Chapter 3

Searching for Associative Writing in the Web

Having defined Associative Writing in the previous chapter as the process of using associative links to integrate a writer's new contributions with existing materials, it was decided that the next step should be to illustrate this definition by describing some examples of integrated hypertexts found in the Web. This chapter therefore describes two investigations which aimed to uncover evidence of Associative Writing in the Web for discussion and inspiration. The first investigation proceeded using a manual inspection of numerous Web sites and pages, based on the recommendations of colleagues. Although some significant examples were found and analysed, the number of examples was far less than anticipated.

In response to this result, a follow-up investigation was initiated with a slight shift of focus: rather than attempt to uncover individual examples of integrated hypertexts by hand through recommendations and ad-hoc browsing, the investigation attempted to conduct a systematic search of a much larger cross-section of the Web, beyond the means of a manual search. It was also hoped that the results of this investigation would help begin to quantify the extent of Associative Writing on the Web at large. To achieve this, evidence of Associative Writing in Web pages archived by the Internet Archive organisation[1] was computationally gathered and reported. The results of both investigations are presented and analysed in this chapter, and also positioned relative to existing studies of Web linking phenomena — some of which have helped inform proposals for further investigation. In order to introduce these investigations, we must first consider *how* to look for evidence of Associative Writing in the often heavily graphic- and navigation-oriented pages of today's Web.

[1] `http://www.archive.org/`

3.1 Where is the Associative Writing?

The Web provides only a single, simple linking mechanism; a point-to-page link that enables writers to connect two Web pages. The growing complexity of Web pages and sites however, has forced Web authors and designers to use these links to achieve a multitude of different functions. This is illustrated by Haas *et al*'s investigation of the use of links in Web pages, which described over 30 different uses of the basic link (Haas and Grams, 1998b) falling under 4 major categories:

1. *Navigation* — links articulating the structure of a Web site, for example *home page*, *previous page*, *next page*, *search*, *table of contents.*
2. *Expansion* — links leading to a detailed presentation of the link anchor, for example *illustration*, *diagram*, *graph*, *definition*, *citation*, *example*, *video clip*, *audio clip.*
3. *Resource* — links to topically related pages.
4. *Miscellaneous* — for example, advertisement links.

Web designers rely on presentation conventions that have emerged through the Web's history to provide visual cues to the reader about which links on the page are being used for which purpose; Web browsers treat all links equally in the display[2] so it is up to the designer to lay out information on the page in such a way as to convey the purpose of each link to the user. Of particular importance is the distinction between what this work terms "functional" and "content" regions of a Web page.

Functional regions on a page serve to expose a site's primary structure, providing local coherence (Moulthrop, 1992) by listing links to nearby pages or media (cf. Haas *et al*'s navigation links). The positioning and layout of functional regions on a Web page have become standard across the most frequently visited sites, and hence adopted by many other designers. Nielsen has documented some of these now *de facto* standards (Nielsen, 1999), which he terms *landmarks* (Nielsen, 1995). Many sites place a horizontal set of tabs across the top of the page to indicate the main areas of their content (Figure 3.1). A popular way of presenting navigation links is to place a coloured strip containing the links down the left (Figure 3.2) or top of the page. Many sites also use a "breadcrumb trail" across the top of the page to situate the current page relative to its parent nodes and to allow users to jump up several levels of the site hierarchy in a single click (Figure 3.1 and 3.3). Some "index‘‘ or "bookmark" pages may consist purely of navigational links to other pages (Figure 3.3). Navigation links provide a function which is orthogonal to the content of the page, so embedding such links in the content of the page makes little

[2] Of course, the visual attributes of links in the browser (for example, colour) can be specified, but there is no standard for doing so; in fact Nielsen recommends that the default visual attributes be allowed to prevail (Nielsen, 1996a).

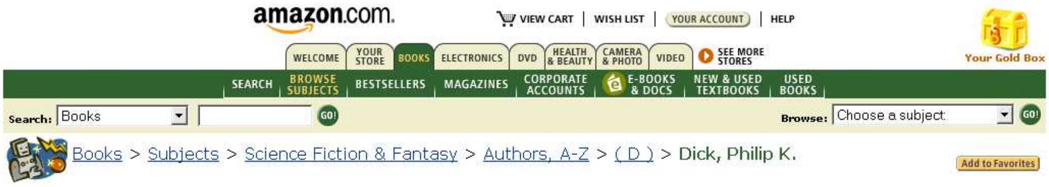
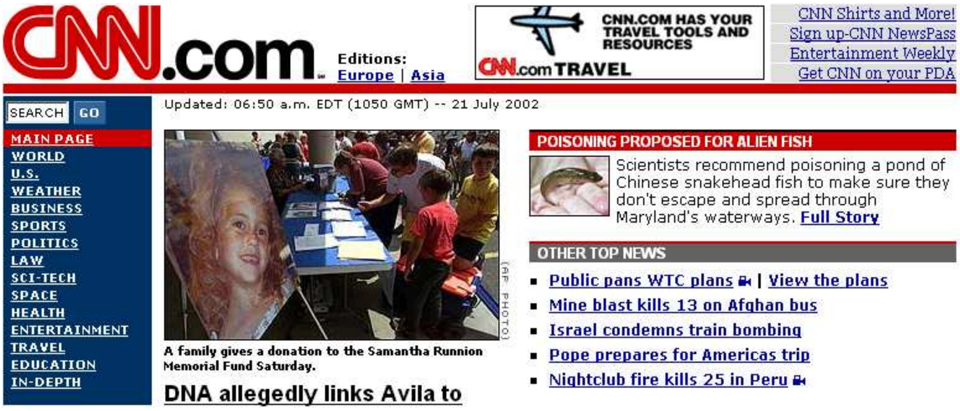


Figure 3.1: Navigation tabs and breadcrumb trail from *Amazon*.



Updated: 06:50 a.m. EDT (1050 GMT) -- 21 July 2002

A family gives a donation to the Samantha Runnion Memorial Fund Saturday.

DNA allegedly links Avila to

Scientists recommend poisoning a pond of Chinese snakehead fish to make sure they don't escape and spread through Maryland's waterways. Full Story

- Public pans WTC plans | View the plans
- Mine blast kills 13 on Afghan bus
- Israel condemns train bombing
- Pope prepares for Americas trip
- Nightclub fire kills 25 in Peru

Figure 3.2: Left-hand navigation strip and lists of navigation links from *CNN*.

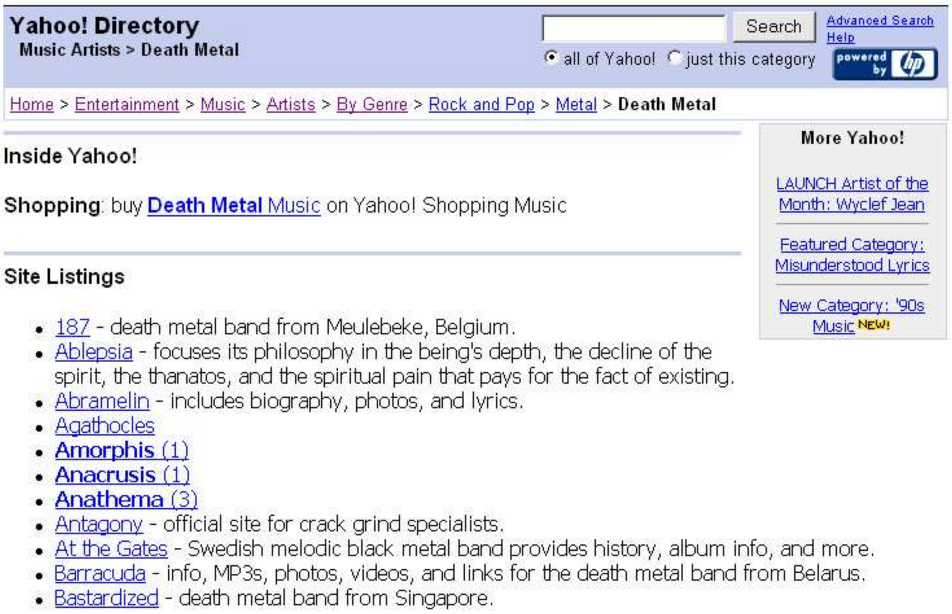
sense; indeed Haas *et al* observed that navigational links were typically isolated from content:

> Some uses of links are becoming conventionalised. For example, many pages provide a navigation bar at the top or bottom of the page, containing links to major sub-pages of the site. These links are isolated from the text or other content of the page. We found, in fact, that 84% of isolated links are for navigation, and that 82% of navigation links are isolated (Haas and Grams, 1998b, pg. 107).

By contrast, the content regions of Web pages remain largely unstructured and are a potentially rich source of Haas *et al*'s expansion and resource links. It is therefore within these content regions that evidence of Associative Writing may be found.

3.2 Investigation 1: Manual Search

In early 2000, a study of the use of linking in the Web was carried out with the aim of finding some interesting examples of Associative Writing (Carr et al., 2000b). The results of this investigation were presented at the 9th International World-Wide Web conference (Carr et al., 2000a).



Yahoo! Directory
Music Artists > Death Metal

● all of Yahoo! ○ just this category

Home > Entertainment > Music > Artists > By Genre > Rock and Pop > Metal > Death Metal

Inside Yahoo!

Shopping: buy **Death Metal** Music on Yahoo! Shopping Music

Site Listings

- 187 - death metal band from Meulebeke, Belgium.
- Ablepsia - focuses its philosophy in the being's depth, the decline of the spirit, the thanatos, and the spiritual pain that pays for the fact of existing.
- Abramelin - includes biography, photos, and lyrics.
- Agathocles
- **Amorphis** (1)
- **Anacrusis** (1)
- **Anathema** (3)
- Antagony - official site for crack grind specialists.
- At the Gates - Swedish melodic black metal band provides history, album info, and more.
- Barracuda - info, MP3s, photos, videos, and links for the death metal band from Belarus.
- Bastardized - death metal band from Singapore.

More Yahoo!

LAUNCH Artist of the Month: Wyclef Jean

Featured Category: Misunderstood Lyrics

New Category: '90s Music NEW!

Figure 3.3: Breadcrumb trail and list of navigational links from *Yahoo!*.

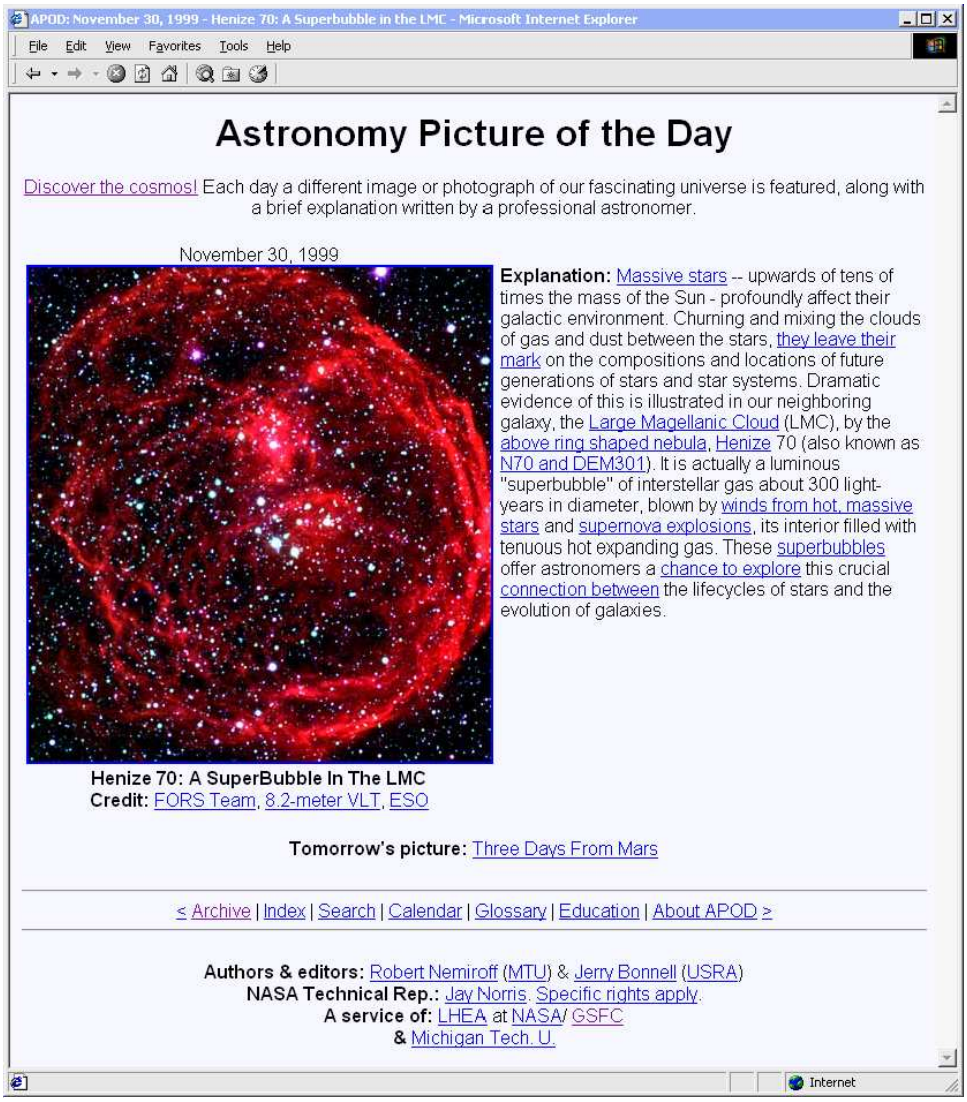
The investigation proceeded with a manual inspection of numerous Web sites and pages, based on recommendations from colleagues and contributors to the hypertext community website *Hypertext Kitchen*[3]. We were surprised to discover just how much of the linking on the many personal home pages, commercial Web sites and educational sites we inspected was functional — structured navigation regions and ad-hoc, non-hierarchical "related information" links presented alongside unlinked content. We did however, discover a small number of scientific and technical sites which demonstrated interesting content linking strategies.

With some of these sites, it proved difficult to make the distinction between functional and content regions. Web Log ('blog') style sites, such as *SlashDot* and *Scripting News*, use links in short 'news' paragraphs but it became difficult to tell whether the links annotated the news items or the sentence-long news items simply annotated the links. A number of other news sites, such as *CNET* and *Wired* provided links within their stories to the home pages (or stock prices) of commercial or institutional bodies that are mentioned, whereas Nielsen's own *Alertbox* site linked from the content of the current "bulletin" any relevant previous bulletins.

NASA's *Astronomy Picture of the Day*[4], a popular site which illustrates and discusses different astronomical phenomena, provided links from each day's content not just to relevant information from previous days, but also to external educational and scientific Web pages which explain or illustrate any key ideas, concepts or technical terms used in the text (Figure 3.4). The online *Scientific American* provided a similar service for its

[3] `http://www.hypertextkitchen.com/`

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


APOD: November 30, 1999 - Henize 70: A Superbubble in the LMC - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Internet

Astronomy Picture of the Day

Discover the cosmos! Each day a different image or photograph of our fascinating universe is featured, along with a brief explanation written by a professional astronomer.

November 30, 1999

Explanation: Massive stars -- upwards of tens of times the mass of the Sun - profoundly affect their galactic environment. Churning and mixing the clouds of gas and dust between the stars, they leave their mark on the compositions and locations of future generations of stars and star systems. Dramatic evidence of this is illustrated in our neighboring galaxy, the Large Magellanic Cloud (LMC), by the above ring shaped nebula, Henize 70 (also known as N70 and DEM301). It is actually a luminous "superbubble" of interstellar gas about 300 light-years in diameter, blown by winds from hot, massive stars and supernova explosions, its interior filled with tenuous hot expanding gas. These superbubbles offer astronomers a chance to explore this crucial connection between the lifecycles of stars and the evolution of galaxies.

Henize 70: A SuperBubble In The LMC
Credit: FORS Team, 8.2-meter VLT, ESO

Tomorrow's picture: Three Days From Mars

< Archive | Index | Search | Calendar | Glossary | Education | About APOD >

Authors & editors: Robert Nemiroff (MTU) & Jerry Bonnell (USRA)
NASA Technical Rep.: Jay Norris. Specific rights apply.
A service of: LHEA at NASA/ GSFC
& Michigan Tech. U.

FIGURE 3.4: Astronomy Picture of the Day for 30th November 1999.

"enhanced articles"[5]. Although both sites share a similar brief on increasing the public understanding of science, *Scientific American*'s British counterpart *New Scientist* was found to provide no links from the content of its articles. The same applied for the online *National Geographic*, a finding also reported by (Fagerjord, 2001).

Both the *Astronomy Picture of the Day* and *Scientific American* sites seemed to illustrate good examples of Associative Writing in the Web, so the next step was to contact the editors of both sites in order to obtain an insight into how these processes were taking place. The editors of the *Scientific American* site reported that as an online ("enhanced") version of a printed text, articles go through separate writing and editorial processes, requiring both writer input to suggest suitable links to existing related materials in the Web, and editorial input to ensure consistency of approach over time.

By contrast, Robert Nemiroff, the co-producer of *Astronomy Picture of the Day*, reported

[5] At the time of writing (September 2002), however, this practice seems to have been curbed.

that pages on the site do not undergo separate writing and editorial processes, and instead are written explicitly to be linked. Each day's content is constructed to function as an "abstract" with links providing all the detailed or background information required by the various readership profiles. As such, the process of Associative Writing is seen by the authors as *easier* than that of writing linear text, because linear text must "express every idea and elaboration that is necessary to the understanding of the subject", a philosophy that reinforces some of the advantages of Associative Writing put forward in the previous chapter.

In an attempt to explain why so few examples were found, the investigation concluded that perhaps Associative Writing is sufficiently at odds with writers' "normal" experience of literacy as to limit its widespread use. However, this explanation seems at odds with the principles on which this work bases an understanding of Associative Writing, which argue that writers "borrow and sew together [text] to create new discourse" (Porter, 1986) and that the real power of documents is their ability to refer to, refute, and elaborate on each other (Hill and Mehlenbacher, 1996). The investigation also suggested that the effort required to locate high quality material to link to may be an issue here: certainly in the case of *Astronomy Picture of the Day*, competent editorial experience and a knowledge of the kinds of material available in a particular subject domain are the key to the site's approach to writing.

3.3 Investigation 2: Automatic Search

After the manual search for evidence of Associative Writing uncovered fewer examples than we had anticipated, a second investigation was initiated in late 2001 with a slight shift of focus: rather than attempt to uncover individual examples of integrated hypertexts by hand through recommendations and ad-hoc browsing, the investigation would attempt to conduct a systematic search of a much larger cross-section of the Web and help give some quantitative indication of the extent of Associative Writing on the Web at large.

Such an investigation has been made possible by the efforts of the Internet Archive, which provides powerful computational access to a massive digital library of Web pages. Collections of Web pages are acquired by the Internet Archive using Web-crawling *robots*, which automatically gather pages from publicly accessible sites — each page is then examined for links to other pages that can be queued for crawling. At the time of the investigation, the Internet Archive was estimated to have archived over 10 billion pages, requiring over 100 terabytes of digital storage. This section reports on the techniques developed by the author to computationally analyse the archived Web pages and identify evidence of Associative Writing. The results of this investigation were presented at the 2002 ACM Hypertext Conference (Miles-Board et al., 2002).

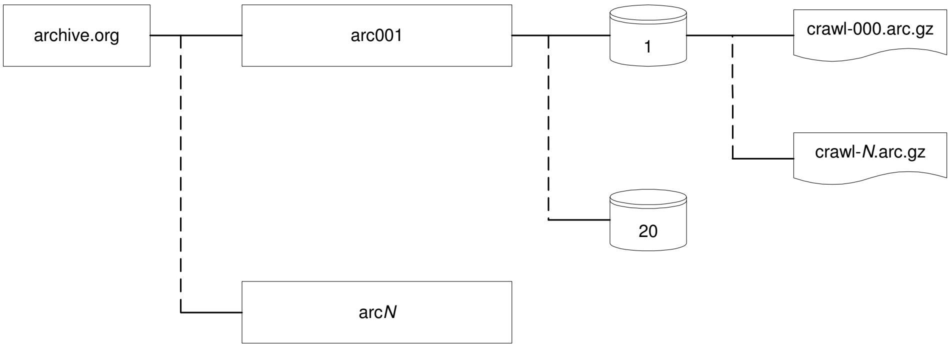


FIGURE 3.5: Overview of the Internet Archive architecture.

Number of links	27
Link density	17%

TABLE 3.1: Characteristics of an 'average' page from the dataset.

3.3.1 Searching the Internet Archive

Although openly accessible to researchers, a working knowledge of the Unix environment and Perl scripting is essential for examining the contents of the Internet Archive. The archive is distributed across many hundreds of machines, each with up to 20 hard disk drives filled with compressed "crawl data-files" (Burner and Kahle, 1996), each containing approximately 10^4 Web pages (Figure 3.5). Unix tools provided by the Internet Archive librarians allow custom analyses to be carried out in parallel across any number of machines, vastly increasing the range and efficiency of analysis.

3.3.1.1 Dataset

100 crawl data-files were randomly selected from the entire archive, collectively archiving a total of 770,992 individual Web pages[6] gathered by the Internet Archive robots between January 1997 and March 2001. Table 3.1 shows a simple profile of an 'average' page from the dataset. Pages were most likely to have been gathered from a `.com` domain (accounting for 53% of the dataset), followed by `.edu` (10%), `.org` (6%), and `.net` (5%). Note that the statistics in Table 3.1 describe the characteristics of an average page with no distinction made between linking in functional and content regions — the next step was to develop techniques for automatically decomposing each page from the dataset into such regions.

[6] The data-files contained many different Web media, including HTML, plain text, images, and video. Of these files, 770,992 were in HTML format and used for analysis.

```
br button caption center code col colgroup dd
dir div dl dt form frame frames frameset h1 h2 h3
h4 h5 h6 hr li ol optgroup option p select table
tbody td textarea tfoot th thead tr ul
```

FIGURE 3.6: HTML elements treated as region boundaries.

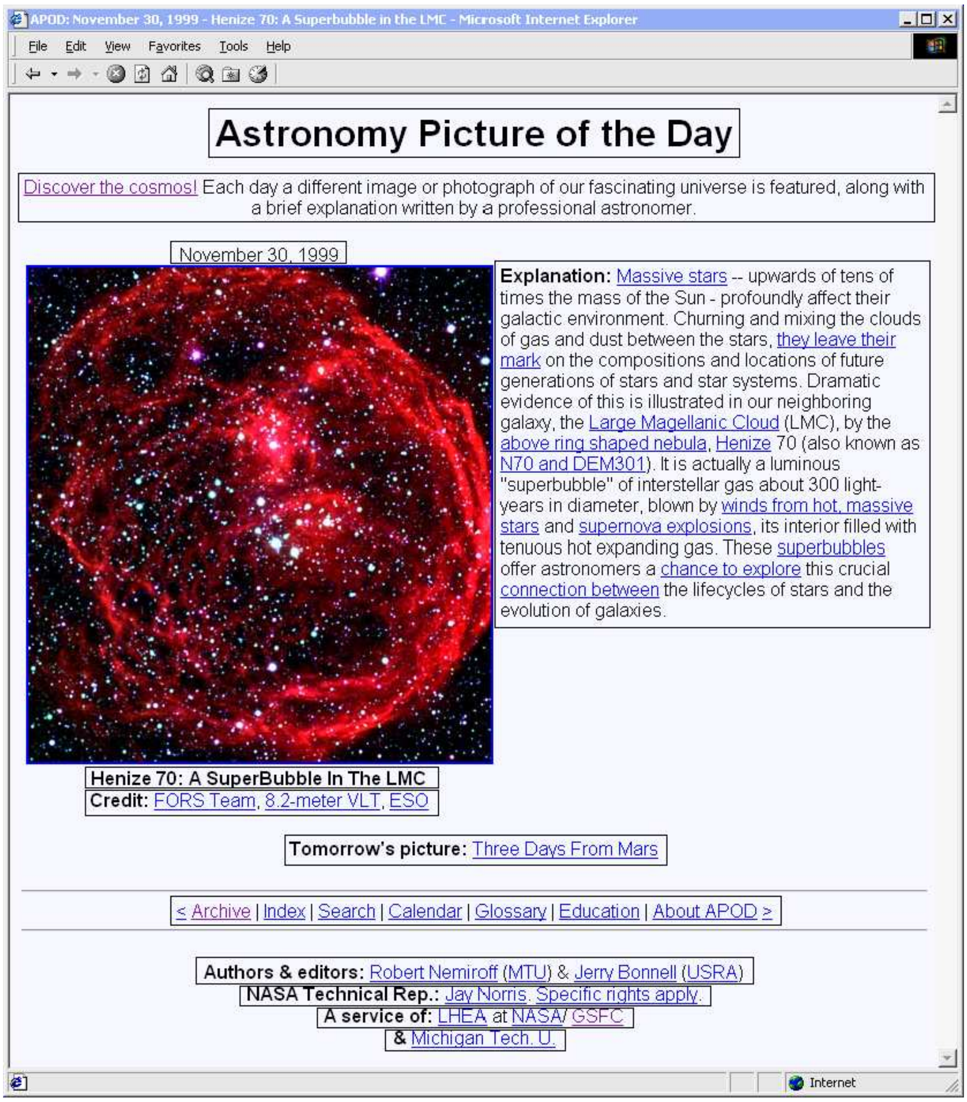
3.3.1.2 Design

In the first investigation, a hybrid approach was used to identify functional and content regions of Web pages: we manually examined how the *Astronomy Picture of the Day* and *Scientific American* pages used visual and structural cues to separate functional from content regions and used this knowledge to “bootstrap” the statistical analysis of content regions reported in (Carr et al., 2000b). However, the size and diversity of the dataset in the second investigation precluded any such hybrid approach. An algorithm was therefore developed which could automatically decompose any Web page into its constituent functional and content regions.

The algorithm works by analysing the HTML structure underlying each page: instead of applying pre-formulated rules for identifying boundaries between regions, it treats any HTML element which causes content to be visually separated from its surroundings as a boundary between one region and the next. The most obvious region boundaries are elements such as headings, paragraphs, and tables; lower-level instructions such as *bold* and *italic* text formatting are ignored. For completeness, Figure 3.6 lists all the HTML elements which have been identified as region boundaries. Figure 3.3.1.2 shows how the algorithm decomposes a page from the *Astronomy Picture of the Day* site into regions.

Having decomposed a page into its constituent regions, the algorithm then identified which regions were functional and which actually contained the content of the page. In the first investigation, a link density “threshold” was tailored specifically to the *Astronomy Picture of the Day* and *Scientific American* pages — regions with a link density above this threshold were treated as functional and excluded from further analysis. The large dataset in this investigation called for a more generic approach, and so number of potential heuristics for automatically identifying functional and content regions were identified:

Link destination Information about link destination may give some clue about intended purpose. Regions containing on-site (local) links may well be functional; regions containing off-site links are perhaps more likely to be content. However, this metric may be highly misleading: many content links in the *Astronomy Picture of the Day* and *Scientific American* sites were previous local articles; the *New Scientist* site was distributed across several servers, so off-site links actually proved to be serving a functional purpose.


APOD: November 30, 1999 - Henize 70: A Superbubble in the LMC - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Internet

Astronomy Picture of the Day

Discover the cosmos! Each day a different image or photograph of our fascinating universe is featured, along with a brief explanation written by a professional astronomer.

November 30, 1999

Explanation: Massive stars -- upwards of tens of times the mass of the Sun - profoundly affect their galactic environment. Churning and mixing the clouds of gas and dust between the stars, they leave their mark on the compositions and locations of future generations of stars and star systems. Dramatic evidence of this is illustrated in our neighboring galaxy, the Large Magellanic Cloud (LMC), by the above ring shaped nebula, Henize 70 (also known as N70 and DEM301). It is actually a luminous "superbubble" of interstellar gas about 300 light-years in diameter, blown by winds from hot, massive stars and supernova explosions, its interior filled with tenuous hot expanding gas. These superbubbles offer astronomers a chance to explore this crucial connection between the lifecycles of stars and the evolution of galaxies.

Henize 70: A SuperBubble In The LMC
Credit: FORS Team, 8.2-meter VLT, ESO

Tomorrow's picture: Three Days From Mars

< Archive | Index | Search | Calendar | Glossary | Education | About APOD >

Authors & editors: Robert Nemiroff (MTU) & Jerry Bonnell (USRA)
NASA Technical Rep.: Jay Norris. Specific rights apply.
A service of: LHEA at NASA/ GSFC
& Michigan Tech. U.

Figure 3.7: Decomposing a Web page into regions of navigation and content.

Number of links The total number of links in a region. No links would indicate an unlinked content region; one or more links would indicate a functional or content region.

Region Length From the observations in the first investigation, one might expect content regions to contain more text than functional regions.

Link distribution The distribution of links in a region. Functional regions seem to exhibit an ordered distribution (for example, a list of links), whereas linking in content regions is far more erratic.

Link density The ratio of linked text (text appearing inside link anchors) to unlinked text. Since functional regions typically contain little other than navigational links, a high link density would tend to indicate a functional region.

3.3.1.3 Procedure

A Perl script automatically examined each of the 770,992 HTML pages in the dataset, applying the following process to each individual page:

1. Divide page into regions according to boundary algorithm described above.
2. Derive the following metrics for each region:
   (a) Total number of links.
   (b) Total length, in words.
   (c) Distribution of links (expressed as the average number of unlinked words between link anchors).
   (d) Link density.
3. Compare these metrics to a set of pre-determined "thresholds":
   (a) The region must contain *at least 4* links.
   (b) The length of the region must be *at least 30* words.
   (c) The average number of unlinked words between link anchors in the region must be *at least 4*.
   (d) The link density of the region must be *not more than 80%*.
4. If page contains *at least 4* separate regions which match these thresholds, report the URL of the page as a positive 'hit', otherwise report the page as a 'miss'.

In order to discover examples of Associative Writing, the metrics were designed to filter out pages which contained several 'well-linked' content regions in the hope that such pages would approximate well to Associative Writing. As it turned out, however, these well-linked content regions were not always evidence of a higher-level Associative Writing strategy in action, a finding that will be discussed shortly. The criteria for reporting a positive "hit" was intentionally set high in anticipation of only a small number hits being reported — these pages could then be manually checked in order to verify the correct operation of the script before initiating further investigations with lower reporting criteria. The larger number of "hits" expected from further runs could then be reported with a greater degree of confidence.

3.3.2 Results

A total of 576 pages from the dataset were reported as positive hits. These pages were manually checked using the *Wayback Machine*[7] (Feise, 2000), a public interface to the

[7] `http://web.archive.org/`

`http://web.archive.org/web/19991013033522/http://artofcheese.com:80/product.htm`

Wayback Machine *archived date* *URL of archived page*

Figure 3.8: Wayback Machine URL addressing scheme.

Internet Archive which allows casual users to search for and access archived versions of Web pages through a normal browser without having to access the underlying storage mechanisms of the archive. For each positive hit, the script constructed a Wayback Machine URL so that the page could be easily inspected in a browser. Figure 3.8 shows the Wayback Machine URL from a reported hit.

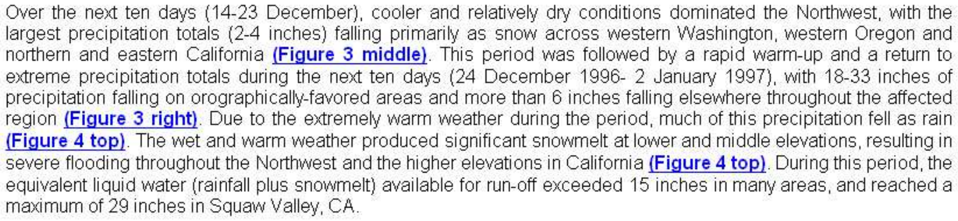
Of the 576 pages reported by the script, 264 pages (46% of the total reported) were confirmed as positive 'hits' (correctly meeting the reporting criteria outlined above). The remaining 312 pages (54%) were identified as *false positives* (incorrectly meeting the reporting criteria).

3.3.2.1 Analysis of Positive Hits

During the manual verification of the results reported from the first run, it was noted that a number of different linking practices were evident in the positive results; well-linked content sections were not always evidence of a higher-level Associative Writing strategy in action, as initially hoped. The distinction between the different kinds of content linking observed is less clear-cut than that between functional and content regions. In order to classify the different kinds of content linking observed, not only did each link anchor and its surrounding context have to be carefully examined, but also each link had to be followed in an attempt to understand the significance of the relationship between the link anchor and the destination page. This examination revealed a total of 58 pages (22% of the positive hits) which were considered to be evidence of an Associative Writing strategy. A small selection of these pages is presented in Appendix A. Many pages also contained evidence of a number of content linking strategies, which were classified into four major categories: structure linking, citation linking, reference linking, and glossary linking. Table 3.2 summarises the observed instances of each type of linking.

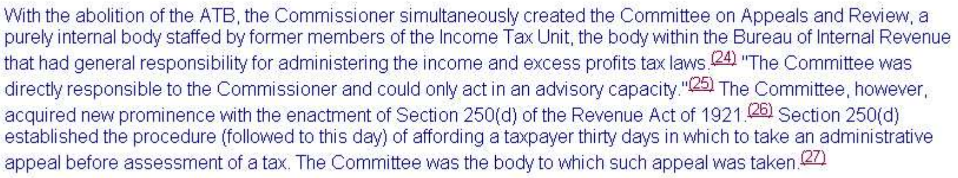
Structure linking 9% of pages consistently linked references to structural features when they appeared in the content (for example, *Figure A*, *Section 2*, *Chapter XL*). The target of these links was usually highly predictable. Pages containing structure linking typically formed parts of larger structured documents, such as essays, manuals, and research papers. Figure 3.9 shows an example of structural linking in a technical report.

Citation linking Citation marks in the content were linked to full citation details (rather than the cited work itself) by 16% of pages. Like structure linking, this practice



Over the next ten days (14-23 December), cooler and relatively dry conditions dominated the Northwest, with the largest precipitation totals (2-4 inches) falling primarily as snow across western Washington, western Oregon and northern and eastern California **(Figure 3 middle)**. This period was followed by a rapid warm-up and a return to extreme precipitation totals during the next ten days (24 December 1996- 2 January 1997), with 18-33 inches of precipitation falling on orographically-favored areas and more than 6 inches falling elsewhere throughout the affected region **(Figure 3 right)**. Due to the extremely warm weather during the period, much of this precipitation fell as rain **(Figure 4 top)**. The wet and warm weather produced significant snowmelt at lower and middle elevations, resulting in severe flooding throughout the Northwest and the higher elevations in California **(Figure 4 top)**. During this period, the equivalent liquid water (rainfall plus snowmelt) available for run-off exceeded 15 inches in many areas, and reached a maximum of 29 inches in Squaw Valley, CA.

Figure 3.9: Structure linking in *NOAA Special Climate Summary, January 1997.*



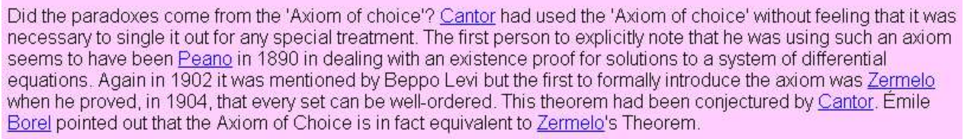
With the abolition of the ATB, the Commissioner simultaneously created the Committee on Appeals and Review, a purely internal body staffed by former members of the Income Tax Unit, the body within the Bureau of Internal Revenue that had general responsibility for administering the income and excess profits tax laws.(24) "The Committee was directly responsible to the Commissioner and could only act in an advisory capacity."(25) The Committee, however, acquired new prominence with the enactment of Section 250(d) of the Revenue Act of 1921.(26) Section 250(d) established the procedure (followed to this day) of affording a taxpayer thirty days in which to take an administrative appeal before assessment of a tax. The Committee was the body to which such appeal was taken.(27)

Figure 3.10: Citation linking in *Judicial Independence: Can It Be Without Article III?*.

was typically confined to pages with technical content and the target of the links was also highly predicatable. Figure 3.10 shows an example of citation linking in an essay.

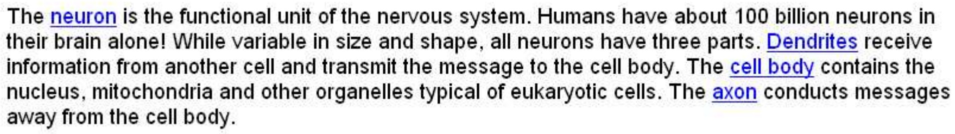
Reference linking 70% of the hits consistently linked proper nouns when they appeared in content regions, for example, linking references to people, products, organisations, and places, to a "home page", page containing further information, or a page which allowed readers to buy the product. Destination Web pages were also referenced directly by name in the link anchor. Figure 3.11 shows an example of reference linking (names of mathematicians appearing in the content are linked to a page on the same site describing their life and work). The targets of reference links were usually predictable (for example, in Figure 3.11, the link anchor *Zermelo* leads to a page about Zermelo), but by no means standardised.

Glossary linking 16% of the hits used a specialised form of reference linking where the target of the link is a definition in a glossary or dictionary, often part of the same site. For example, each biological term in Figure 3.12 is linked to its definition in a biological dictionary on the same site.



Did the paradoxes come from the 'Axiom of choice'? Cantor had used the 'Axiom of choice' without feeling that it was necessary to single it out for any special treatment. The first person to explicitly note that he was using such an axiom seems to have been Peano in 1890 in dealing with an existence proof for solutions to a system of differential equations. Again in 1902 it was mentioned by Beppo Levi but the first to formally introduce the axiom was Zermelo when he proved, in 1904, that every set can be well-ordered. This theorem had been conjectured by Cantor. Émile Borel pointed out that the Axiom of Choice is in fact equivalent to Zermelo's Theorem.

Figure 3.11: Reference linking in *The Beginnings of Set Theory.*



The neuron is the functional unit of the nervous system. Humans have about 100 billion neurons in their brain alone! While variable in size and shape, all neurons have three parts. Dendrites receive information from another cell and transmit the message to the cell body. The cell body contains the nucleus, mitochondria and other organelles typical of eukaryotic cells. The axon conducts messages away from the cell body.

FIGURE 3.12: Glossary linking in *The Nervous System.*

Content linking strategy	% of positive hits using strategy
Reference linking	70
Associative Writing	22
Glossary linking	16
Citation linking	16
Structure linking	9

TABLE 3.2: Summary of observed instances of content linking strategies.

3.3.2.2 Related Work in Link Taxonomies

The purpose of this section is to position the analysis of positive hits relative to the work of Haas *et al* (briefly described at the beginning of this chapter), and other link taxonomies described in the literature. The link taxonomies reported by Haas *et al* (Haas and Grams, 1998a,b) and Fagerjord (Fagerjord, 2001) are also empirical, and so methodologies can also be compared.

Haas *et al* carried out a content analysis of 75 randomly selected Web pages, informing a link taxonomy of 4 major categories: navigation, expansion, resource, and miscellaneous. Navigation and expansion links accounted for over 80% of the observed links; resource links, seeming closest to Associative Writing, accounted for only 15%. In contrast to the investigations reported here which focus on linking strategies in content regions of pages, Haas *et al* examined *all* links in their dataset, compiling and correlating results from two independent investigators. However, the dataset examined by Haas *et al* was significantly smaller; 75 compared to approx. 780,000 pages.

Fagerjord also describes an analysis of a relatively small dataset, in the comparison of a selection of Web "features" on the National Geographic Web site with their magazine and television counterparts (each feature appeared in each of the three presentation mediums). In analysing these features, Fagerjord divided links into three categories:

1. Navigation links that articulate a site's structure.
2. Presentation links used to display the next or previous part of a text, or start a movie.
3. Relation links that jump to another place in the hypertext that is related in some way to the present page or paragraph.

	Haas *et al*	Fagerjord
Structure		presentation
Citation		
Reference	expansion	
Glossary	expansion	
Associative W.	resource	related
Functional	navigation	navigation

Table 3.3: Correlating observed linking strategies with other empirical work.

Fagerjord reported that he found no relation links in the National Geographic sites he analysed, only navigation and presentation links, concluding that the sites were actually "more linear" than the film and magazine counterparts. Table 3.3 shows the correlations between the link categories observed in the investigation reported here (with the addition of a 'functional' link category for completeness) and those of Haas *et al* and Fagerjord.

There have been many other proposals for link taxonomies in the literature, mostly created in the context of hypertext systems. Haas *et al* note that in terms of linking pages (or nodes) together, there is a crucial difference between hypertext systems and the Web (Haas and Grams, 1998b): A hypertext system is generally created with a single purpose or theme, and although several authors may work on the system for several years, there is generally a sense of unity and similarity of style throughout the system. Pages on the Web, however, share no such unity or purpose of style; writers can link to any other page on the Web, regardless of the difference between the source and target page in terms of genre, style, intended audience, or even language or culture. Therefore, in comparison with taxonomies from various closed, application- or domain-specific hypertext systems, a classification of Web linking is much more general in order to account for the types of relationships between pages found on the Web. However, as Table 3.4 shows, there is some correlation between the observed linking practices reported here and such *a priori* taxonomies.

Trigg One of the earliest reports of research using any form of typed links is Trigg's dissertation (Trigg, 1983), recently reappraised by (Bernstein, 2001). Trigg's Textnet hypertext system included 75 different link types, which were broadly classified into two categories: normal and commentary links. Normal links, largely rhetorical, connected nodes on the basis of argument structure or discourse, for example: citation, background, explanation, example, futurework, refutation, support. Commentary links connected comments and criticisms to the work, for example: comment-critical, relatedwork-ignores, argumentation-redherring, style-boring.

DeRose DeRose proposed a link taxonomy to represent the interconnections between different versions, translations and annotations of ancient texts in a hypertext net-

	Trigg	DeRose	Baron *et al*	Cleary & Bareiss
Structure		implicit		
Citation	citation†	implicit		
Reference	explanation, summarisation, detail..	implicit, annotational	rhetorical	
Glossary	explanation, summarisation, detail..	implicit	rhetorical	
Associative W.	normal, commentary	associative	content-based	conversational associative
Functional		organisational		

†Both taxonomies share a citation link type, but Trigg's intention is to link directly to the cited work rather than details of it (possibly using more specialised citation links such as *pioneer* and *eponym*).

Table 3.4: Correlating observed linking strategies with other a priori work.

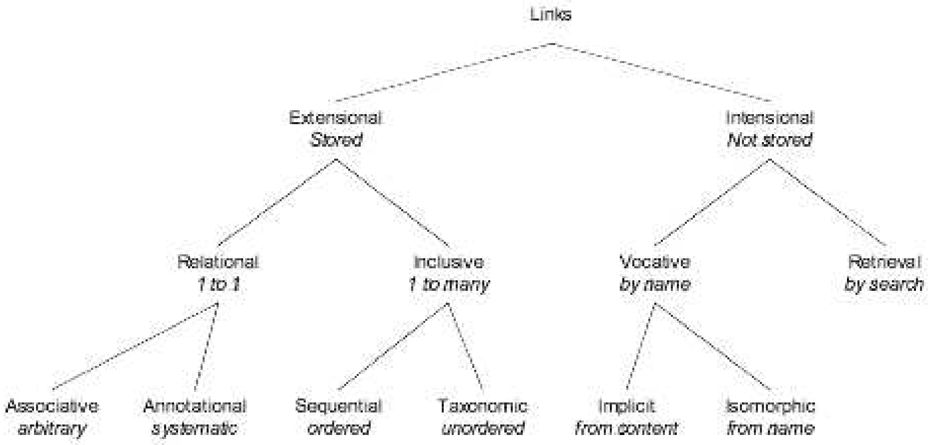


Figure 3.13: DeRose's link taxonomy (DeRose, 1989).

work (DeRose, 1989). Although created with this specific domain in mind, DeRose felt that his taxonomy (Figure 3.13) was generic enough to be applied to other domains.

Four of DeRose's more conceptually complex link types bear further explanation:

Associative Connect single locations together in entirely *unpredictable* ways.

Annotational Connect single locations together in *predictable* ways, representing connections from portions of a text to *information about the text*, such as the presence of linguistic, thematic, or other phenomena.

Implicit Inferred when a particular document element is invoked by name in the content of a document. Examples are links between terms used in the text and their dictionary definitions, and references to named sections (for example, *Chapter 2*).

Isomorphic Inferred when a particular document element appears as an element name in a document. A simple example is a link that connects the same structural element (for example a section) in two different translations of the same document.

The two main classes, extensional and intensional, are separated by the dimension of storage — extensional links are based on the writer's suggested uses of the hypertext and need to be stored; intensional links can be automatically inferred and therefore do not need to be stored. Links are also divided into different categories according to their "ended-ness" — relational links are 1 to 1 (binary), inclusive links are 1 to many — and in the case of inclusive links, whether the end points are ordered or unordered. These dimensions make it difficult to draw overall comparisons with the (purely semantic) link types informed by investigation 2 since Web links are binary and rarely inferred; indeed Haas *et al* point out that the utility of such implicit links depends on shared knowledge among hypertext authors and readers (Haas and Grams, 1998b) — such assumptions are more likely to hold in a closed hypertext system than in the unconstrained arena of the Web.

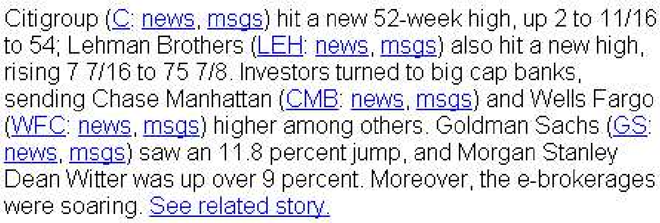
Baron et al Baron *et al* proposed a taxonomy of link types for application in a study of the use of a hypertext reference manual (Baron et al., 1996). The taxonomy consisted of two major categories: organisational and content-based links. Organisational links were those used specifically for navigation through the hypertext, for example tables of content, and directional cues such as previous and next page. Content-based links dealt with specific relationships between nodes, and included semantic links such as similar, contrast, and part/kind of, rhetorical links such as definition, explanation, illustration, and summary, and pragmatic links such as warning, prerequisite, and example.

Cleary & Bareiss Cleary & Bareiss (Cleary and Bareiss, 1996) describe a link taxonomy based on a simple theory of conversation: "at any point in a conversation, there are only a few general categories of follow-up statements that constitute a natural continuation rather than a topic shift." The resulting set of eight "conversational associative categories" are presented as binary alternatives under four broader classes:

- Refocusing: Context/Specifies
- Comparison: Analogies/Alternatives
- Causality: Causes/Results
- Advice: Opportunities/Warnings

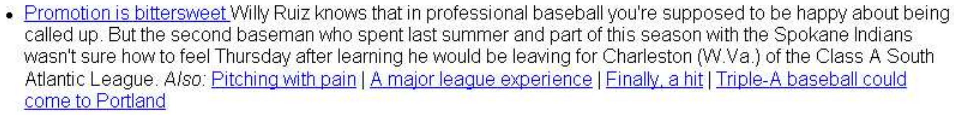
3.3.2.3 Analysis of False Positives

The diagnosed "false positive" results stemmed from two particular problems:



Citigroup (C: news, msgs) hit a new 52-week high, up 2 to 11/16 to 54; Lehman Brothers (LEH: news, msgs) also hit a new high, rising 7 7/16 to 75 7/8. Investors turned to big cap banks, sending Chase Manhattan (CMB: news, msgs) and Wells Fargo (WFC: news, msgs) higher among others. Goldman Sachs (GS: news, msgs) saw an 11.8 percent jump, and Morgan Stanley Dean Witter was up over 9 percent. Moreover, the e-brokerages were soaring. See related story.

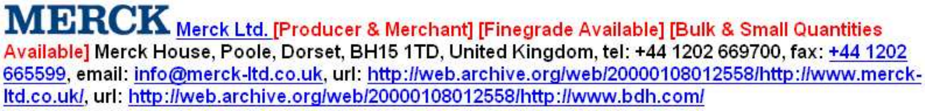
Figure 3.14: False positive: *CBS MarketWatch, October 28 1999* (16 links, 100 words, mean distance between links 5 words, link density 18%).



- Promotion is bittersweet Willy Ruiz knows that in professional baseball you're supposed to be happy about being called up. But the second baseman who spent last summer and part of this season with the Spokane Indians wasn't sure how to feel Thursday after learning he would be leaving for Charleston (W.Va.) of the Class A South Atlantic League. *Also:* Pitching with pain | A major league experience | Finally, a hit | Triple-A baseball could come to Portland

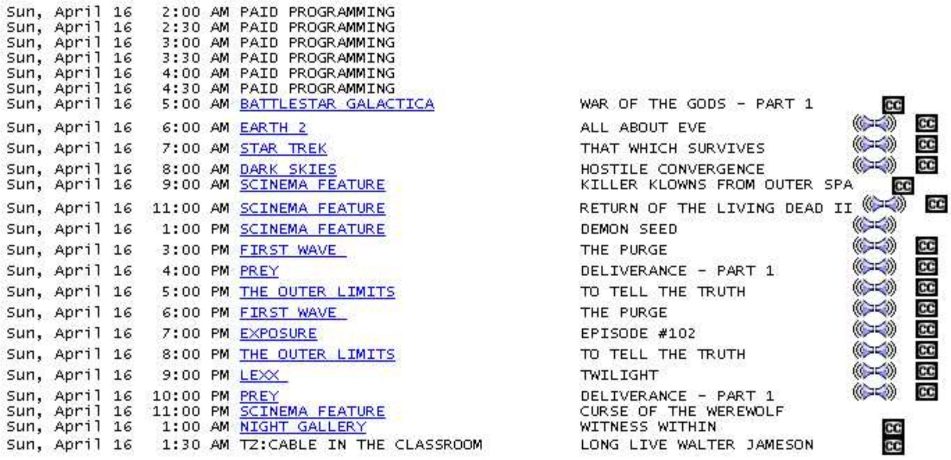
Figure 3.15: False positive: *Spokane.net News Stories July 4 1999* (5 links, 78 words, mean distance between links 10 words, link density 24%).

- 59 pages (19% of the false positive results) contained regions that matched the selection criteria, but on closer inspection the linking strategy in these regions served a functional purpose. For example, Figure 3.14 shows a region from a *CBS Marketwatch* page which contains a small embedded navigation menu next to each stock symbol reported in the content. Figures 3.15 and 3.16 are typical of many false positive results, where the "content regions" are in fact annotated functional links.
- 253 pages (81% of the false positives) were found to have been incorrectly reported due to an oversight in the design of the script. These pages made extensive use of preformatted text to exercise explicit control over the layout of text on the page. The script treated each area of preformatted text as single region of the page, so large runs of preformatted text containing multiple links were able to meet the reporting criteria. Innovative uses of preformatted text included lottery numbers (Figure 3.18), family trees (Figure 3.19 — at least 60 pages from different sites appeared to have been created using the same genealogy software package), legacy documents, timetables, program code, television schedules (Figure 3.17), newgroups, and FAQs.



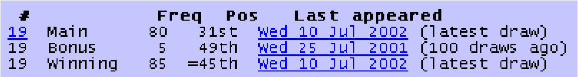
MERCK Merck Ltd. [Producer & Merchant] [Finegrade Available] [Bulk & Small Quantities Available] Merck House, Poole, Dorset, BH15 1TD, United Kingdom, tel: +44 1202 669700, fax: +44 1202 665599, email: info@merck-ltd.co.uk, url: http://web.archive.org/web/20000108012558/http://www.merck-ltd.co.uk/, url: http://web.archive.org/web/20000108012558/http://www.bdh.com/

Figure 3.16: False positive: *Sourcerer - Magnesium Nitrate Suppliers* (5 links, 37 words, mean distance between links 5 words, link density 22%).



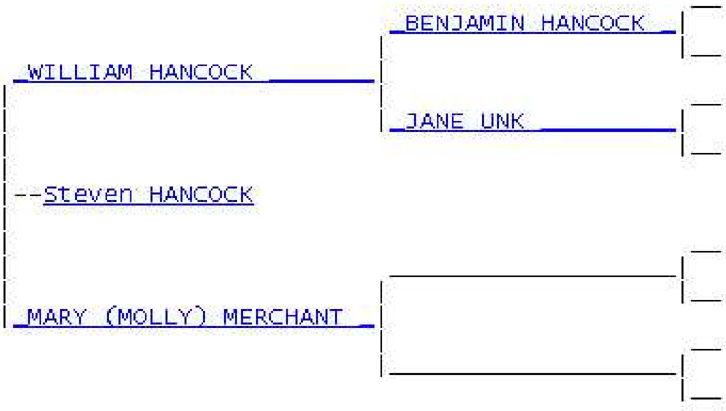
```
Sun, April 16   2:00 AM PAID PROGRAMMING
Sun, April 16   2:30 AM PAID PROGRAMMING
Sun, April 16   3:00 AM PAID PROGRAMMING
Sun, April 16   3:30 AM PAID PROGRAMMING
Sun, April 16   4:00 AM PAID PROGRAMMING
Sun, April 16   4:30 AM PAID PROGRAMMING
Sun, April 16   5:00 AM BATTLESTAR GALACTICA           WAR OF THE GODS - PART 1          CC
Sun, April 16   6:00 AM EARTH 2                        ALL ABOUT EVE                        CC
Sun, April 16   7:00 AM STAR TREK                      THAT WHICH SURVIVES                  CC
Sun, April 16   8:00 AM DARK SKIES                     HOSTILE CONVERGENCE                  CC
Sun, April 16   9:00 AM SCINEMA FEATURE                KILLER KLOWNS FROM OUTER SPA       CC
Sun, April 16  11:00 AM SCINEMA FEATURE                RETURN OF THE LIVING DEAD II         CC
Sun, April 16   1:00 PM SCINEMA FEATURE                DEMON SEED
Sun, April 16   3:00 PM FIRST WAVE                     THE PURGE                            CC
Sun, April 16   4:00 PM PREY                           DELIVERANCE - PART 1                 CC
Sun, April 16   5:00 PM THE OUTER LIMITS               TO TELL THE TRUTH                    CC
Sun, April 16   6:00 PM FIRST WAVE                     THE PURGE                            CC
Sun, April 16   7:00 PM EXPOSURE                       EPISODE #102                         CC
Sun, April 16   8:00 PM THE OUTER LIMITS               TO TELL THE TRUTH                    CC
Sun, April 16   9:00 PM LEXX                           TWILIGHT                             CC
Sun, April 16  10:00 PM PREY                           DELIVERANCE - PART 1                 CC
Sun, April 16  11:00 PM SCINEMA FEATURE                CURSE OF THE WEREWOLF
Sun, April 16   1:00 AM NIGHT GALLERY                  WITNESS WITHIN                    CC
Sun, April 16   1:30 AM TZ:CABLE IN THE CLASSROOM      LONG LIVE WALTER JAMESON          CC
```

FIGURE 3.17: Preformatted text: Television schedule from *Scifi.com ScheduleBot April 16, 2000* (17 links, 248 words, mean distance between links 12 words, link density 13%).



```
 #              Freq  Pos    Last appeared
19  Main         80    31st  Wed 10 Jul 2002 (latest draw)
19  Bonus         5    49th  Wed 25 Jul 2001 (100 draws ago)
19  Winning      85   =45th  Wed 10 Jul 2002 (latest draw)
```

FIGURE 3.18: Preformatted text: Lottery statistics from *UK National Lotto Number Palace* (4 links, 36 words, mean distance between links 5 words, link density 36%).



```
                                      _BENJAMIN HANCOCK _|
                                     |                   |__
 _WILLIAM HANCOCK ________|
|                        |                              __
|                        |_JANE UNK __________|
|                                                       |__
|
|--Steven HANCOCK
|
|                                                       __
|                          ___________________|
|                        |                              |__
|_MARY (MOLLY) MERCHANT _|
                         |                              __
                         |___________________|
                                                        |__
```

FIGURE 3.19: Preformatted text: Genealogical information from *Steven Hancock Family Tree* (5 links, 44 words, mean distance between links 5 words, link density 34%).

3.3.2.4 Discussion: Implications for Future Searches

This investigation succeeded in uncovering some evidence of Associative Writing, a sample of which is presented in Appendix A. In the manual verification of the reported 'hits' from the dataset, it was found that pages matching the thresholds demonstrated several different content linking strategies. In total, evidence of Associative Writing accounted for at approximately 0.01% of the entire dataset (although lower reporting thresholds are likely to increase this estimate).

The observed content linking strategies seem to correlate well with those reported by other empirical studies and also with *a priori* taxonomies designed for hypertext systems, although the observed link types were more generic compared to the specific link semantics described by the closed hypertext link taxonomies. Independent verification of different content linking practices, as demonstrated by (Haas and Grams, 1998b), may be useful in future investigations. It may also be useful to be able to identify the different content linking strategies 'automatically' in order to carry out a more focused search for Associative Writing. Each linking practice may have properties which lend themselves to automatic identification, for example, citation link anchors often take the form of "[2]", and structure link anchors the form "Section *n*", although these usages are by no means standard.

The relatively large number of 'false positives' reported by the investigation can be used to inform updates to the search algorithm. These false positives were largely due to an oversight in the design of the algorithm which considered each run of preformatted text (even those encompassing several paragraphs) as a single region. The correct approach would have been to split the preformatted text into paragraphs and then treat each paragraph as a separate region.

However, the remaining false positive results may be more difficult to occlude from the set of reported results: regions which met the requirement thresholds for 'linked content region' were found on careful inspection to in fact serve a functional purpose. The reporting thresholds were purposefully set high in investigation 2 in anticipation of performing a manual verification of the results. It is reasonable to assume that future investigations using lower reporting thresholds will uncover more examples of Associative Writing (pages from the *Astronomy Picture of the Day* and *Scientific American* sites would not have met the thresholds set in investigation 2), but that these results will be polluted by such 'functional content regions' — results obtained from larger output sets (beyond the means of thorough manual verification) may therefore be more difficult to state with confidence.

Some further weaknesses in the investigation procedure can also be highlighted:

Pseudo-random dataset The dataset used in this investigation was created by randomly selecting 100 crawl data-files from the entire Internet Archive. However, since the Internet Archive robots work by following links from each Web page they encounter (and archive), each datafile is likely to contain a number of pages from the same site, and therefore the dataset may not be a representative sample of the Web at large. However, the storage architecture and sheer volume of information in the Internet Archive makes it difficult to select pages truly at random. Perhaps a compromise would be to iteratively select a number of pages at random from crawl data-files (also randomly selected), until a dataset quota is met. It may also be necessary to check that the dataset contains no duplicates (the same page archived at different times). A more representative dataset would allow future results to be stated with a greater degree of confidence.

Link distribution The distribution of links in a region was identified as a useful metric for distinguishing linked content regions from functional regions. A basic assumption was that navigational regions would exhibit an ordered distribution, whereas linking in content regions would be far more erratic and random. However, the link distribution of a region was expressed as the average number of words appearing between link anchors, which in itself does not capture the 'orderliness' of a linked region well. (Zhu et al., 2002) demonstrate how a Markov model can be used to analyse and predict Web site navigation patterns — this approach could be adapted to help identify linked content regions in Web pages. Given a sequence of inter-link distances (distances in words between link anchors), the Markov model could be used to predict the distance to the next link. We would expect the prediction to fail often in content regions (unordered link distribution) and succeed often in navigation regions (ordered link distribution).

3.4 Directions for Future Investigation

The aim of this section is to discuss how other related studies of Web linking phenomena have helped propose possible future directions for uncovering and quantifying evidence of Associative Writing in the Web.

3.4.1 Web Connectivity Analysis

Both investigations described in this chapter focused on the use of associative links within content regions of individual pages. In contrast, Web Connectivity Analysis is the study of the use of linking on the Web as a whole (Heylighen, 2000). According to Web Connectivity analysts, all links on the Web are in principle equivalent — the Web itself does not express any preference for one link or one document above another — however, a lot of implicit information about the relative importance of links is contained in the *connectivity* or pattern of linkages between pages. Writers normally only link to other pages that are relevant to the general subject of their writing, and of sufficient quality. Therefore, high quality documents, containing clear, accurate and useful information, are likely to have many links pointing to them, while low quality documents will get few or no links. Hence there is a preference implicit in the total number of links pointing to a page, although no explicit preference function is attached to the link. The creator of a page *p*, by including a link to page *q*, has in some measure *conferred authority* on *q* (Kleinberg, 1999a). This implicit "endorsement" is produced collectively, by the group of all Web authors.

Extracting this implicit information from the connectivity of the Web has had important implications for information retrieval. Traditional search engines built giant indices allowing users to quickly retrieve the set of all Web pages containing a given word or string. A topic of any breadth typically resulted in several thousand or even several million relevant Web pages. Furthermore, the most authoritative pages on a topic often don't contain the search term. (Chakrabarti et al., 1999) point out, "there is no reason to expect the home pages of *Honda* or *Toyota* to contain the term 'japanese automobile manufacturers', or the home pages of *Microsoft* or *Lotus* to contain the term 'software companies'." Connectivity analysis, in contrast allows the most "definitive" or "authoritative" Web pages on a topic to be discovered. Communities of thematically related pages on the Web can be characterised by the way in which they link to their most central, prominent members. These prominent sources serve as a form of broad-topic summary of a much larger underlying ensemble, condensing an enormous amount of information down to a more tractable representation (Kleinberg, 1999b).

A number of algorithms exist to extract information about authoritative pages from the link structure of the Web. The PageRank algorithm (Brin and Page, 1998) is used by the popular *Google* search engine. Such search engines have introduced a "political economy" of links in the Web (Walker, 2002). Since a link from *p* to q translates to a precise PageRank value, a link from *p* to *q* has a clearer value to *q* than the content of *q*'s page has to *p*'s readers: the author of *p* "pays" for *q*'s content with the link. The HITS (Kleinberg, 1999a; Lempel and Moran, 2001) and Clever (Chakrabarti et al., 1999) algorithms identify two types of pages: *authorities* that are linked to by many good *hubs*, and *hubs* which are linked to many good *authorities*. Authorities on topics are the most prominent sources of primary content. Hubs are high-quality guides and resource lists that direct users to recommended authorities (Kleinberg, 1999b).

The identification of hubs and authorities in Web Connectivity Analysis could be useful for uncovering evidence of Associative Writing in the Web: hubs may be a good starting point for search/analysis. Kleinberg's attempts to offset the effect of navigation links in the identification of hubs and authorities are particularly relevant:

> "We distinguish between two types of links...We say that a link is trans-

> verse if it is between pages with different domain names, and intrinsic if it is between pages with the same domain name...Since intrinsic links very often exist purely to allow for navigation of the infrastructure of the site, they convey much less information than transverse links about the authority of the pages they point to. Thus, we delete all intrinsic links from the graph..., keeping only the edges corresponding to transverse links.." (Kleinberg, 1999a).

Although observations from investigation 1 might suggest that "intrinsic" (functional) links aren't always navigational (and that "transverse" links aren't always content links), the removal of intrinsic links from the analysis seems to indicate that hubs would be good sources of associative links. However, the hubs reported by Kleinberg do typically take the form of long index pages or lists of links rather than pages created using an Associative Writing strategy. Even so, this technique may be useful for "homing in" on areas of interesting linking practices in the Web.

The addition of pages and links to the Web is a distributed, asynchronous, complex and continual process (Pennock et al., 2002). Yet, when examined as a whole, discernible linking patterns emerge, some of which are shared with other social and biological networks (Barabasi and Albert, 1999). The distribution of the number of links to and from a Web page has been shown to follow a power law over many orders of magnitude. Power law scaling can be attributed to a "rich get richer" (or "winner takes all") mechanism: as the Web grows, the probability that a given page receives a link is proportional to the current connectivity of that page, leading to a relatively small number of sites receiving a disproportionately large share of links. However, (Pennock et al., 2002) demonstrate that, among collections of Web pages of the same type, the distribution of inbound links deviates strongly from a power law at small connectivities, implying that, relative to their community, winning pages don't quite "take all" — less popular pages still attract a considerably higher proportion of links than would be the case under a power law distribution.

This assertion is encouraging since it implies that some well thought out linking strategies are taking place beneath the linking "bandwagons", and that the hub-based approach proposed above may indeed uncover some examples of Associative Writing.

3.4.2 Dominance and Connectedness

(Jackson, 1997) demonstrates how network analysis can be applied to the interpretation of linking structures on the Web using indices of *dominance* and *connectedness.* Dominance is the deviation from equality of the distribution of links among nodes. In a system with high dominance, most links will connect to a select number of nodes. Connectedness is the ratio of actual connections existing within a collection of pages to the

	High Dominance	Low Dominance
High Connectedness	Satellite structure	Hypertext/Associative structure
Low Connectedness	Index structure	Linear, narrative structure

Table 3.5: Identifying various linking structures using network analysis (Jackson, 1997).

total number that are possible: a ratio of 1.00 would indicate a structure in which each node is connected bidirectionally to every other node. Table 3.5 shows how measures of dominance and connectedness can be used to identify various structures, described below:

- *High connectedness and high dominance* indicates a network with a high number of links, but a very skewed distribution of those links. Such a network might exhibit a "satellite" structure in which a few dominant pages are central.
- *High connectedness and low dominance* indicates a network with a high number of links distributed evenly across nodes. In such a structure, users may move from any one node to another, at any time — this is the ideal structure to support the associative movement proposed by the original hypertext vision (Bush, 1945).
- *Low connectedness and low dominance* indicates a network with few links distributed evenly across nodes, consistent with linear narrative offering few paths through the pages (for example, an article divided into sections, with "next" and "previous" links between each section).
- *Low connectedness and high dominance* indicates a network with a few links concentrated among a few pages, for example an information repository with a central page acting as an index or "list of links" to all available information.

Jackson notes that as a "hypertext index", *connectedness* would be useful to researchers interested in assessing the extent to which Web designers are creating structures capable of supporting associative thought (Jackson, 1997). Measuring the dominance and connectedness of collections or communities of pages may therefore help "home in" on areas of the Web deserving more thorough investigation.

A further consideration for future work is that the complex social dynamics of the Web may influence the way in which writers create links; for example, corporations don't want users to leave their site and so don't provide links to competitors who may be offering lower-priced services. In investigation 2, the dataset was chosen at random without consideration of these factors — were we looking for evidence of Associative Writing in places where one wouldn't expect to find it anyway? Perhaps a better approach would be to take a collection of pages that we would *expect* to be interconnected (for example, a community of non-commercial pages on the same topic rather than competing commercial pages) and measure the *dominance* and *connectedness* of the community as an indication of Associative Writing strategies taking place — the Web Connectivity Analysis techniques outlined here could be used to help identify these communities in the first instance.

3.5 Summary

This chapter has described two investigations which aimed to uncover evidence of Associative Writing in the Web. Today's Web is often heavily graphic- and navigation-oriented, so to introduce these investigations, the question of *how* to look for evidence of Associative Writing was first considered and the notion of "functional" and "content" regions introduced. Functional regions — or "landmarks" (Nielsen, 1995) — on a Web page serve to expose a site's primary structure, by listing links to nearby pages or media (home page, previous page, next page, search page, contents page), whereas content regions remain largely unstructured are a potentially rich source of associative links. It is therefore within these content regions that evidence of Associative Writing may be found. The first investigation proceeded using a manual inspection of the content regions of numerous Web sites and pages, based on the recommendations of colleagues. Although some significant examples were found, in the form of pages from the *Astronomy Picture of the Day* and *Scientific American* sites, prompting a discussion of the editorial process behind both sites, the number of examples of Associative Writing uncovered was far less than anticipated.

The second investigation therefore focused on conducting a systematic search of a much larger cross-section of the Web than could be investigated by hand. This investigation took advantage of the vast digital repository of Web pages stored by the *Internet Archive*, from which a dataset of 100 crawl data-files (archiving a total of 70,992 HTML pages) was selected to form the basis of analysis. A generic algorithm was developed to automatically split each page into its constituent functional and content regions, and to determine whether the page potentially showed evidence of an Associative Writing strategy. In the manual verification of the reported 'hits' from the dataset, it was found that the pages demonstrated several different content linking strategies, and that in total, evidence of Associative Writing accounted for at approximately 0.01% of the entire dataset. The observed content linking strategies were correlated with those reported by other empirical Web studies and also with *a priori* taxonomies designed for hypertext systems. Future improvements to the search algorithm were suggested, including those informed by a number of reported 'false positive' results. Related studies of Web linking phenomena, such as the notions of 'hubs' and 'authorities' (Kleinberg, 1999a) and 'dominance' and 'connectedness' (Jackson, 1997), were also used to suggest alternative approaches for future consideration.