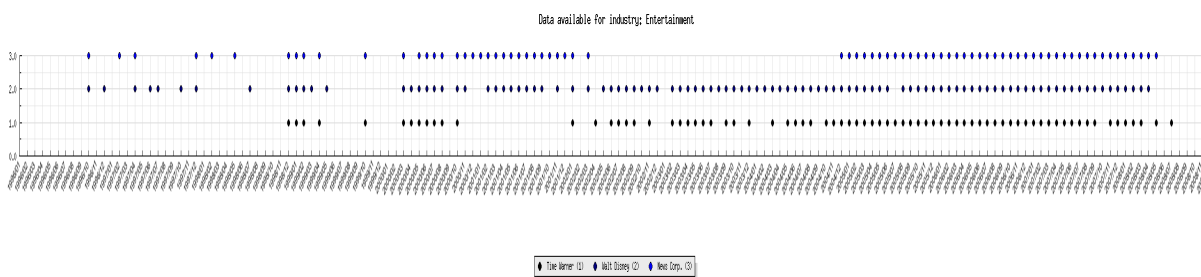
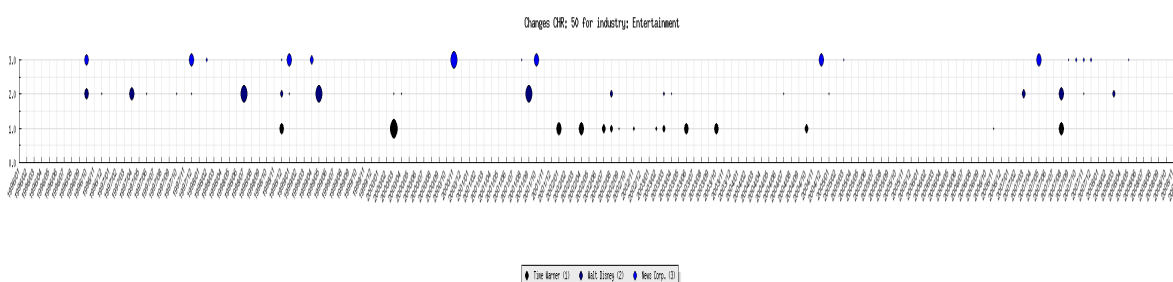


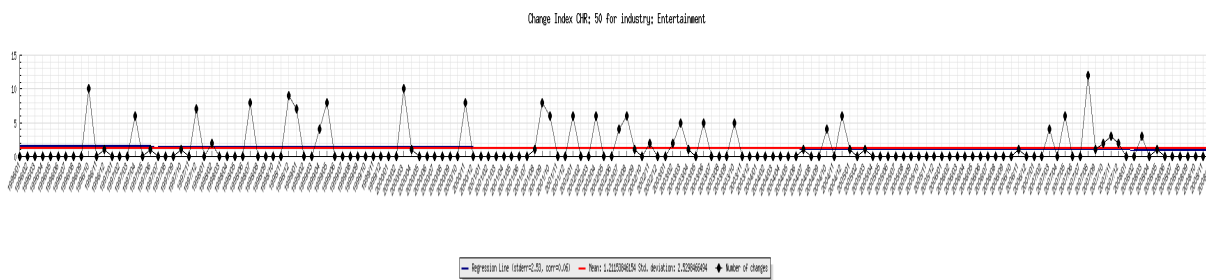
Figure 4 shows a bubble graph, which depicts the changes found. The axes are the same as in the previous graph and each bubble represents a change in the website, a larger circle suggests that more changes have taken place. It is important to note that the size of the circle does not scale with the degree of change. As it has been already mentioned, a larger circle means that more changes have occurred, but as the different indicators of change do not relate, it does not necessarily mean that a greater change has taken place. On first inspection of the bubble graph there appear to be clusters of change, particularly in the months 1998/12, 2001/09 and 2007/08. Another obvious pattern is the rather lasting change in Time Warner's web page form 2002/01 to 2003/10. This could possibly be a result of faulty data delivered by the archive, such as redirecting request to an alternative date, confusing the change detection algorithm.

**Figure 4: Changes**



The change index graph (Figure 5) shows the aggregated total of change indicators for each month. To analyse any potential trends in change over time, a regression analysis was performed. This is represented by the blue horizontal line. The correlation between time and number of changes was identified as 0.06 which indicates no correlation between time and number of changes. The red line represents the mean of the data set, almost overlapping the blue regression line, a further indicator that no correlation exists.

**Figure 5: Change Index**



## 5. Conclusions and further research

This paper presented a methodology for evaluating web site change over time. Our early findings have potentially significant implications for both theory and practice and consequently further work is proposed. Future research could extend the methodology further by adding more items to be measured in the evaluation framework. These could be potentially grouped instead of looking at technologies independently (e.g. web1.0 versus web2.0 technologies). Also, experimenting with the weight factors may offer more realistic results and should be considered in the future. The revised methodology could then be compared to other evaluation methods (e.g. that of Coursaris and Papagiannidis (2009) looking at political web sites), in order to test the relative effectiveness and efficiency. In addition, it could be applied to many different industries and markets, examining the diffusion innovation process for new web technologies among them.

## References

- Artail, H., and Fawaz, K., (2008) 'A fast HTML web page change detection approach based on hashing and reducing the number of similarity computations', *Data & Knowledge Engineering*, 66 (2), pp.326-337
- Bauer, C., and Scharl, A., (2000) 'Quantitative evaluation of web site content and structure', *Internet Research: Electronic Networking Applications and Policy*, 10 (1), pp.31-43

- Calzarossa, MC., and Tessera, D., (2008) 'Characterization of the evolution of a news website'. *Journal of Systems and Software*, 81 (12), pp. 2336-2344
- Casalo, L., Flavian, C. and Guinaliu, M (2008) 'The role of perceived usability, reputation, satisfaction and consumer familiarity on the web site loyalty formation process', *Computers in human behaviour*, 24 (2), pp. 325-345
- Cho, J., and Garcia-Molina, H., (2000) 'The evolution of the web and implications for an incremental crawler'. In *proceedings of the 26<sup>th</sup> international conference on very large databases*
- Coursaris C; Papagiannidis S. (2009) Online political marketing in Greece: An evaluation of the 2007 national elections and two case studies. *Computers in Human Behavior*, 25(4), pp. 853-861.
- Cyr, D., Head, M., and Larios, H. (2010) 'Colour appeal in web site design within and across cultures: A multi-method evaluation', *Human –Computer Studies*, 68 (1-2), pp.1-21
- Fetterly, D., Manasse, M., Najork, M., and Wiener, JL. (2004) 'A large scale study of the evolution of web pages', *Software-Practice and Experience*, 34 (2), pp.213-237
- Koehler, W. (2002) Web page change and persistence – A four year longitudinal study', *Journal of the American Society for Information Science and Technolohg*, 53 (2), pp.162-171
- Olsina, L., and Rossi, G., (2002) 'A quantitative method for quality evaluation of web sites and applications', *IEEE Multimedia*, 9 (4) pp. 20-29
- Taylor, MJ., McWilliam, J., Forsyth, H. and Wade, S. (2002) 'Methodologies and web site development: a survey of practice', *Information and Software Technology*, 44 (6), pp.381-391
- Teo, HH., Oh, LB., Liu, C. and Wei, KK. (2003) 'An empirical study of the effects of interactivity on web user attitude', *International Journal of Human-Computer Studies*, 58 (3), pp.281-305
- Xie, ZC., and Barnes, SJ. (2008) 'Web site quality in the UK airline industry: A longitudinal examination', *Journal of Computer Information Systems*, 49 (2) pp. 50-57
- Zhang, P. and von Dran, GM. (2002) 'User expectations and rankings of quality factors in different web site domains', *International Journal of Electronic Commerce*, 6, (2) pp. 9-33