

Convention says...

*This paper was presented at the **Flexible Hypertext Workshop**
held in conjunction with
The Eighth ACM International Hypertext Conference (Hypertext'97)*

Einat Amitay

Centre for Cognitive Science
einat@cogsci.ed.ac.uk

Jon Oberlander

Human Communication Research Centre
J.Oberlander@ed.ac.uk

The University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
SCOTLAND

abstract

Hypertext is of interest to workers in a broad range of fields, including HCI, educational software design, text generation and information retrieval. Recent studies, such as Chen and Rada (1996), indicate that because there is no single formalised method for designing hypertext, there are significant discrepancies between experiments using hypertext systems. In this article we claim that hypertext design - both manual and automatic - can reliably be informed by studying the robust conventions that have already established themselves on the World Wide Web. We present here a few basic conventions in hypertext design, extracted from a corpus of HTML files.

Introduction

For more than a decade there has been a growing interest in the study of hypertext. Since the mid-eighties there have been hundreds of articles and studies published involving the theory and application of this multi-dimensional electronic text. Some of the articles published study the cognitive aspects of hypertext, others examine its practical aspects. However since there are so many hypertext systems, designs and theories, it's not easy to see the wood for the trees.

Chen and Rada (1996) compare 23 published experimental studies and doctoral dissertations on hypertext, which appeared between 1988 and 1993. They compare users, tasks and tools, finding many discrepancies between results. In their discussion they point out that "*...experimental design vary as the hypertext systems vary. Some systems have only the most primitive hypertext facilities implemented. For instance a page of a document is displayed on the screen at a time and hypertext links are provided at a page-to-page level. Some systems are advanced and include graphical browsers, a persistent window to display the current focal point, footprints, and search facilities.*" They note also that users' characteristics affect the results and that these characteristics should be addressed in future research. The authors conclude by inviting researchers to formulate precedents for hypertext experiment designs. Since the studies considered were published, there has been no substantial change in experimental design. Considering, for instance, McDonald and Stevenson (1996; in press), Stanton (1994), Baron, Tague-Sutcliffe, Kinnucan and Carey (1996), and Wenger and Payne (1996), it is clear that little effort has yet been made to equate such experimental conditions. It is true that benchmark measures had already been formulated by Anderson, Berre, Mallison, Poerter and Schneider (1990) but these are not widely used, and are inadequate for recent hypertext systems.

Theory and implementation of hypertext design also suffer from lack of consistency. Theoretical papers such as Recker and Ram (1994), and system design illustrations such as Nanard and Nanard (1995) and Maioli and Vitali (1994) refer to an undefined user, ignoring individual cognitive differences, and variations in knowledge domains. Nygren, Allard and Lind (1995), and Nygren (1996) point out that hypertext design should aim to improve the ability of skilled users to navigate within the systems. Nygren (1996) defines the difference between the design for novice and the design for

skilled users: “A design fit for the untrained user will focus on ease of learning, while a design fit for the expert will focus on efficiency in daily use.” She also claims that the studies done today do not satisfy the demand for better daily user interface design, but aim instead toward the novice user. Since the use of computers is very widespread, and since menus and hypertext are now an integral part of any much simple software, we would like to focus on this audience of skilled users.

We adopt an approach analogous to that suggested by Suchman (1987), in the context of planning: “*common-sense notions of planning are not inadequate versions of scientific models of actions, but rather are resources for people’s practical deliberations about action ... Rather than attempting to abstract action away from its circumstances and represent it as a rational plan, the approach is to study how people use their circumstances to achieve intelligent action. Rather than build a theory of action out of a theory of plans, the aim is to investigate how people produce and find evidence for plans in the course of situated action.*” (pp 49-50). By analogy, we can usefully look at the way in which hypertext designs are accomplished on a day-to-day basis, and extract generalisations from this study. Such a naturalistic, descriptive approach presents a viable alternative to more rationalistic, prescriptive views on hypertext design. Arguably, there are already robust conventions regarding the use and production of hypertext, followed by millions of people every day, embodied in the World Wide Web. The WWW allows people to create their own sites and to join the virtual community. Over the years people have evolved certain conventions regarding this means of communication. Consumption of existing hypertext designs has been translated into production of further designs. The chicken and the egg problem - whether conventions are caused by, or give rise to, the existing software - is not of our concern. We simply want to assert that there are already regularities and patterns within the WWW which are the fruit of a simple language evolution rule: maximum communication with minimal effort.

We suggest here a new method for informing hypertext design. In this article we would like to begin with the smallest unit: the anchors themselves. In many designs of hypertext systems attention is directed to the spatial linking. Even studies such as McDonald and Stevenson (1996; in press), and Wright (1993), which focus on the text within a node do not relate to the content of the actual links. Their conclusions relate to the structure of the document and to its form, but never to its content, or to the content of the referring anchors (but see Baron, Tague-Sutcliffe, Kinnucan, and Carey (1996) for a suggested classification of anchor content). Many of the studies actually put the blame for the “lost in hyperspace” phenomenon on the referential space between documents, and not on the labelling of the links within the document. Since one of our goals is to study whether this assumption is true, we decided to examine, amongst other patterns, the regularity of anchors within hypertext documents. In order to use the endless resources offered by the Web we simply retrieved and extracted information from it. This was done by creating a corpus of HTML files and analysing the data. The data we present here is only the tip of the iceberg.

Gathering the corpus data

The Parker Corpus comprises 845 HTML files. In order to gather the files randomly, URLs were retrieved from the HotBot search engine by looking for the word “Parker”. The URLs were then fed into a program that “fetched” each file and placed them under one directory. The corpus was “cleaned” of error messages and, as much as it was possible, of HTML frames, which use different files for code.

Two methods were used to retrieve the anchors from the files. One method was a simple UNIX command (`grep -i '</a' *.html | sed 's;</[aA]>.*;;' | sed 's;.*>; ;')` which created a problem because of the nature of the commands and the way people write HTML files: UNIX commands related to one line only, while markups for links were spread over two or even three lines. The other method was to use a Perl program in order to avoid the line problem. Neither method was perfect, but while the UNIX command gave less information than there actually was, the Perl program gave more information than was needed (such as font types etc.). The anchors retrieved with the Perl program were then cleaned by hand taking out HTML markers and extracting text from textual images.

The HTML files consist of 2,294,569 words, which are stored in slightly more than 18 Mbytes. From this corpus of HTML files we extracted a smaller corpus of anchors. This anchors corpus consists of 1.7 Mbytes, representing 133,341 words. We assumed that using a name, such as “Parker”, to retrieve the URLs would bias the results, since we expected most files to be personal Homepages of some sort. But as the information accumulated we realised that this was not the case. Since we did not aim at any specific domain, the results presented here are as general as possible. We intend to gather more specific information in the future, and to try and formulate domain specific patterns. Such patterns could be the use of indexes and table of content, the insertion of content anchors (as oppose to structure anchors) within documents, and the use of domain-dependent vocabulary. The results described here outline general patterns and tendencies in HTML document design, maintained by Web users all around the world.

Results from the corpus data

One of the most sought-after questions, presented in various mailing lists, is how many outgoing links (as opposed to internal links) there are in an average Web document. We would like to leave this question open for later discussion, since the definition of outgoing links and internal links depends on screen size, font size, and page design. Any given link might imply a physical leap from one document to another, or from one chunk of text at the beginning of the document to the acknowledgements at the end of the same document. For this reason we include all links, whether outgoing or internal, in our analysis, making no distinction between their potentially different destinations.

The first question we addressed was - how many non-verbal icons are used in the average Web document. The answer to this question is quite surprising, only 2.6% of the retrieved anchors were actually textless. The definition of textless is that there are no printed characters on the icon. This finding is very interesting because it demonstrates that the convention precedes the science: In a recent study by King, Boling, Anelli, Bray, Cardenas and Frick (1996) it is concluded that *"...buttons with both pictorial symbols and text labels resulted in significantly less user confusion than did buttons with pictorial symbols only. Buttons with text labels only also produced significantly less confusion, compared to those with pictorial symbols only"*. It is unlikely that any of the authors of the 845 HTML files we have gathered has ever read an article related to hypertext design. We suggest that the use of "ISMAP", which is a clickable map of textual and pictorial icons, is becoming very common. It is possible that this combination of text and picture will increase in use at the expense of textless icons. One of the reasons for the low percentage of the use of images as links might be the loading speed. Since loading documents with many images slows their retrieval time, the authors of Web documents usually try to avoid this phenomenon by inserting fewer images to their creations. Another aspect of the loading speed is that if links are to be presented as images they might appear after the text, or in very busy hours of the day it may be the case that the images will not appear at all. Users are apparently well aware of the problem and the consequence is that fewer images are being used as anchors.

The next question addressed was how many links there are in an average document. At first we calculated the overall number of links, including textless icons. The number we have computed seems to be quite high: 36.87 links. Looking for deviations from this number we have found within our corpus one file which consists of more than 2000 links! It is interesting to note that these links are not presented as a list of items but as embedded anchors within the text. There are many documents with more than 200 links, and of course, there are some documents with no links at all. Calculating the average number of verbal-links per page yielded the result of 35.9 links per page. A verbal-link is usually marked in the HTML language between two angled brackets. When the text appears inside an icon it is usually specified in the markup as `ALT="text"`. We extracted this information manually from the corpus. From now on only verbal-links will be analysed, in order to have a better understanding of the language convention with which users create their anchors.

The number of words per anchor might be one of the most important findings of this study. One of the major distractions in hypertext designs can be the length of the phrase chosen as anchor. We found that the average stands at 2.89 words per anchor, and this time there were almost no large deviations from the average. People tend to make the message short, and more important, they tend to distinguish it from other texts or anchors by using nouns, proper names, and adjectives, omitting determiners, prepositions and apostrophe 's'.

The use of determiners is quite consistent: the definite article is used if the anchor refers to a proper name such as a title of a book, say "The Wind in the Willows". Determiners are omitted from the anchor when they are not part of the proper name, or when they might distract the reader from the core idea. For example:

- Go to the `Sample Chapter` (file 537)
- For new material- check out the `Library Fall Programs` (file 127)
- Next: The Titans ride the `Highway to Hell` (file 230)
- Advertising on these pages is subject to the `FederalFair Housing Act.
` (file 624)

Verb usage is also intriguing and we have noticed the tendency to use imperatives and present participle morphology such as 'climbing', 'jogging', 'reading', 'downloading', etc. These forms appear very often in vertical lists of links, apparently to shorten the message and to make it as compact as possible. For example:

- `Sailing on the Internet` (file 613)
- `Historical Exploring` (file 104)
- `Backpacking` (file 724)

We intend to study the syntactic and semantic structure of anchors in the future. We would like to investigate this issue in the light of the Baron, Tague-Sutcliffe, Kinnucan, and Carey (1996), and Trigg (1983) models, which try to define a taxonomy of link types.

Conclusions

There are many interesting trends we can retrieve from the Parker Corpus. We have chosen to present here only some very basic ones in order to demonstrate how even such simple analysis of data can solve design problems and justify decisions. In the future we would like to analyse the discourse markers people use within Web pages, and the language conventions they follow in choosing words for anchors. Results are applicable in at least three areas. First, in dynamic hypertext generation, systems should produce anchors that resemble naturally occurring anchors; version 1.1 of our ILEX system, for instance, produced anchors that are both unnaturally long, presented in a repetitive phrase format and contain determiners, and this behaviour will be modified in subsequent versions. Secondly, text-to-hypertext generation systems could especially benefit from applying these conventions, since they try to create human-like linking. Finally, it is very unlikely that all hypertext designs would follow the conventions presented here, but we suggest that a certain measure of unity should be applied. Thus, designers of hypertext systems should bare in mind that there are already robust conventions regarding hypertext design, followed by millions of people, all around the world.

References

- Anderson T., Berre A., Mallison M., Porter H., and Schneider B. (1990). The hypermodal benchmark. in F. Bancilhon, C. Thanos, and D. Tschritzis (eds.), **Advances in database technology-EDBT '90**. London: Springer-Verlag.
- Baron L., Tague-Sutcliffe J., Kinnucan M. T., and Carey T. (1996). *Labelled, Typed Links as Cues when Reading Hypertext Documents*. Journal of the American Society for Information Science. 47(12):896-908.
- Chen C., and Rada R. (1996). *Interacting With Hypertext: A Meta-Analysis of Experimental Studies*. Human-Computer Interaction. 11(2):125-156.
- King K. S., Boling E., Anelli J., Bray M., Cardenas D., and Frick T. (1996). Relative Perceptibility of HyperCard Buttons Using Pictorial Symbols and Text Labels. Journal of Educational Computing Research. 14(1):67-81.
- Maioli C., and Vitali F. (1993). Anchors and Paths in Hypertext Publishing System. Technical Report UBLCS-93-3.
- McDonald S., Stevenson R. J.(1996). *Disorientation in hypertext: the effects of three text structures on navigation performance*. Applied Ergonomics. 27(1):61-68.
- McDonald S., Stevenson R. J.(in press). Hypertext, navigation and cognitive maps: the effects of a map and a contents list on navigation performance as a function of prior knowledge. in D. Harris (ed.) **Engineering Psychology and Cognitive Ergonomics: Interaction of theory and application** (in press). Avebury Technical.
- Nanard J., and Nanard M. (1995). *Adding macroscopic semantics to anchors in knowledge-based hypertext*. International Journal of Human-Computer Studies. 43:363-382.
- Nygren E. (1996). From Paper to Computer Screen - *Human Information-Processing and User Interface Design*. CMD, Uppsala University. (<http://delfi.cmd.uu.se/papers/diss188/summary.html>)
- Nygren E., Allard A., Lind M. (1995). Effects of figural patterns on trend assessment. Report no. 56, CMD, Uppsala University.
- Recker M., and Ram A.(1994). Cognitive Media Types as Indices for Hypermedia Learning Environments. **Proceedings of the AAAI-94 Workshop on Indexing and Reuse in Multimedia Systems**.
- Stanton N., A. (1994). *Explorations into Hypertext: Spatial Metaphor Considered Harmful*. Education & Training Technology International. 31(4):276-294.
- Suchman L., A. (1987). **Plans and Situated Actions: The problem of human machine communication**. Cambridge University Press: GB.
- Trigg R. H. (1983). A network-based approach to text handling for the online scientific community. PhD Thesis, University of Maryland. Also technical report TR-1346.
- Wenger M., J., and Payne D., G.(1996). *Comprehension and retention of nonlinear text: Considerations of working memory and material-appropriate processing*. American Journal of Psychology. 109(1):93-130.
- Wright P. (1993). To Jump or Not to Jump: Strategy Selection While Reading Electronic Texts. in C. McKnight, A. Dillon, and J. Richardson (ed.), **Hypertext: A Psychological Perspective**. (1993). Ellis Horwood LTD. Chichester.