Do HTML Tags Flag Semantic Content?

HTML tags are designed to support only the display of Web page content, but this study quantifies their feasibility as proxies for semantic content as well.

Jonathan Hodgson
St. Joseph's University

Experience shows that existing search engines find too many pages and that it is hard, without actually reading the pages found, to determine their usefulness. The use of HTML structural elements to evaluate the contents of a Web page has been addressed both in products and in research literature. For example, Brin and Page[1] applied an algorithm that evaluates the relevance of Web pages using tagged texts in the construction of the Google search engine. Other researchers have used the number of links to and from a Web page as a way of determining "authoritative" Web pages (see the sidebar "Related Work in Tagging and Semantics"). However, none of this work provides data on the relation of the textual content of tagged text to the semantic content of a Web page. Although mechanisms for flagging semantic content do exist, as with the Extensible Markup Language, my goal here was to see if authors using HTML tags used them in a similar manner. This is of interest because most Web pages are written using only HTML and because the tags in XML are not as yet standardized.

I have examined a number of Web pages to try to quantify whether, consciously or otherwise, authors use headings or the highlighted text accompanying a link to indicate the subject matter of the corresponding page.

The following Web page elements were studied:

- *Link (anchor) text:* Text enclosed by the tag `<a href= ...> ...</a>` was evaluated for clues to page contents.
- *Heading text:* Text enclosed by heading tags `<h1> </h1> ...<h6> </h6>` was evaluated for clues to page contents. The idea was that this text would serve a purpose similar to that of a book's table of contents.
- *Metatag and comment text:* Text enclosed in the metatag and comments enclosed between `<!–` and `–>` were evaluated because metatags often provide keywords.

My investigations focused on empirically evaluating the feasibility of using key Web page elements as proxies to indicate page contents. Additionally, I compared text retrieval by keyword. This

 1089-7801/00/$10.00 ©2001 IEEE

article describes the methods I used and summarizes results. My empirical results suggest that text in HTML headings and in anchor texts is useful for indicating Web page content for logic programming.

Approach

The experiments were done using a Sun Sparc 5, (primarily because this machine provided me with the faster network connection). The extraction of tagged text was done using a handwritten analyzer coded in standard Prolog. Prolog was chosen for its support of grammar rules.

Because tag relevance was a concern, I initially used a binary evaluation to distinguish highlighted texts that seemed to carry no content—such as the phrase *click here*—from those with semantic content, such as *machine learning*. Later, I used a technique that accepted several degrees of informativeness. Also, the results were parceled out to evaluators—students and faculty possessing varying degrees of knowledge about logic programming and the domains of the selected Web pages.

An unanticipated problem, particularly relevant for the potential of automatic Web page analysis, is the degree to which browsers forgive HTML format infractions. The code for a Web page analyzer needs therefore to accept as many of these infractions as possible (that is, to be as permissive as the average browser), if it is to cover real-world Web pages. Certain HTML infractions that browsers permit, like those shown in Table 1 (next page), required me to modify my analyzer to accommodate them, by adding rules that accepted the infractions as "legitimate HTML." It was this analyzer that was used for the studies.

Method

I obtained the data set for these experiments, first, by querying the Google search engine (http://www.google.com/technology/index.html) to seek sites pertaining to *logic programming*. This topic yielded primarily academic sites that were predominantly text based.

The first 1,000 pages returned from the query were then extracted and downloaded. For pages that included frames, I downloaded the files containing all the frame texts. In this process, the number of pages was winnowed down to 553 because either the URL no longer existed, the attempt to contact or download the URL timed out, or the URL contained syntax errors.

I created two databases from these pages. The first one contained file summary data: URL and URL title (if any), number of frames in the URL, number of links and headings, and number of metatagged texts and comments.

The second database contained the following tagged texts:

- The URL for the file
- A flag indicating whether or not the URL has frames
- The title of the URL (if any)
- The list of links with both linking text and the URL or the linked-to file
- The texts of the headings
- The texts of the comments and metatagged texts
- A list of keywords that appear in the file

To identify keywords that were added to this second database, I created, and hand reviewed, a sort-

Related Work in Tagging and Semantics

Researchers have used HTML tags in constructing search engines. For example, Laser[1] uses emphasized text to improve the ranking of retrieved pages. Google[2] uses both links to a page and headings within a page to calculate page rankings. Both these systems are exploiting the semantic content carried in tagged text, but they are not concerned with authorial practice, which is the focus of the work reported in the main text here.

Other researchers[3,4] have used the number of links pointing to a page to establish its "authoritativeness." This work complements the use of link text to indicate page content.

The Dublin Core (http://purl.org/dc/) is a comprehensive effort to provide ontologies for use in document markup. The Extensible Markup Language (http://www.w3.org/XML/) offers a way of tagging document contents to carry their semantic information. The work reported here investigates current practices with the smaller system of tags represented in HTML, albeit not a system designed for semantics.

Other researchers have treated Web pages as programs.[5,6] The use of headings as surrogates (proxies) is a simple version of this idea.

References

1. J. Boyan, D. Freitag, and T. Joachims, "A Machine Learning Architecture for Optimizing Web Searches," *Proc. AAAI Workshop Internet-Based Information Systems*, AAAI Press, Menlo Park, Calif., 1996; also available online at http://www.lb.cs.cmu.edu/afs/cs/project/reinforcement/papers/boyan.laser.ps.
2. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks ISDN Systems*, vol. 30, 1998, pp. 107-117.
3. S. Chakrabarti et al., "Automatic Resource List Compilation by Analyzing Hyperlink Structure and Associated Text," *Computer Networks ISDN Systems*, vol. 30, 1998, pp. 65-74.
4. H.P. Frie and D. Steiger, "The Use of Semantic Links in Hypertext," *Information Process and Management*, vol. 31, no. 1, 1995, pp. 1-13.
5. S.W. Loke and A. Davison, "LogicWeb: Enhancing the Web with Logic Programming," *J. Logic Programming*, vol. 36, no. 9, 1998, pp. 195-240.
6. E. Pontelli and G. Gupta, "W-ACE: A Logic Language for Intelligent Internet Programming," *IEEE Int'l Conf. Tools with Artificial Intelligence*, IEEE Press, Piscataway, N.J., 1997, pp. 2-10.

Table 1. Examples of HTML infractions that browsers permit.

Problem	Description	Result
Faulty bracketing of comments	The correct form is <!— ……— >. However, browsers permit the trailing — > to be concatenated to the last token in the comment.	The HTML is not accepted by a strict parser/analyzer.
Accepting misplaced HTML tags	Example: linking text <a href="def.html" more linking text This is not part of the link.	The browser interprets the single as terminating both opening <a> tags
	The user sees "linking text more linking text" in the link.	The reader does not see that there are actually two separate links.
Using the header tag, instead of the <font> tag, for font sizing	Example: <h3> Some text <h2> Now it's larger </h3> Not a header Output: **Some text Now it's larger** Not a header	Causes the appearance of unbalanced and inconsistent headers, which means that a strict HTML parser would reject the document although it looks OK.

ed list of words from approximately 100 Web pages related to logic programming. Words relating to logic programming were kept as keywords.

To assess Web-page text informativeness, I created two questionnaires and distributed them to the evaluators. Each questionnaire contained 25 texts randomly drawn from either the link or the header texts. Evaluators selected from one of four questionnaire categories:

- *No help:* The text gave no guidance as to the document's semantic content. Stand-alone numbers fell into this category.
- *Little help:* The text gave minimal guidance, meaning that contextual information, if available, would make the text helpful. An example would be the stand-alone word *home*.
- *Helpful:* The text gave some information on the document content. Names of individuals fell into this category.
- *Very helpful:* The text gave a clear impression of the likely contents. An example is *WWW Virtual Library: Logic Programming*.

The database contains 10,487 anchor texts; of these, forty 25-item samples were evaluated. The database contains 1,865 header texts, of which twenty-nine 25-item samples were evaluated.

In checking the sampling validity, some links were found to have no text, probably because an image was used. The processing discarded images. Based on the samples, one would estimate the predicted percentage links with no text as 5.6. The actual percentage was found to be 5.77.

Results

Results are presented as statistics and as distribution graphs, followed by an evaluation of tagged text informativeness.

Links and link text statistics

For the 553 analyzed URLs, the pages generally contained more links than headers. The average number of links was 18.96 with a standard deviation of 23.68. The largest number of links was 325; the smallest, zero. The mean was 16.97 with a standard deviation of 14.56; this excluded the 8 pages with the largest number of links, which ranged from 99 to 325. The maximum of the remaining pages was 76. In general, pages related to logic programming contain a significant number of links to other pages, making the question of the informativeness of these links a reasonable one to ask.

The average number of headers was 3.37 with a standard deviation of 5.03. The largest number of headers was 42; the smallest number was again zero. The mean was 2.91 and the standard deviation was 3.46; this excluded the 9 pages with the largest number of headers, which ranged from 27 to 48. The maximum of the remaining pages was 21. The presence of a number of headers in almost all Web pages on logic programming indicates that authors do organize their pages into sections of some kind.

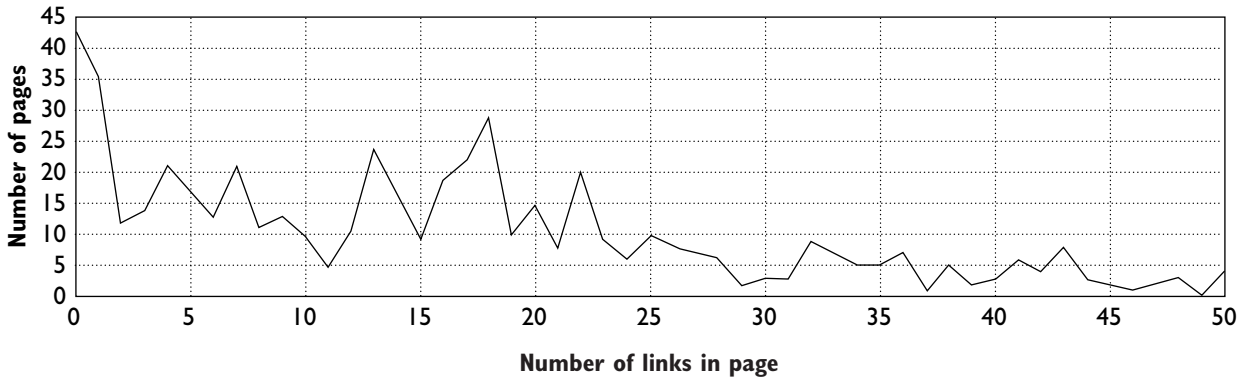

Figure 1. The number of sampled pages, 533, and the frequency distribution of those pages having a given number of links.

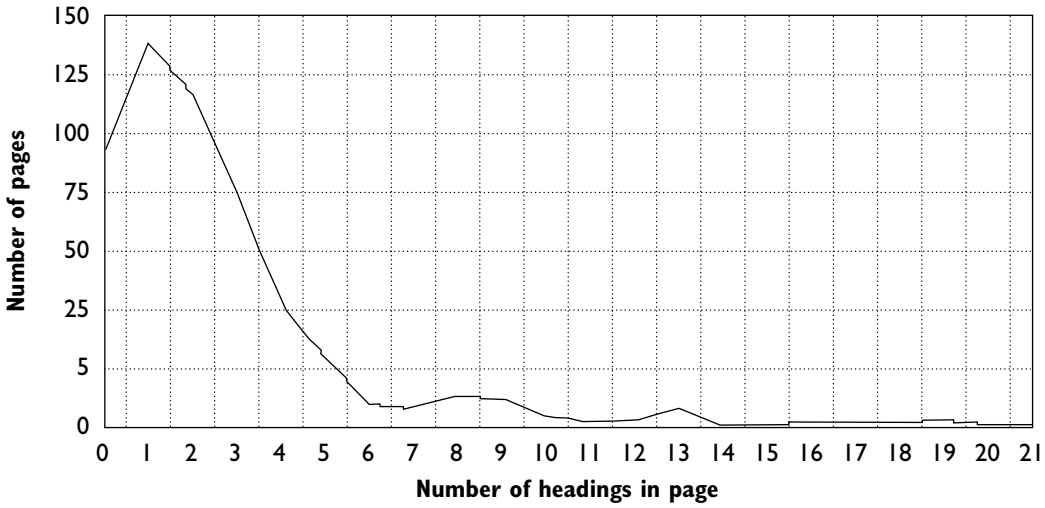

Figure 2. Frequency distribution of 544 sample pages having a given number of headings.

Web-page links appeared about five and a half times more frequently than headers, as Figure 1 shows. The graph in Figure 1 was based on the 533 pages with 50 or fewer links; the other 20 pages had between 57 and 325 links.

Figure 2, which shows the frequency distribution for all headings, is based on the 544 pages containing 21 or fewer headings. The other 9 pages had between 27 and 42 headings.

Tag content evaluations

Tables 2 and 3 itemize the evaluations of tagged texts from samples of 25 pages.

Link texts. The "no text" column in Table 2 took into account that we did not try to evaluate images, which were discarded by the program that created the databases of tagged texts.

The more helpful the text, the longer it generally was—for example, a text phrase such as *knowledge representation and reasoning* was more helpful than *ILPNet2*. It's likely that there would be some tension between longer link texts and readability of long text; however, one study has shown that examining surrounding text is a valid approach to increasing the informativeness of tagged text.[2]

Header texts. Web authors frequently use headers such as *Introduction* and *Discussion* for document organization rather than to indicate content. Not surprisingly, evaluators found headings generally less useful than link text, as Table 3 shows.

The average percentage of very helpful headings was much lower (4.55) than the very helpful link texts (17.6). Again, the more helpful the heading,

Table 2. Informativeness evaluations of link texts. Sample size was 25 text items.

	Statistical element			
Informativeness	Mean	Standard deviation	Mean %	Example text
No text	1.40	1.28	5.60	
No help	4.03	2.62	16.10	previous
Little help	4.23	2.58	16.90	FAQ
Helpful	10.95	3.05	43.80	Standard ML of New Jersey
Very helpful	4.40	2.73	17.60	ALP Newsletter Archive

Table 3. Evaluations of heading texts.

	Statistical element			
Informativeness	Mean	Standard deviation	Mean %	Example heading
No help	13.76	2.68	55.03	Exercises
Little help	6.69	2.56	26.76	Program Committee
Helpful	3.41	1.55	13.66	What is logic programming used for?
Very helpful	1.14	1.09	4.55	An analysis of refinement operators in inductive logic programming

Table 4. Text retrieval comparison by heading, link, and keyword.

Keywords	URLs retrieved by link	URLs retrieved by heading	Pages containing keyword
Logic, logic	420	265	548
Prolog, prolog	29	120	183
clause	0	1	76
backtracking	1	2	24
declarative	8	16	73
goal	2	6	51
grammars	2	1	31

the longer and more explicit it generally was. It's unclear how to expand header context to improve less helpful text. Choosing a fixed number of words near the heading in the Web page, for example, does not seem likely to be helpful.

Metatags and comments. Most HTML editors insert metatags automatically. A small sample of metatagged and comment texts (225) yielded no semantic content in 84 percent of them. This somewhat counterintuitive result is because most metatagged and comment information inserted automatically by Web page composition tools is more concerned to detail the tool used than provide metainformation on the content itself. Because only 6.5 percent were rated helpful or very helpful, most as index or keyword data, I did not pursue this evaluation further.

Tagged-text retrieval

To determine the effect of using keywords in the tagged-text approach in comparison with using keywords in a full-text search, I used single keywords from the list prepared for the tagged-text database. A text was retrieved

- if the keyword appeared in a Web page header,
- if the word occurred in anchor text corresponding to a Web link, and
- if the word occurred in the body of the Web page text.

I selected keywords for their relevance to *logic programming*. Table 4 summarizes the results. The database used was constructed from the original 553 pages used earlier.

Aside from the words *logic*, *prolog*, and possibly *declarative*, the retrieval rate based on heading or link text is low compared to retrieval by whole-text search. Since few pages used metatags to give keyword references, this suggests that, with few exceptions, most keywords do not appear in headings or link text. Indexes based on just these two HTML tags and keywords in metatags are therefore unlikely to be entirely satisfactory.

Conclusions

This exploratory work shows that 61percent of Web page link text was rated helpful or very helpful in indicating semantic contents. Text retrieved on the basis of all a document's headers and tags showed that the resulting abstracts, or outlines, are most useful for Web pages in which headers contained informative content. In general, authors of Web pages should be encouraged to make use of informative headings and link text in their Web pages.

Future Work

The initially favorable results from the study of link text suggest that this is a fruitful direction for future research. Effectiveness of headings as proxies for URL contents is a particular concern since more than a quarter of the headings we evaluated were of little help in that regard. I experimented with document headings as part of a proxy for the document itself, hoping that this would be similar to expanding the window around link text. In fact, the results were quite mixed and more research is required.

The results are documented at http://www.sju.edu/~jhodgson/Webproj/compare.html. For each Web page analyzed, this site lists the URL title (if any), the headers within the page (in order of appearance), and the text associated with links, from other pages within the database. These proxies' informativeness varies widely. Many proxies, however, would adequately guide users wanting to know if the page is worth pursuing further.

Acknowledgments

Thanks to the anonymous referees who commented on earlier article drafts, to the students and faculty in the Mathematics and Computer Science Department at Saint Joseph's University who filled out the evaluation questionnaires, and to the INRIA Rocquencourt for hospitality during part of this work.

References

1. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," available online at http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm (current as of 7 Dec. 2000).
2. C. Lawrence and L. Giles, "Context and Page Analysis for Improved Web Search," *IEEE Internet Computing*, vol. 2, no. 4, July/Aug. 1998, pp. 38-46.

Jonathan Hodgson is a professor of mathematics and computer science at St. Joseph's University in Philadelphia. His research interests include logic programming and its application to Web page analysis. Hodgson received a PhD in mathematics from Cambridge University, England. He is a member of the IEEE Computer Society.

Readers can contact Hodgson at jhodgson@sju.edu.