Web Page Design: Implications of Memory, Structure and Scent for Information Retrieval

Kevin Larson and Mary Czerwinski
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
+1-425-936-8259, +1-425-703-4882
a-kevlar@microsoft.com, marycz@microsoft.com

ABSTRACT
Much is known about depth and breadth tradeoff issues in graphical user interface menu design. We describe an experiment to see if large breadth and decreased depth is preferable, both subjectively and via performance data, while attempting to design for optimal scent throughout different structures of a website. A study is reported which modified previous procedures for investigating depth/breadth tradeoffs in content design for the web. Results showed that, while increased depth did harm search performance on the web, a medium condition of depth and breadth outperformed the broadest, shallow web structure overall.

Keywords
Information retrieval, memory, depth and breadth in information design, web page design, hypertext, searching.

INTRODUCTION
The goal of this study was to discover principles for the design of multiple hyperlinks on a web page for information retrieval tasks. Of particular interest was the optimal depth versus breadth of the hyperlinks' distribution across expertly categorized web content, with an emphasis on the importance of structure. In addtion, an effort was made to tie the findings both to current research in information retrieval and web design for large information spaces. The differential effects of short-term memory and visual scanning were examined as cognitive covariates in the experiment.

Optimal Number of Menu Items
There has been a vast amount of work exploring the optimal number of items in a menu design [8, 2, 15, 6, 5]. Most studies concluded that breadth was better than depth when it came to organizing menu contents, although the breadth of the menus examined has varied from study to study.

D. Miller [8] tested four structures with 64 bottom level nodes: 2^6 (six levels of depth each with two items of breadth), 4^3 (three levels of depth each with four items of breadth), 8^2 (two levels of depth each with eight items of breadth), & 64^1 (64 top-level items). The 8^2 condition allowed the fastest acquisition and fewest errors of the four structures. D. Miller suggested that depth of a hierarchy should be minimized, but not at the expense of display crowding. He also mentioned that the level of breadth that tested well (8^2), fit well within the range of G. Miller's [9] 7+/-2 finding on the limits of short-term memory.

Snowberry, Parkinson & Sisson [15] replicated and extended D. Miller's [8] study by examining the same for depth/breadth tradeoff conditions (2^6, 4^3, 8^2, and 64^1). They included an initial screening session during which subjects were administered memory span and visual scanning tests in an effort to tease out their contributions in subjects' performance data. They found that memory span was not predictive of performance in any of the conditions, but that visual scanning was predictive of performance, especially in the deepest hierarchies.

Kiger [6] tested five structures with 64 bottom level nodes (2^6, 4^3, 8^2, 16x4, 4x16) and collected both performance and preference data. The 4x16 structure (four top level items each containing 16 items) had the fastest reaction times, followed closely, and not reliably different from 16x4 and 8^2. The 4x16 structure also had the fewest errors, followed closely, and again not reliably different from the 8^2 and 16x4 conditions. Subjectively, subjects favored the 8^2 structure when asked about both ease of use and preference. For both ease of use and preference, the 4^3 and 4x16 conditions followed behind by a non-reliable difference.

Jacko & Salvendy [5] tested six structures varying both depth and breadth without controlling for the size of lowest-level search area. The structures (2^2, 2^3, 2^6, 8^2, 8^3, and $8^{6)}$ were measured for reaction time, error rates, and subjective preference. Jacko & Salvendy found reliable differences in reaction time for depth, breadth, subjects, and the depth by breadth interaction. There were reliable differences in accuracy and perceived complexity only for depth. Relating these findings back to complexity theory [1], they concluded that as you increase breadth and/or depth, reaction time, error rates, and perceived complexity will all increase. The cognitive substrate governing this

CHI 98 Los Angeles CA USA
Copyright 1998 0-89791-975-0/98/4..$5.00

complexity was assumed to be short-term memory (although neither short-term memory nor attentional contributions were uniquely factored out in this study). This stands in contrast to Snowberry et al.'s conclusions that found visual scanning to be more predictive of performance, especially in deep structures.

The Importance of Category Structure

D. Miller [8] and Kiger [6] created categories that had identical bottom level nodes in each of their structures. While on the surface this was a sound experimental control, it had the side effect of creating somewhat unwieldy category structures. For instance, "Science News: Psychology", "Science News: Biology", "Science News: Physics", and "Science News: Chemistry" all existed together on a level of one structure where the four sciences were all under the category "Science News" in another structure. In effect the same category structure existed in both hierarchy levels, but counted as one level of depth in the first instance, and two levels of depth in the second instance. Note that this experimental control could be related to the theoretical notion of "scent" [12], where scent has been described as conveying distal target information via category labeling. According to Information Foraging Theory [13], scent is the amount of remote indication a user can derive from an information structure's design and labeling about the relative location of a target. Therefore, in these studies, some structures may have performed better than others because they had stronger scent for the target at the top levels of the hierarchy (i.e., their category labels were more distinctive at the top level). Couched in these theoretical terms, the views of the information structures were examined, but not their navigability [4]. Snowberry, Parkinson, and Sisson [15] showed the strong advantage for structurally grouping like objects. As mentioned earlier they re-ran the D. Miller [8] study, but they did so with one additional structure. In addition to the 2^6, 4^3, & 8^2 structures, they ran 2 forms of the 64^1 structure. One form of the 64^1 structure was similar to D. Miller's, with items randomly grouped on screen. The other 64^1 structure included items grouped into coherent categories. Snowberry et al. found that when categorical grouping was utilized, there was an advantage for the broad 64^1 structure in both accuracy and speed, and these advantages were not eliminated with practice. This was not true for the randomly organized 64^1 structure.

Landauer & Nachbar [7] presented a predictive equation for amount of time to make a decision given X number of extremely grouped alternatives. They created 4 different structures (2^{12}, 4^6, 8^4, and 16^3) with 4096 bottom level nodes each. The bottom level nodes were the numbers 1 through 4096, and at higher levels the numbers were grouped into equal categories, such as 1-2048 and 2049-4096 at the top level of the structure with 2 items of breadth. Landauer and Nachbar found that the reaction time for any given screen would increase linearly with number of alternatives, but that total search time decreased with higher numbers of alternatives per screen because there were fewer levels of depth. They also found the same results with 4096 alphabetized words, and came out in favor of designing broad, shallow displays based on these results.

Training

An important issue in both D. Miller's and Snowberry et al.'s studies was that subjects were given the hierarchies to study for rather long periods of time, and given extensive feedback during testing when correct or incorrect categorical choices had been made during searches. This could have contributed to subjects' relying much more heavily upon a learned structure than they normally would or could when traversing the World Wide Web. Since most of the studies reviewed showed significant effects of training on performance, these studies were clearly examining best case performance. Given that the World Wide Web is constantly being altered and extended, how generalizable are these findings to web design? Will broad, shallow web pages really provide a performance advantage when searching through unfamiliar or loosely related constructs?

The Magic Number 7

Four decades ago, G. Miller [9] offered a general rule of thumb that the span of immediate memory is about 7 +/- 2 items. When people are asked to distinguish between different tones, if the number of tones presented is over about 5, their accuracy at this task decreases rapidly. When asked to recall a series of unrelated words or numbers, people fail when the size of the series increases to 6 or 7. In other words, the span of immediate memory imposes severe limitations on the number of items we are able to receive, process and remember. Although there has been much controversy over the "magic number 7", as it is often referred to, G. Miller recognized that by organizing items into categorical units or chunks, we can at least stretch an apparent short-term memory bottleneck.

A commonality with the D. Miller, Kiger, and Jacko & Salvendy papers is that the breadth of the structures with superior performance falls in the range of 7 +/- 2. Kiger [6] says, "...the data seem to indicate both preference and performance advantage for broad, shallow trees. Interestingly, the tree structure resulting in best user performance used a menu breadth that falls within G.A. Miller's [9] 'seven plus or minus two' estimation of short-term memory capacity... As a general principle, the depth of a tree structure should be minimized by providing broad menus of up to eight or nine items each." In essence, G. Miller's findings that people are only able to make quick, accurate decisions with a small handful of objects at a time has had wide support across studies, and may provide useful guidance in the design of web hyperlinks across pages.

Breadth on the Web

While limiting the breadth of items in menu design has been standard practice, the same is certainly not true of

information design on the web, or with some of the newer information visualization techniques. There are numerous examples of structures with enormous breadth. One need not travel far to see structures with breadths wider than what were covered in the studies mentioned above. For example:

http://www.cnet.com/ – technology magazine
http://www.yahoo.com/ – indexed content
http://www.cnn.com – newspaper
http://www.lycos.com – search engine
http://www.cs.uh.edu/~clifton/index.html – encyclopedia
http://www.slate.com/ – political magazine

Zaphiris & Mtei [17] examined the depth/breadth tradeoff on the web. In their study they attempted to replicate Kiger's structures, but using web hyperlinks. They found that of the 5 structures tested (2^6, 4^3, 8^2, 16x4, 4x16) the 8^2 structure was the fastest to search, followed by the 4x16 structure (with non-reliable differences). Subjects ranked the structures in order of ease of use from easiest to hardest: 16x4, 4x16, 4^3, 8^2, and 2^6, though there were no reliable differences among the top 4. Zaphiris & Mtei state, "...overall our results are in agreement with those of Kiger [6] where it has been proven that access time is proportional to depth in menu selection".

While there is overwhelming evidence that structures with a breadth less than G. Miller's magical number are not optimal (when depth is high), there is less evidence that structures with greater breadth are to be recommended (if depth is reasonable). Snowberry et al. and Landauer and Nachbar provided the first evidence in the literature that suggests a broad, shallow structure that does not fall within Miller's magic number of seven might be optimal for menu design. It remains unclear which of these design principles (memory-constrained or depth-constrained design) holds more weight, and how tightly they are coupled to the category structure of the information space.

In addition, there has been the aforementioned movement in the information retrieval literature to provide the user interface artifacts of "scent" [12] optimally for the end user in web design. According to these theoretical perspectives, the design challenge is to distribute "scent" optimally throughout a well-partitioned information structure. We describe an experiment to see if large breadth and decreased depth is preferable, both subjectively and via performance data, while attempting to design for optimal scent throughout different structures of a website. In the discussion section, we will attempt to link our findings to the theoretical notions of scent and any corresponding design issues.

METHODS

Subjects

19 subjects were taken from the Microsoft database of people who identify themselves as willing participants in computer related studies. The subjects were all experienced computer and web users: users who had used windows computers for at least two years, had used the web for at least one year, and were using the web at least twice per week. Subjects were rewarded with Microsoft software for participating in the study.

Materials

The visual scanning and memory span pre-tests were chosen from the Kit of Factor-Referenced Cognitive Tests [3]. The Memory Span pre-test chosen was the Auditory Number Span Test (MS-1); the visual scanning pre-test presented to subjects was the Finding A's Test (P-1) from the Perceptual Speed sub-tests. Due to time constraints, only one fifth of each sub-test was included in this study. The visual scanning sub-test took 30 seconds to complete, while the memory span test was verbally presented to subjects, at the rate of one digit per second. The test administration took approximately 3 minutes.

Both the content and the categorization scheme that subjects searched were pulled from the Encarta® encyclopedia. There were three different categorization structures, each with 512 bottom level nodes. The three different structures were 8x8x8 (8 top-level categories, each with 8 sub-levels, and 8 content level categories under each sub-level), 16x32 (16 top-level categories, each with 32 content level categories), and 32x16 (32 top-level categories, each with 16 content level categories). One problem with previous studies on this topic is that category "soundness" or naturalness was confounded across depth and breadth conditions. In order to create natural categories, therefore, it was not optimal to use identical content level categories for each structure. The emphasis was to create category labels that would be sensible to users, or to maintain good scent throughout the structures. Instead of choosing 512 bottom level nodes, then fitting different structures to those items, three sensible structures were created and populated with items that naturally belonged to the structures. Because of this, only a quarter of the total 512 items in each structure (128 items) appeared in all three structures. These 128 overlapping items therefore became the set of possible search targets for the study. A nagging problem for us was that these structures were not user tested to control for category soundness, which would have been optimal. Instead, due to time constraints, we invited an editor to establish the category contents for each structure, which resulted in categories that appeared natural to us. This therefore resulted in at least an initial effort to tease apart the effects of structure, scent and category soundness. A portion of each of the three category structures is included in Appendix A.

Since the same 128 targets appeared in each of the hierarchies, subjects' semantic knowledge of the target as well as the target word's length and frequency was controlled. Because items outside of the target search set varied from hierarchy to hierarchy, subject's semantic knowledge, as well as word frequency and length, were not controlled among those non-target items. Instead, we

relied on the fact that the Encarta encyclopedia has gone through five generations of content refinement to create understandable category structures. As stated above, the items used in each structure were picked from Encarta by an editor with the instructions to pick items that are representative of each category.

Each web page was marked with title information indicating where this page was located in the hierarchy. The top page of each hierarchy was marked either 'hierarchy 1:', 'hierarchy 2:', or 'hierarchy 3:'. Second level pages were marked 'hierarchy X: appropriate page title:' where X was replaced with the appropriate hierarchy number and the appropriate category title was filled in. On second level pages 'hierarchy X' was a link back to the top-level page of the hierarchy. If there were three levels, the third level was titled 'hierarchy X: appropriate sub-level category name: appropriate bottom level category name:' and the appropriate sub-level category name and hierarchy X were both backwards links. Under the category name on each page was a vertical list(s) of all the items in a randomized order. The items were formatted on the web page following conventional web style guidelines for optimal scanning [10] and for more efficient view traversability [4]. If the page had 8 or 16 items, they were arranged in a single column. If the page had 32 items, they were arranged in two columns, so that scrolling was never necessary. Example pages from each hierarchy are included in Appendix B.

Procedures

Each subject performed 8 searches in each structure for a total of 24 searches. The search target was always one of the 128 bottom level nodes that overlapped with all three hierarchies. Target items for each hierarchy were chosen at random from the 128 possible targets with two restrictions: each subject would search for a target only once regardless of hierarchy, and, within a hierarchy, no more than one target was chosen per category. The 24 trials were not blocked by hierarchy, but presented in a random order. Subjects were given both the search target and the hierarchy to search in directly preceding each trial, and were not told about future targets. Each trial's target and hierarchy information was presented to the subject on paper next to the computer so it was constantly available as a reference material, with one target and hierarchy presented per page. At the start of each trial, the subject was asked to turn the page in order to see the new target and hierarchy to be used for the next trial.

Subjects were told the purpose of the experiment ahead of time. They were told that we were interested in determining the optimal number of links on a page and that we were having them perform searches in three different hierarchies to explore this issue. They were asked to perform all searches as quickly as possible while making as few mouse clicks as possible. Subjects were not told about the experimenters' expectations of the results. Subjects were warned that the hierarchies did not contain cross-referenced material, so they may choose a logical pathway to a target and find that an alternative path to the target is not provided there. They were told that when this happens, they should try to find another route to the target.

Data for three kinds of analysis was collected: lostness measures, reaction times, and subjective ratings. Smith [14] defined lostness in hyperspace as distance on the hypotenuse of a right triangle where one side of the triangle is the number of different nodes accessed over the number of total nodes accessed (minus 1), and the other side is number of nodes required to complete the task over the number of different nodes accessed (minus 1). Lostness scores can be helpful in identifying when subjects are effectively "going around in circles". In addition, reaction time was measured. There was a start screen for each trial with links to 'hierarchy 1', 'hierarchy 2', and 'hierarchy 3'. The reaction time measure was initiated when the subject pressed the link to the appropriate structure and stopped when the link for the target item was selected. The web server was located on the subjects' computer, in an effort to control for any problems with differential web download and lag times.After the subjects finished all 24 trials, subjective preference and rank order responses were collected for the three hierarchies. Finally, subjects answered the following five subjective questions on a 5 point Likert scale about each hierarchy: "I liked this structure", "Right when I started I knew what information was available", "It was easy to get where I wanted in this structure", "This structure is easy to use", and "This structure feels familiar". Subjects were allowed to go back and review the hierarchies while answering all but the first subjective question.

RESULTS

Reaction Times

Figure 1 shows that on average, subjects completed search tasks fastest in the 16x32 hierarchy (Avg. RT=36 seconds, SD=16), second fastest in the 32x16 hierarchy (Avg. RT = 46 seconds, SD = 26), and slowest in the 8x8x8 hierarchy (Avg. RT = 58 seconds, SD = 23).

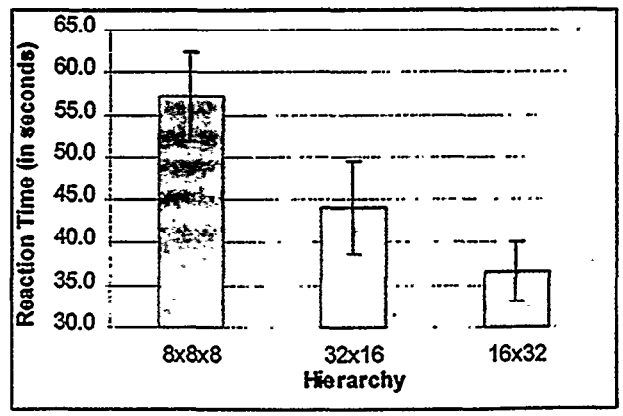


Figure 1: Average reaction time for each hierarchy.

A one-way Analysis of Variance (ANOVA) of the reaction time data revealed a significant main effect of hierarchy [F(2, 36) = 6.34, p<.01]. A post-hoc analysis using the Bonferroni method showed that the 8x8x8 hierarchy was significantly slower than both of the broader, shallower, hierarchies. However, there was no significant difference between the 16x32 and the 32x16 hierarchies for reaction time. No meaningful pattern in the error data was observed in this study, as subjects were for the most part required to continue searching until the target item was found. However, if the subject took longer than five minutes to find a target, the experimenter asked the subject to abandon the search. In the fourteen cases when this happened, the reaction time was recorded as five minutes. The 8x8x8 hierarchy, which had the longest reaction times, had the most time-outs (9 timeouts for 8x8x8, 2 for 16x32 and 3 for 32x16).

Lostness

An analysis of the number of unique and total links visited in comparison to the "optimal" path was performed. Smith [13] defined lostness as any score greater than 0.5 and any score less than 0.4 as not lost. On average, subjects were not lost in the 16x32 hierarchy (Lostness score = 0.38, SD = 0.19), were somewhat lost in the 32x16 hierarchy (Lostness score = 0.49, SD = 0.25), and were very lost in the 8x8x8 hierarchy (Lostness score = 0.63, SD = 0.20).

An ANOVA showed that there was a significant main effect of lostness, F(2,36)=9.22, p<.001. Post-hoc analyses (Bonferroni method) showed that the 8x8x8 hierarchy contributed to significantly more users being lost than did the other two hierarchies. The 32x16 hierarchy was marginally more likely to contribute to lostness than was the 16x32 hierarchy (p=. 067).

Memory and Visual Scanning Ability

The pre-test scores for visual scanning and memory span tasks were correlated with reaction times across the different hierarchies. Although the relationships were non-significant, the memory span scores were slightly more correlated with subjects' reaction time overall, r = -0.27, than were the visual scanning scores, r = -0.12. A negative correlation means that better memory span and visual scanning scores were predictive of faster search times.

Further analyses were carried out checking for whether or not memory or scanning pretest scores were more likely to be predictive of one of the three hierarchies. Table 1 shows the correlation between the cognitive tests and reaction time in each hierarchy. Greater memory and scanning ability was most highly correlated with faster reaction times in the 16x32 hierarchy. Memory span scores had no correlation with reaction time in the 8x8x8 hierarchy, and poorer scanning ability correlated with faster performance in that hierarchy. None of these correlations were statistically reliable.

	8x8x8	*32x16*	*16x32*
Memory	0.02	-0.31	-0.35
Scanning	0.16	-0.18	-0.32

Table 1: Correlations between visual scanning and memory span and reaction time in each of the hierarchies

Subjective Ratings and Preference Measures

Subjects filled out subjective ratings for each hierarchy, as well as a forced-choice preference after completing their searches in all of the hierarchies. In the forced choice preference data, Table 2 shows that the 32x16 hierarchy was ranked as most preferred more often than the other two hierarchies.

	8x8x8	*32x16*	*16x32*
Best	6	11	2
Second Best	2	3	14
Worst	11	5	3

Table 2: Rank ordered votes for the three hierarchies

The average score on a 5 point Likert scale and standard deviation for each subjective questionnaire item are provided in Table 3. Although the 32x16 hierarchy had the highest average ratings overall, there were no significant differences across any of the questionnaire items for the three hierarchies.

	8x8x8	*32x16*	*16x32*
Liked Hierarchy	3.0 (1.5)	3.4 (1.4)	3.1 (1.1)
Availability	3.2 (1.5)	3.4 (1.3)	3.4 (0.8)
Easy to navigate	3.2 (1.5)	3.5 (1.3)	3.1 (1.0)
Easy to use	3.4 (1.6)	3.5 (1.2)	3.1 (1.1)
Familiar	3.4 (1.5)	3.5 (1.1)	3.3 (1.1)

Table 3: Average subjective measures for the three hierarchies on a five point Likert scale (S.D. in parenthesis)

DISCUSSION

The reaction time and lostness data together paint a clear picture that subjects performed best with the 16x32 hierarchy and worst with the 8x8x8 hierarchy. This corroborates previous findings that demonstrated that increasing the levels of depth hurt user performance during search. Both of the hierarchies with two levels of depth resulted in better user performance than did the hierarchy with three levels of depth. But the findings stand in contrast to recent web design and information visualization techniques that herald increased breadth of items to extremely large sizes on the top page of a website. Although not statistically reliable, the hierarchy with 32 top-level items resulted in not only slower search times, on the average, but also more subjects feeling "lost in hyperspace".

Our findings that memory span was a slightly better performance predictor than visual scanning differed from Snowberry et al.'s [15] findings (although these cognitive pretests were not found to be reliably predictive of performance in either study). Snowberry et al. found that

visual scanning was more predictive of performance than memory span. They also found a slightly better correlation between memory span and accuracy in the deepest structure. In fact, detailed error analyses in Snowberry et al.'s [15] study determined that subjects performed less well in the deep structures due to forgetting the target, or because the category labels at the top levels of the deep structures were too general for subjects to remember the correct traversal paths to a target. We found that memory span was more predictive of performance in the two hierarchies with less depth/greater breadth. The differences between the findings in the two studies are likely the product of different methodologies. In Snowberry et al., subjects were required to remember the target item (taxing short-term memory), as well as index the target, retain that index in short-term memory, and map the index to a response. Under these high cognitive load conditions, only people with large memory spans may have correctly retained the target through deep structures. In our experiment, subjects were given paper instructions presenting them with the target item, and this paper remained with the subject throughout the trial. Therefore, in the present study, both memory span and visual scanning may have been taxed more in the large breadth hierarchies, explaining why people with higher scores on these tests exhibited accelerated performance.

As Snowberry et al. found, the subjective ratings of the three hierarchies did not always match the performance data. In a forced choice preference question, most subjects preferred the 32x16 hierarchy over the other two hierarchies (though there was a cluster of five subjects who selected the 8x8x8 hierarchy as their favorite and 32x16 as their least favorite). While the 32x16 hierarchy scored slightly better on average for the five Likert scale questions on appeal, there were no large differences between the three hierarchies.

The magic number 7 +/- 2

At the onset of this study the authors were convinced that short-term memory limitations would play an important role in users' ability to learn and remember the structure of a website, to aid information retrieval. Instead we have shown that memory is only one variable in this debate, at least over the short course of time that users searched through the web pages used in this study. The present results demonstrated that depth, or the number of levels inherent to a web structure, was a stronger determinant of performance, and that three levels of depth resulted in significantly more problems during searching than two, regardless of breadth. So, our findings are consistent with those reviewed earlier that favored breadth over depth, even with our structures that were expertly organized to deliver optimal scent.

We did find stronger correlations between memory span and the two hierarchies with greater breadth. Subjects with better memory abilities were able to perform better in these hierarchies. There was no correlation between memory span and performance for the 8x8x8 hierarchy. Apparently the breadth of the 8x8x8 hierarchy was small enough that it did not tax users' memory, so subjects with better or worse than average memory performed equally well.

Implications for design

This study was an investigation of the effects of memory, response mapping, structure, and scent on designing web sites for efficient information retrieval. We have tried to couch the last twenty years of research on depth and breadth in menu design in a current theoretical framework for information design on the web. The danger of generalizing from earlier research to web design is that there may be a tendency to assume that broader, shallower web site designs are always preferable. The current study has demonstrated that, for one well-organized, large information space, our moderate level of breadth (the 16x32 structure) may actually afford optimal user performance. The results of this study map nicely into the information foraging and effective view navigation lines of research. As has been demonstrated in previous work on that topic, it is extremely difficult to distribute residue, or scent, throughout an information structure effectively [4]. For this study and its expertly organized content, the 16x32 and 32x16 information structures most likely afforded optimal performance because their category labels were more distinct at the top levels (better scent) than those of the 8x8x8 hierarchy. The 8x8x8 structure suffered from the fact that subjects had to make another categorical decision at the second level of the hierarchy. It seems reasonable that the 16x32 hierarchy performed better than the 32x16 hierarchy (though non-reliable in this study) because there were fewer categorical judgments to be made at the top level. As justification for this claim, researchers in cognitive science have long modeled decision making behavior as a more time-intensive cognitive process than simple visual search [16].

In summary, one implication for design based on the current set of results is that web designers need to balance the number of categorical decisions made for their information structure against the number of items needing to be visually searched on the web page. To help designers with understanding this tradeoff, the authors wish to emphasize the need to consider the layout as well as the semantics and labeling of web content. More research on matching category soundness and labeling to a user's understanding of the information space should augment our understanding of how to best design for large-scale information spaces, such as the World Wide Web.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Ben Shniederman, Andrew Sears and Julie Jacko for their guidance in the early stages of this research. Early comments from George Robertson and Sue Dumais were very helpful in refining rough drafts of this paper. We would also like to offer special thanks to Linda Loba for her help with developing the web hierarchies.

REFERENCES

1. Campbell, D. J. (1988). Task complexity: a review and analysis. Academy of Management Review, 13, 40-52.
2. Card, S. (1984). Visual search of computer command menus. In Bouma, H. & Bouwhis, D. (Eds.) *Attention and Performance X, Control of Language Processes*: Hillsdale, N.J.
3. Eckstrom, R. B., French, J.W., Harman, H.H. & Derman, D. (1976). *Kit of factor-referenced cognitive tests.* Educational Testing Service, Princeton, N.J.
4. Furnas, G. W. (1997). Effective view navigation. In Proceedings of *CHI '97 Human Factors in Computing Systems* (Atlanta, GA, April, 1997), ACM press, pp. 367-374.
5. Jacko, J. A. & Salvendy, G. (1996). Hierarchical menu design: Breadth, depth, and task complexity. Perceptual and Motor Skills, 82, 1187-1201.
6. Kiger, J. I. (1984). The depth/breadth tradeoff in the design of menu-driven interfaces. International Journal of Man-Machine Studies, 20, 201-213.
7. Landauer, T. K. & Nachbar, D. W. (1985). Selection from alphabetic and numeric menu trees using a touch screen: Breadth, depth, and width. In Proceedings *of CHI '85 Human Factors in Computing Systems*, ACM press, pp. 73-78.
8. Miller, D. P. (1981). The depth/breadth tradeoff in hierarchical computer menus. Proceedings of the *Human Factors Society*, 296-300.
9. Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.
10. Nygren, E. & Allard, A. (1996). Between the clicks: Skilled users' scanning of web pages. Paper presented at *Designing for the Web: Empirical Studies*, Redmond, WA.
11. Parkinson, S. R., Sisson, N., & Snowberry, K. (1985). Organization of broad computer menu displays. International Journal of Man-Machine Studies, 23, 6, 689-697.
12. Pirolli, P. (1997). Computational models of information scent-following in a very large browsable text collection. In Proceedings *of CHI '97 Human Factors in Computing Systems* (Atlanta, GA, April, 1997), ACM press, pp. 3-10.
13. Pirolli, P. & Card, S. (1995). Information foraging in information access environments. In Proceedings of the Conference on Human Factors in Computing Systems, CHI '95 (pp. 51-58), Denver, CO: ACM Press.
14. Smith, P. A. (1986). Towards a practical measure of hypertext usability. Interacting with Computers, 8, 4, 365-381.
15. Snowberry, K., Parkinson, S. R., & Sisson N. (1983). Computer display menus. Ergonomics, 26, 7, 699-712.
16. Sternberg, S. (1966). High-speed scanning in human memory. Science, 153, 652-654.
17. Zaphiris, P. & Mtei, L. (1997). Depth vs Breadth in the Arrangement Web Links. Available at http://otal.umd.edu/SHORE/bs04/.

Appendix A

Comparison of semantic content in the 3 hierarchies. 64 (of 512) bottom level categories are shown from each structure. The 16 (of 128) items that appear in each of the three hierarchies are bolded.

32x16 Hierarchy. Four of the top 32 levels are shown. Items that match items in other hierarchies are bolded.

Birds	Booby, Cardinal, Condor, Eagle, Emu, Flamingo, Great Horned Owl, Hummingbird, Ibis, Jay, Loon, Meadowlark, Parrot, Puffin, Quail, Swan
Fish	Anchovy, Aquarium, Carp, Eel, Goldfish. Great White Shark, Halibut, Lamprey, Piranha, Porcupine Fish, Salmon, Sea Horse, Stingray, Swordfish, Tropical Fish. Tuna
Invertebrate Animals	Abalone, **Ant, Barnacle, Caterpillar, Cockroach, Earthworm,** Firefly. **Jellyfish,** Krill, Leech, Nautilus, Portuguese Man-of-War, **Scorpion, Snail.** Swallowtail. Symbion
Plants	**Bamboo. Botanical Garden, Cactus, Conifer,** Diseases of Plants. Evergreen, Forest. Heath, Lavender, **Mistletoe, Oak,** Peony, **Plant Propagation.** Rhizome. Wheat. **Yew**

16x32 Hierarchy. Two of the top 16 levels are shown. Items that match items in other hierarchies are bolded.

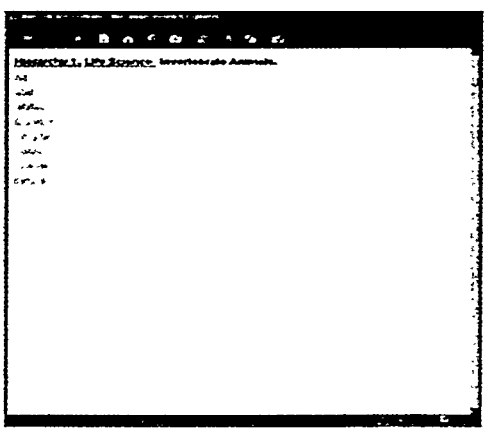
Invertebrate Animals	Abalone. **Ant**, Bark Beetle. **Barnacle**. Bivalve. Butterflies & Moths. **Caterpillar**. Centipede. **Cockroach**. Daddy Longlegs. Dragonfly, **Earthworm**. Firefly. Gastropod. Gnat. Hermit Crab. Invertebrate. **Jellyfish**. Krill, Leech. Mollusk, Moss Animals, Nautilus, Portuguese Man-of-War, Sand Dollar. **Scorpion**. Sea Cucumber, **Snail**. Swallowtail. Symbion, Walkingstick, Water Bug
Plants	African Violet, Alder. Apricot, Aster, **Bamboo**. Banana, Belladonna, Bleeding Heart, **Botanical Garden**, **Cactus**, Cedar, Coffee, **Conifer**, Diseases of Plants, Dragon's Blood, Evergreen, Forest, Grafting, Heath, Indigo Plant, Insectivorous Plants. Lavender, Licorice. Marigold. **Mistletoe**. **Oak**, Onion. Peony, **Plant Propagation**. Rhizome, **Yew**, Wheat

8x8x8 Hierarchy. One of the top 8 levels are shown. All 8 items under the top level are shows, as are all the bottom level items under each. Items that match items in other hierarchies are bolded.

Life Sciences	Algae & Fungi	Bracket Fungi, Fungi, Lichen, Mushroom, Red Algae, Rust, Seaweed, Slime Molds
	Birds	Condor, Eagle, Great Horned Owl, Hummingbird. Jay, Parrot, Puffin, Quail
	Fish	Anchovy, Carp, Eel, Goldfish, Piranha, Salmon, Stingray, Tuna
	Invertebrate Animals	Ant, Barnacle, Caterpillar, Cockroach, Earthworm, Jellyfish, Scorpion, Snail
	Mammals	Bat, Hedgehog, Kangaroo, Porcupine, Reindeer, Sloth, Whale, Wolf
	Plants	Bamboo, Botanical Garden, Cactus, Conifer, Mistletoe, Oak, Plant Propagation, Yew
	Reptiles & Amphibians	Adder, Alligator, Anaconda, Cobra, Gila Monster, Newt, Python, Sea Snake
	Viruses, Monerans, & Protists	Acquired Immune Deficiency Syndrome, Amoeba, Bacteria, Bacteriophage, Flagellates, Protozoa, Retrovirus, Syphilis

Appendix B: Screen Shots.

1) A bottom level page from the 8x8x8 structure. 'Hierarchy 1' is a link back to the top level page, and 'Life Science' is a link back to the life science sub-category.



2) A bottom level page in the 16x32 hierarchy. 'Hierarchy 3' links back to the top level node, and each of the items in the column link to content.