Looking for Linking: Associative Links on the Web

Timothy Miles-Board
University of Southampton
Southampton, UK
tmb99r@ecs.soton.ac.uk

Leslie Carr
University of Southampton
Southampton, UK
lac@ecs.soton.ac.uk

Wendy Hall
University of Southampton
Southampton, UK
wh@ecs.soton.ac.uk

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*navigation*

General Terms

Measurement

Keywords

Associative Linking, Internet Archive. Link Taxonomies

1. INTRODUCTION

Non-trivial hypertexts (containing more than one node) use links to implement their internal structure. On the Web *navigation bars* have become ubiquitous, defining functional regions on a web page that expose a site's primary structure, listing nearby pages or media (home page, next page, previous page, search, related links).

By contrast, *associative linking* [2] takes place in the *content* regions of Web pages and may be used to interlink related concepts from the domain semantics, expose argumentation structures, add glossary functions or reveal instructional components according to various secondary informational schemas or controlling "applications".

In this paper we describe an attempt to identify the latter kind of links on the World Wide Web, as the preliminary stage of recognising and classifying "good" linking practices that go beyond the merely organisational infrastructure common to the Web.

These associative linking practises are exemplified by NASA's *Astronomy Picture of the Day* (APOD) archive[1], a popular website which illustrates and discusses different astrological phenomena. Each day's text is linked to relevant information from previous days, and also to external educational and scientific Web pages which explain or illustrate any key phrases and technical terms used in the text. The authors of the site explain that each text is "written explicitly to be linked", with the content "constructed as an abstract, with links providing all the detailed or background information required by the various readership profiles."

[1] `http://antwrp.gsfc.nasa.gov/apod`

HT'02, June 11-15, 2002, College Park, Maryland, USA.

	Links	Density
Mean	27	0.17
Median	14	0.12
Standard Deviation	60.66	0.17

Table 1: Distribution of links and link density

2. THE SEARCH FOR ASSOCIATIVE LINKS

The Internet Archive[2] provides a digital library of crawled Web sites, consisting of over 10 billion pages. It provides an ideal resource for examining linking practices on the Web in recent years. We assume that a well-linked hypertext is something which can not only be recognised when seen by a human, but which has identifiable and measureable attributes. The challenge of this investigation was therefore to identify characterics of this practice in order that examples may be identified programmatically without manual intervention.

As a first step, a script was written which collected statistics describing approximately 500,000 web pages (selected at random) from the archive. The results are summarised in Table 1. Subsequently a profile of a well-linked web page was constructed, assuming that a well-linked web page was the exception rather than the norm (an assumption supported by previous observations [1]): a well-linked page would have more links than average and a higher than average link density[3].

Given this profile, we initially expected that those web pages best matching the profile were well-linked. However, we discovered that this was a flawed assumption since web pages typically exhibit *multiple linking practices simultaneously.* For example, our "profile" of well-linked hypertexts rewards pages with a large number of links, and penalises those with fewer links. "Bookmark lists" for example, could masquerade as examples of deep linking, while genuine deeply linked texts (with comparatively fewer links) could be overlooked.

The next step was therefore to "purify" the data by ignoring non-associative linking practices. We devised a simple set of rules which segment a page into a number of smaller *micro pages.* These rules are based on the way in which

[2] `http://www.archive.org`
[3] Link density is the ratio of linked text to unlinked text

Functional Section	Content Section
High link density Ordered link distribution	Medium/Low link density Random link distribution

Table 2: Comparative characteristics of functional and content sections

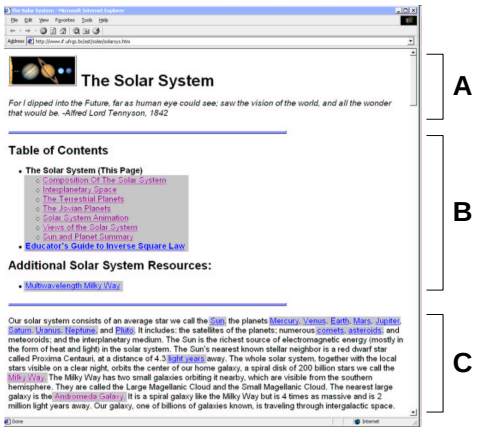

Figure 1: Segmenting a Web page (linked text highlighted for clarity).

readers visually parse pages into spatial areas of content and function, and segment a page according to HTML tags which visually isolate segments (for example, paragraphs, headings, tables, forms). Each micro page could then be classified as either a functional or content section, using heuristics described in Table 2. Figure 1 shows how a page is segmented by our script. In this example, sections A (no links) and B (high link density and ordered distribution of links) are ignored; section C is identified as a well-linked content section.

The script was configured to look for web pages containing at least 4 micro pages with: *at least 4 links*, *at least 30 words*, an *average of 4 words between links*, and a *link density of no more than 80%*. From a total of 770,992 randomly selected pages, 576 were identified as meeting this criteria. A further visual inspection elimated false positives - these have subsequently informed improvements to our page segmenting heuristics. The genuine associatively-linked pages contained a number of common linking practices:

Reference links A number of pages consistently linked proper nouns when they appeared in the narrative. Observed examples included linking the names of people, products, places, and the titles of other web pages to home pages, product descriptions etc.

Deep links Links on keywords or concepts which tied together resources with an aim to provide a deeper understanding of the topic in hand.

Structural links Observed within well-linked structured pages, such as manuals, essays, and papers. References to figures, sections, chapters, adjacent pages (e.g. next/previous/top) etc. were consistently linked.

Citation links Also frequently observed in formal or technical pages. Citations (or endnotes) in the narrative were consistently linked.

Glossary links These were observed both in pages which linked terms to a glossary (or dictionary), and within glossary pages themselves.

The most frequently observed linking practice was reference linking (occuring in 84% of the well-linked pages), followed by deep linking (37%), glossary/dictionary linking (35%), citation/endnote linking (32%), and structural linking (17%).

3. RELATED AND FUTURE WORK

De Rose proposes a taxonomy in which hypertext links are categorised as either extensional (stored individually) or intensional (inferred) [2]. Arguably, the majority of links on the Web must fall into the former category, since they are embedded in web pages. However, our structural, citation, and glossary links are similar to De Rose's (intensional) implicit links. Our reference links may also be similar (for example, in the case of the name of an organisation linked to its home page). Our deep links are typical of De Rose's (extensional) associative links.

Haas *et al.* empirically investigated the use of hypertext links on web pages [4], informing a link taxonomy of 4 major categories: *navigation*, *expansion*, *resource* and *miscellaneous*. Our structural, citation, and glossary links are examples of navigation links, and our reference and deep links examples of expansion links.

Not all hypertext researchers agree that inline linking from the content section of a document is a good thing. The HDM hypermedia design model [3] focuses on the connections between 'atomic pages' and discourages the use of inline linking so that the structure of the hypertext is clearer to both author and reader. The Yale C/AIM style guide expresses strong concern about associative linking on the WWW, claiming that overuse or poor placement of links in the content of web pages can disrupt the narrative flow by inviting readers to go elsewhere [5].

In the future, we hope to develop heuristics which allow different types of well-linked hypertext to be recognised, and so a true measure of associative linking practises on the Web can be obtained.

4. ACKNOWLEDGEMENTS

Thanks to the authors of the APOD Web site, Robert Nemiroff and Jerry Bonnell for their help in answering questions about the site.

5. REFERENCES

[1] L. Carr, W. Hall, and T. Miles-Board. Writing and reading hypermedia on the web. Technical Report ECSTR-IAM00-1, University of Southampton, Southampton, UK, February 2000.

[2] S. J. De Rose. Expanding the notion of links. In *Proceedings of the Hypertext '89 Conference on Hypertext, Pittsburgh, Pennsylvania, USA*, Usability, Links, and Fiction, pages 249–257, 1989.

[3] F. Garzotto, L. Mainetti, and P. Paolini. Hypermedia design, analysis and evaluation issues. *CACM*, 38(8):74–84, 1995.

[4] S. W. Haas and E. S. Grams. A link taxonomy for web pages. In *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, pages 485–495, 1998.

[5] P. Lynch and S. Horton. Imprudent linking weaves a tangled web. *IEEE Computing*, 30(7):115–117, July 1997.