Submitted for publication: 4.10.99

Information Retrieval Patterns from Hypermedia Presentations

Joan C. Nordbotten

Dept of Information Science, University of Bergen,
N-5020 Bergen, Norway
hhtp://www.ifi.uib.no/staff/joan

Abstract

Organizations increasingly use hypermedia presentations to give the general public information traditionally disseminated using pamphlets, videos, or visits to the organization. Unfortunately, little is known about how hypermedia information is retrieved. The following presents a 3 year study of information retrieval patterns from a hypermedia presentation which was initially implemented as a stand-alone museum exhibit and subsequently placed on the Internet. The underlying hypothesis for the study has been that *information retrieval patterns used for stand-alone hypermedia presentations will be similar to those used for web presentations.* If so, then the retrieval patterns observed for stand-alone presentations can be used to develop web presentations. Our data does not support this hypothesis. Information retrieval sessions for the web presentation were significantly shorter and more focused than for the stand-alone system.

Information Retrieval Patterns from Hypermedia Presentations

Public and private organizations are increasingly using web (World Wide Web) presentations for dissemination of information about their organization, products, and/or services. The hypermedia format of web documents is particularly useful for presentation of information about topics or artifacts that are otherwise presented using multiple media, such as images, text, tables, charts, and/or video. Hypermedia presentations can contain a few to millions of pages and are found as stand-alone exhibits, or information centers, in public areas of buildings and as web presentations on the Internet. An example is from the French Ministry of Culture which has started an ambitious project to make more than 22.5 million museum documents available on the web (Mannoni, 1996, 1997).

Hypermedia technology supports associative information retrieval and can facilitate information gathering (Bush 1945, Nelson 1967, Shneiderman 1992). In hypermedia presentations, information is structured as a set of inter-linked pages, often supplemented with orientation aids such as indexes, navigation bars, and/or context maps. While these structures are intended to facilitate access to information, little is known as to whether the information reaches its intended audience and whether the recipients receive the information that they need and/or have requested (Day 1995, Futers 1997). Researchers anticipate a number of problems. As the number of inter-linked documents and path selections increases, *user disorientation* and *cognitive overload* may hinder information gathering (Conklin 1987, Preece 1994). Link structures may actually hinder location of specific information (MacKenzie 1996).

For the general public, gathering information entails retrieval of interesting document sets (Futers 1997). Providing effective support for information gathering from hypermedia presentations requires an understanding of how the intended public retrieves information. Actually, little is known about how web users retrieve hypermedia information. How is the information selected? How much is viewed? How long does the user dwell with an information presentation? Answers to these questions will enable information providers to improve presentations and tailor presentations to different recipient groups.

In the following, we report and compare 4 studies of information retrieval patterns from a hypermedia presentation that has been available to the general public for 3 years in both stand-alone and web implementations. The dominant user group differs in each study, allowing a comparison of information retrieval patterns across both user groups and implementation medium. The underlying hypothesis has been that *users of stand-alone hypermedia presentations would have similar information retrieval behavior to those using web presentations.* If so, studies of stand-alone information retrieval behavior, which are relatively easy to conduct, could be used to develop a framework for the design of hypermedia presentations for Internet users.

Our study does not support this hypothesis. Internet users are significantly more thematically focused, goal directed, than users of the stand-alone system. They select significantly fewer pages and rely more often on direct link access.

General Approach for this Study

We have studied information retrieval patterns for general public users of a small hypermedia presentation. The presentation was initially set up in August, 1996, as a computer-based, stand-alone exhibit in the Museum of Natural Science. The exhibit was part of a university wide anniversary celebration presenting research themes from each of the universities schools. It was moved to the School of Social Science in the late fall of 1997 to be used as an information center. After translation to English, the exhibit was placed on the web in Jan. 1999 where it can be viewed at http://nordbotten.com/museum.

We have analyzed log data from 4 separate periods, each with a different dominant user group. The details of each study period will be discussed separately after a general presentation of the study at large. A discussion of the whole study concludes this paper.

Participants

Since the goal of this project was to study information retrieval patterns of the general public, no recruitment of subjects was preformed. The exhibit has been available to anyone who visited the museum during 1996-1997, the School of Social Science in 1997-1998, or searched the web from 1999. Information searchers and browsers determined if they would activate the exhibit, which topics they would see, the number of exhibit pages they would retrieve, and the length of time they would spend.

Visitors to the natural science museum that housed the exhibits for the university anniversary celebration were the subjects of the 1st and 2nd studies. Our assumption that these visitors would be similar to web users is based on the expectation that both groups contain casual browsers, looking for something interesting, as well as goal oriented seekers of specific information.

Students and visitors to the School of Social Science were the subjects of the 3rd study. We assumed that these exhibit users would find particular interest in an exhibit developed by researchers of the social sciences and thus reflect those web users seeking specific information.

The persons behind the web exhibit visitors are unknown.

The hypermedia presentation

Six topics, developed by researchers in the social sciences, were formed as sets of hypermedia pages and included in the stand-alone exhibit. Each topic presentation was introduced with a general description page containing embedded text and image links to inter-linked detail pages describing particular aspects of the topic. Most topic pages include 2 images with accompanying text, as shown in the example in Figure 1. Each page is thematically self contained and there was no scrolling on the implementation machine. Topic presentations ranged from 2 to 8 pages.

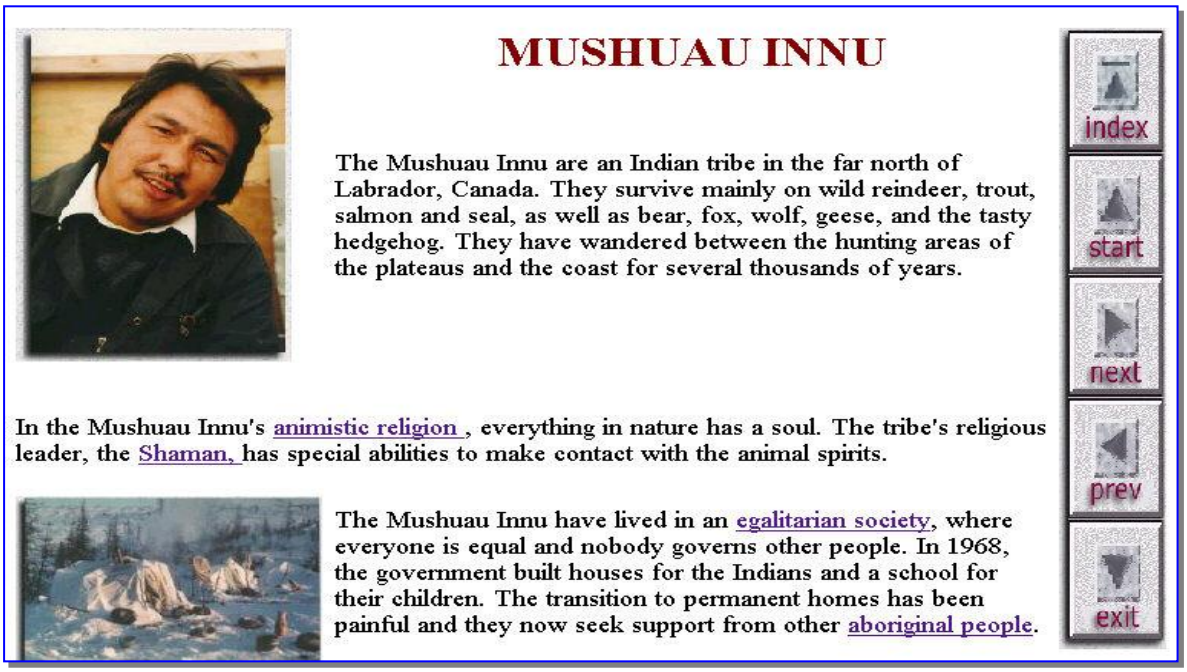

Figure 1. The initial page of the "Mushuau Innu" topic presentation[1].

All topic pages contain a navigation bar with buttons for <exhibit index>, <topic start>, <next topic page>, <previous topic page>, and <exit>. The <exit> button calls a questionnaire, requesting the following data from the viewer: gender, age group (<20, 20-40, and >40), whether the exhibit was difficult to navigate, and whether the exhibit was interesting. In addition to the topic presentations, the exhibit contains a cover page, a 2-level hierarchical index, and an overview index. The later, shown in Figure 2, is accessible from each topic page. A time out, set at 45 seconds, assured that the stand-alone exhibits were restarted at the cover page if a user left the exhibit without using the <exit> button.

Changes for the web presentation include: content translation to English, elimination of the time-out feature, and replacement of the 2-level, hierarchical index by the overview index in order to reduce the path length to topic pages, from 3 to 2 pages.

The exhibit was initially implemented on a stand-alone PC with touch screen input using a WebSite™ server with a Netscape™ browser. The WebSite™ server also administers the web presentation.

[1] Note the alternative navigation aids - the navigation bar, supporting serial navigation, and the embedded hypertext links allowing nonlinear page selection.

Figure 2. Topic content and index layout[2].

Procedure

In the stand-alone exhibit, user sessions were initiated by touching the cover page causing a transition to the theme index, identical to the left 2 columns of Figure 2. Thereafter, pages were selected by touching an active image, text, or button link. The 2-level index structure required 2 index page selections: theme selection followed by topic selection from a topic index, to reach the 1st topic selection, giving a minimum path length to a topic of 3 page selections. Sessions completed on return to the cover page.

Log data were analyzed for 4 periods, in the (a) fall 1996, (b) summer 1997, (c) fall 1998, and (d) spring 1999. The browser cache for the stand-alone exhibits was set to null, ensuring that the server logged each page selection. The log data used for this study includes; the name of the requested page, the date and time for its selection, the name of the calling machine, and the calling/previous exhibit page.

Data preparation for the analysis of information retrieval patterns included separation of sessions, generation of a page transition matrix, and calculation of session length in time and number of pages. System implementer sessions were excluded from the analyses.

[2] The index sequence for this study shown in the figure was the one presented to the 2nd to 4th groups. In the initial index sequence, "Other Cultures" preceded "Business and Trade".

The following definitions have been used for analyses of the stand-alone sessions:

- Session start: transition from the cover page to the theme index
- Session end: transition from any page to the cover page.
- Time out: 45 seconds display time, thereafter return to cover page.
- Session length: sum of pages requested.
- Topic session: a session containing at least 1 topic page.
- Session time: calculated from the first initiation of the theme index to the return to the cover page. The last page time was ignored if it equaled the system reset time under the assumption that the visitor had left the exhibit.

A combination of source machine identification and date/time of the exhibit access identified web sessions.

Study Results

The results reported here are given in Tables 1 and 2 and Figures 3 and 4, placed at the end of the paper. Tables 1 and 2 give the session characteristics and topic selection profiles for each study group, respectively. Note that in Table 1, session length for Internet users (row 4) includes only initial page requests since the user cache cannot be set to null. Session time for the Internet users has not been included in the analyses, since it includes network transport times for page construction that are not applicable for the stand-alone systems.

Characteristics of topic page selection, based on the 3 longest topic presentations, are given in Figures 3 and 4. Figure 3 shows the percentage of user sessions containing the 1st to 4th topic detail pages. Figure 4 shows the percentage of topic detail page selections that were accessed using an embedded text or image link. Note that it is not possible to determine from the log if access to the 1st detail page is through an embedded link or via the <next> button of the navigation bar.

A more detailed presentation and discussion of the results is given under the presentation of each of the 4 study periods. Discussion of their implications is given following the presentation of the 4 individual studies.

1st Study – Youths and School Children[3]

Data for the 1st study period were collected during the fall of 1996, shortly after the exhibit was installed in the natural science museum.

Participants

Museum admission data for the study period showed that 35, 37, and 28% of the visitors were adults, youths with school identification, and school classes from pre-school through 9th grade, respectively. Since no visitor information was available for the exhibit, we have

[3] A more detailed presentation of this study is given in (J.Nordbotten & S.Nordbotten, 1997).

assumed that the 331 users had the same age distribution as the general museum population, i.e. that the majority, 65%, were young people of school/university age[4].

The Exhibit

The sequence of themes in the exhibit index was "Other Cultures", "Business and Trade", and "Public Information Systems". Note that in order to test the high correlation between topic selection and index placement, the index layout was later changed to the sequence shown in Figure 2. The topic sequence within each theme was as shown in Figure 2. The <exit> questionnaire had to be discontinued because of usage problems.

Procedure

Exhibit activation, data collection, and session analyses were performed as described earlier in the "General Approach" section.

Results and Discussion

Session characteristics for the 1st study are given in the 1st row of Table 1. There were 331 session starts during the study period. Of these, 32% contained only index pages, indicating no topic interest. Some of these sessions can be attributed to museum attendants who checked each day that the exhibit was functioning. Others can be attributed to young children who were unable and/or unwilling to read about research topics.

Characteristics for a typical session, were calculated as the average of the 225 session characteristics, though individual sessions varied widely, as follows:

-	Session time:	45 seconds,	ranging from 3 - 348 seconds,
-	Session length	6 pages,	ranging from 3 to 50 pages,
-	Topics per session	1.1,	ranging from 1 to 6 (all), and
-	Topic length	2.5 pages,	ranging from 1 to 8 (all).

The typical session was short for all attributes, indicating only moderate interest in the exhibit content.

Topic selection frequencies (shown in the 3rd column of Table 2) are strongly correlated to topic placement in the indexes, the correlation coefficient is >0.9. That is, topic selection was top-down in the index choosing most frequently the 1st topic in the 1st theme followed by 1st topic in 2nd theme, and so on. Topic interest, measured in number of detail pages selected, fell from 62% to 51%, 41%, and 29% for the 2nd to 4th pages, respectively (shown in the left-most column in each set of Figure 3). Finally, more that 75% of the detail page selections were done using the <next> button (left-most column in each set of Figure 4). These characteristics indicate that these viewers were reading the exhibit in a serial manner.

[4] This was also the impression of the museum attendants who worked in the exhibit room.

2nd Study - Adult Tourists[5]

The 2nd study sought to confirm the strong correlation between topic selection and index placement, and to study the information retrieval characteristics for a 2nd population.

Participants

The 2nd study period was during the summer of 1997, while the exhibit was still located at the natural science museum. Museum admission data confirmed that the dominant visitor groups were adults and families with children. Almost 50% of the visitors were non-Norwegian tourists.

The Exhibit

The initial exhibit index, used for the 1st study, was reordered to check the high correlation between topic selection and index sequence. The resulting sequence is shown in Figure 2. The <exit> questionnaire was still not available.

Procedure

Exhibit activation, data collection, and session analysis were the same as used for the 1st study and as described in the "General Approach" section.

Results and Discussion

Of the 374 session starts, fully 50% were not topic sessions. Most of these sessions can be attributed to the high percentage of foreign tourists who would not have been able to read the exhibit text. Children, without parental help, may also have contributed to some of these sessions.

The typical topic session characteristics for the 2nd study (given in the 2nd row of Table 1) show an increase in interest as measured by:

- Session time: increased to 55 seconds,
- Session length increased to 7 pages,
- Topics per session increased to 1.6 topics,
- Topic length remained the same 2.5 pages.

However, these increases are not significantly different from the topic session characteristics from the 1st study.

Topic selection frequencies (Table 2) are still highly correlated, >0.8, with index placement. Selection of detail pages (Figure 3) fell to under 50% for the 3rd detail page, and embedded link usage was still under 25% (Figure 4). Again, though topic interest was somewhat higher, the difference is not statistically significant.

Though the adult dominant group appeared to be more interested in the exhibit than the youth dominant group of the 1st study, topic sessions were still short in both time and number of pages viewed. Topic selection was still strongly correlated to index placement and navigation was predominantly serial.

[5] A more detailed presentation of this study is given in (J.Nordbotten & S.Nordbotten, 1999).

3rd Study - Social Science Students

Presumably, visitors to a natural science museum would not be expecting an exhibit based on topics from social science research and therefore would not have a special interest in exploring its content. After the exhibit was moved to the School of Social Science, a 3rd study was made to determine if the match between exhibit content and location would be reflected in the information retrieval patterns. In particular, if the high correlation between topic selection and index placement would be retained or if topic content would be a stronger determinant for topic selection.

Participants

The 3rd study period was during the fall of 1998. Exhibit visitors were students and faculty of the social sciences.

The Exhibit

The exhibit index sequence was the same as that for the 2nd study period. The <exit> questionnaire was reinstalled and additional embedded links were added to some of the detail pages. Otherwise the exhibit content was unchanged. Though the <exit> questionnaire was selected in 53% of the sessions, no questionnaires were submitted to the system during the study period.

Procedure

Exhibit activation, data collection, and session analyses were performed as described earlier.

Results and Discussion

Interest in exhibit content, as measured by the percentage of topic sessions, 78% of exhibit starts, the length of time spent, and the number of pages viewed, appears to have increased (Table 1). Topic session characteristics show that:

- Session time: increased to 79 seconds,
- Session length increased to 8 pages,
- Topics per session decreased from the 2nd study to 1.4 topics,
- Topic length increased to 2.8 pages.

However, less than 50% of topic selections contained detail pages (Figure 3). And, topic selection frequencies (Table 2) are still correlated, >0.75, with index placement. One characteristic of significant difference, p=0.000009, is the use of the embedded links for detail page selection (Figure 4).

Though these variations are interesting, an ANOVA analysis showed no statistical support for considering the information retrieval patterns as significantly different.

4th Study – Internet Users

The underlying hypothesis for our study has been that users of stand-alone hypermedia presentations would use similar information retrieval patterns as those used by Internet users of web presentations. Placement of the museum exhibit as a web presentation on the Internet has allowed us to test this hypothesis. The exhibit location was announced through research workshops and conferences, research papers, and as links from the authors' home pages. In addition, 7 Internet crawlers visited the site 385 times during the study period.

Participants

Data for the 4th study were collected during the spring of 1999. The <exit> questionnaire was selected 3 times without resulting in any data submissions. Thus user demographics are unknown, since the exhibit has been open to anyone surfing the Internet.

The Exhibit

The exhibit index sequence was the same used for the 2nd and 3rd studies. The <exit> questionnaire was installed and additional embedded links were added to some of the detail pages. Other than language translation to English, the exhibit content was unchanged.

Procedure

Exhibit access was open in the sense that, in addition to access through the index pages, users could access detail pages directly by using Internet search engines. This necessitated a redefinition of a topic session. Under the assumption that a single page retrieval without follow up indicates an uninteresting 'hit' for the information requester, a *web topic session is defined as one that contains at least 2 pages including at least 1 topic page.*

Since there is no opportunity to set a user browser's cache to null, the log data collected was limited to initial page requests, making full path analysis, including backtracking, impossible. Session length is therefore understated.

Session times were dropped from the inter-study comparisons. Calculation of session time gave uncertain results in that some page changes occurred at very long intervals, >10 minutes, indicating that other activities were happening in parallel with viewing exhibit pages. Further, network transport caused relatively long page construction times compared to the stand-alone presentations.

Results and Discussion

Of the 87 exhibit starts during the study period, 60% started topic sessions as defined above. Only 5 sessions, 10%, began at the exhibit cover page. These sessions were longer than those for the stand-alone exhibits, averaging 2.2 topics and 9 unique pages.

Most sessions, 79%, retrieved their 1st exhibit page from a keyword search using a search engine. 77% of the topic pages started sessions from the search engines. Over 90% of the topic sessions contained only 1 topic. Selection frequency of the detail pages (Figure 3) reflects the number of direct entries. What's notable is that use of the embedded text and image links for further exhibit navigation is under 40%.

General Discussion

Knowing how users retrieve information can help support the design of effective information presentations. The primary objective of this study has been to gather data about how the general public retrieves information from hypermedia presentations, both stand-alone, as used in museums and information centers, as well as web presentations. Summary data from the studies is given in Tables 1 and 2, and Figures 3 and 4.

Sessions for the stand-alone exhibits were begun whenever a transition from the cover page to the 1st theme index page was recorded. For Internet users, sessions were begun with the retrieval of the 1st exhibit page. In our study, 36 % of the 1114 sessions, contained only index pages or only 1 exhibit page for Internet users, indicating that these visitors found nothing of interest in the exhibit. The 714 topic sessions, defined as containing a minimum length of 2 pages, containing at least 1 topic page, have formed the basis for this study of information retrieval patterns.

Log data has been studied from 4 periods, each with different dominant users: school age youths, adult tourists, social science students, and Internet users. The first 3 groups used a stand-alone hypermedia presentation with a hierarchic index structure. Though each of these groups showed more interest in the exhibit than the previous group, measured in time spent and number of pages selected, the differences were not significant. A dominant characteristic of stand-alone information retrieval is that topic selection is strongly correlated, >80, with topic placement within the index structure and, for most viewers, page selection is serial, through use of the <next> navigation button. Further, we found that sessions were short in both number of topics selected, 1.3, pages viewed per topic, <3, and total session time <1.3 minutes. All of which indicates that designers of stand-alone hypermedia presentations for the general public should pay particular attention to topic sequence in the indexes, limit the number of detail pages, and order detail pages by importance to the information presentation.

In our study, Internet users did not use the information retrieval patterns used by stand-alone viewers. Nearly 80% retrieved initial topic pages by using a keyword search through a search engine. Topic selection frequency was significantly different, $p=0.0003$, from the index guided selection of the stand-alone viewers. Also frequency selection of the topic detail pages varied significantly, $p=0.006$, between the 2 groups. This latter is also a consequence of the keyword search, which frequently selected a detail page rather than the initial topic presentation page. Finally, we observed that the topic sessions were very focused, 90% contained only pages relevant to the keyword topic. It appears that designers of hypermedia web presentations should focus of short, self-contained page sets, where any one could be an entry point to the presentation. Indexes are not necessary as entry guides, but can be useful for the interested viewer to gain an overview of the presentation content.

In conclusion, it appears that studying stand-alone exhibits, where knowledge of user demographics is possible to obtain, does not give sufficient information for design of web presentations. Our attempt to solicit demographic information from users of the web exhibit was unsuccessful.

Acknowledgments:

This project was begun as part of the anniversary celebration activities for the University of Bergen and the School of Social Sciences. Thanks are extended to staff, faculty, and students of the Bergen Museum, School of Social Science, and the Department of Information Science for their help in the construction and test of the electronic exhibits. Special thanks are extended to Professor Svein Nordbotten for all project support.

References

Bush, V. (1945). As we may think. Atlantic Monthly July 176:1, 101-108.

Conklin,J. (1987). Hypertext: An introduction and survey. IEEE Computer, 20:1,17-41.

Day,G., (ed.) (1995). Discussion. Proceedings. Museum Collections and the Information Superhighway. Science Museum, London. http://www.nmsi.ac.uk/infosh/discuss.htm.

Futers, K. (1997). Tell Me What You Want, What You Really, Really Want: a look at Internet user needs. Mda. http://www.open.gov.uk/mdocassn/eva_kf.htm.

MacKenzie, D. (1996). Beyond Hypertext: Adaptive Interfaces for Virtual Museums. http://www.dmcsoft.com/tamhpapers//evaf.htm.

Mannoni,B. (1996). Bringing Museums Online. Communications of the ACM, 39(6), 100-105.

Mannoni,B. (1997). A Virtual Museum. Communications of the ACM, 40(9), 61-62. See also http://www.culture.fr/lumiere/documents/files/imaginary_exhibition.html

Nelson, T.H. (1967). Getting it out of our system. In G.Schechter, (ed.) Information Retrieval: A Critical Review. (pp. 191-210). Thompson Books.

Nordbotten, J. & Nordbotten, S. (1997) Information Dissemination using Hypertext Exhibits. Norsk tidsskrift for Bibliotekforskning, Nr.10, 76-90.

Nordbotten, J. & Nordbotten, S. (1999) Search Patterns in Hypertext Exhibits. Proceedings of HICSS 32, Maui, HI, USA, Jan. 4-8, CD, IEEE ISBN 0-7695-0001-3.

Preece, J. et.al. (1994). Human-Computer Interaction. Addison Wesley.

Shneiderman, B. (1992). Designing the User Interface - Strategies for effective Human-Computer Interaction (2nd ed). Addison-Wesley.

Tables and Figures:

Table 1. Exhibit Session Profiles[6]

	Sessions[a]			Session characteristics[b]		
Dominant visitor	Total	Topic (n)	%	No. topics	No. pages	No. seconds
Youth	331	225	68	1.1	6	45
Adult / tourist	374	187	50	1.6	7	55
Social science student	322	250	78	1.4	8	79
Internet user	87	52	60	1.1	4[c]	-[d]
Sums/averages	1114	714	64	1.3	6.3	60

Table 2. Topic Selection Profile by Visitor Group

Themes	**Topics**	**Topic selection frequencies (%)**				
		Youth	**Adult**	**Social science students**	**Internet users**	**All users**
Other cultures	Mushuau Innu	35	28	22	21	27
	Palestinians	14	12	14	7	13
Business and Trade	Banking & Society	22	23	24	13	23
	Maritime safety	8	8	8	4	8
Public information systems	Inca statisticians	18	20	18	30	20
	Modern IT systems	4	10	13	13	9

[6] Note. Data were collected from the system server log of a stand-alone system for the first 3 groups during the fall of 1996, summer of 1997, and fall of 1998, respectively. The 4th group data were collected from the web-server log during the spring of 1999.

[a] Up to 50% of the exhibit sessions contained only index pages. These have been omitted from further selection path analysis. Thus n has been chosen for each group as those sessions in which at least one topic was selected.

[b] Session characteristics give the averages for sessions with at least one topic selection.

[c] Includes initial page selection only. Backtracking occurs at the user site without server logging.

[d] Session time for Internet users has not been included for the inter-session analysis, since it includes net transport during page construction, transmission restarts, and parallel activities which make these times incompatible with session times for stand-alone systems.

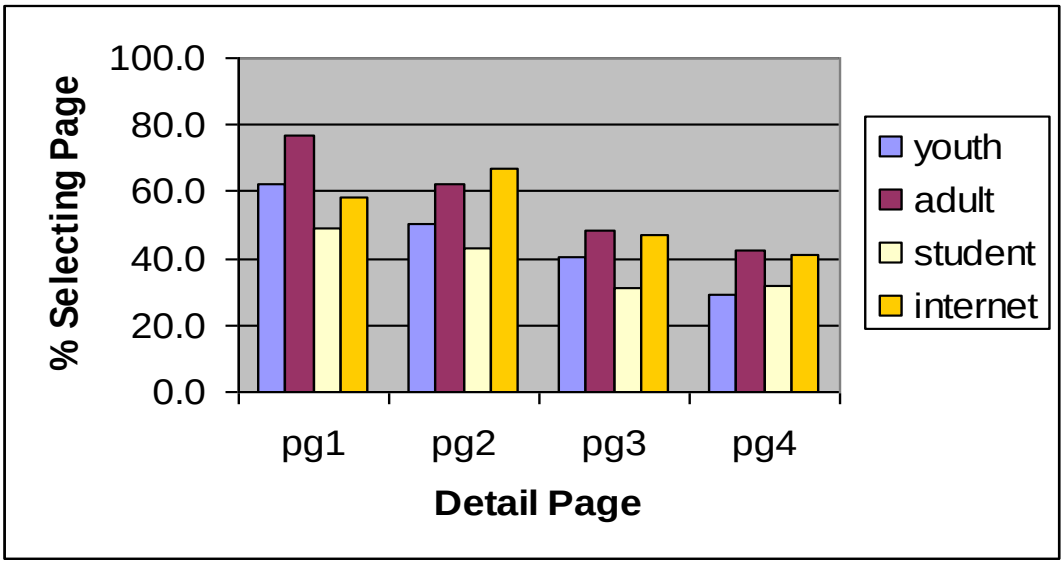

Figure 3. Percentage of topics containing each detail page by user group[7].

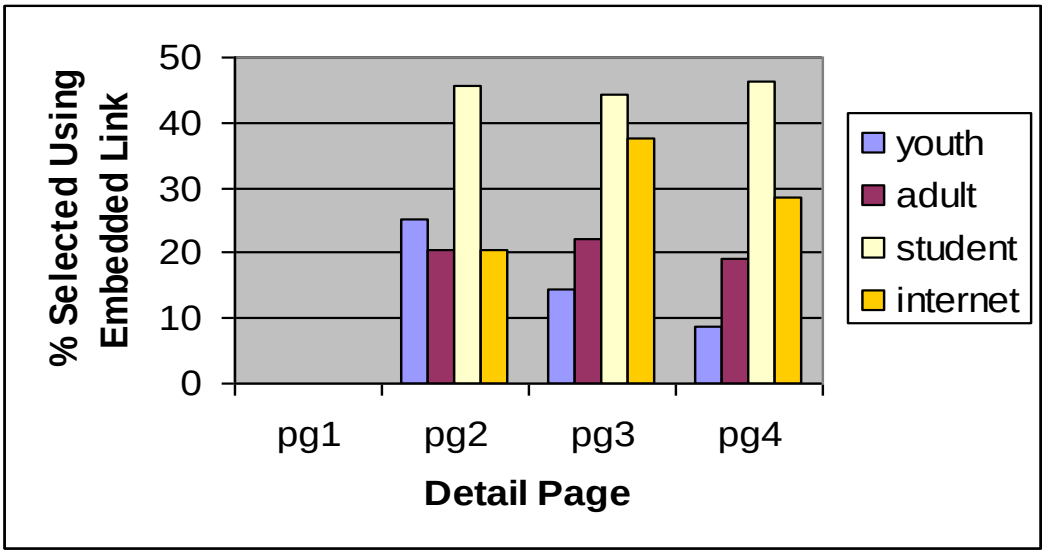

Figure 4. Percentage of detail pages accessed using embedded links by user group[8].

[7] Note that this chart is based on an analysis of the 3 topics containing at least 5 hypermedia pages.
[8] Note that it is not possible to determine from the log data if the 1st detail page has been retrieved using the text link or the <next> button in the navigation bar, see Figure 1.