HTML Tags as Extraction Cues for Web Page Description Construction

Timothy C. Craven
The University of Western Ontario, London, Ontario, Canada

craven@uwo.ca

Abstract

Using four previously identified samples of Web pages containing meta-tagged descriptions, the value of meta-tagged keywords, the first 200 characters of the body, and text marked with common HTML tags as extracts helpful for writing summaries was estimated by applying two measures: density of description words and density of two-word description phrases. Generally, titles and keywords showed the highest densities. Parts of the body showed densities not much different from the body as a whole: somewhat higher for the first 200 characters and for text tagged with "center" and "font"; somewhat lower for text tagged with "a"; not significantly different for "table" and "div". Evidence of non-random clumping of description words in the body of some pages nevertheless suggests that further pursuit of automatic passage extraction methods from the body may be worthwhile. Implications of the findings for aids to summarization, and specifically the TexNet32 package, are discussed.

Keywords: HTML; extracting; metadata; summarization; computer software; World Wide Web.

Introduction

The background to the research reported on in this paper is the author's ongoing programme aimed at developing a computerized abstractor's assistant (Craven, 1988 ; Craven, 1991; Craven, 1993 ; Craven, 1996 ; Craven, 1998a). In addition to a simple word processor and other general writer's tools, the package integrates tools, such as an automatic extractor, related specifically to the task of summarizing. Apart from the author's own work, Paice (1994) has given a list of desirable features for such a package. A hybrid system, in which some tasks are performed by human abstractors and others by software, appears to be an appropriate short term goal, since purely automatic abstracting methods (Endres-Niggemeyer, 1998 ; Paice, 1990 ; Paice, 1994 ; Pinto & Galvez, 1999) do not show immediate promise of totally superseding human effort.

With a view specifically to applying the computerized assistant to summarizing Web pages, this development work has lead to an investigation (Craven, 2000) into how people and organizations in fact summarize their own Web pages. An underlying assumption has been that author created descriptions will will reflect features that authors and other users may consider desirable. It is not assumed thereby that all descriptions reflect the same desirable features, nor that all authors consider the same features desirable, nor even that all descriptions contain only desirable features.

The investigation has focused to date on descriptions stored on the pages themselves in meta tags. More exactly, the descriptions studied are those that appear as values of the "content" attribute of

Section Editor: Amanda Spink

HTML "meta" tags the value of whose "name" attribute is "description". Variations on this pattern, such as making "description" the value of the "http-equiv" attribute or "dc.description" the value of the "content" attribute have been found to be very rare and have therefore been ignored. "Meta" tags are supposed to be found in the "head" element in HTML, but some ill-formed pages may include them without specifying a "head" tag.

Other aspects of the content of Web pages have been studied by various researchers; for example, page layout of home pages (King, 1998); characteristics of anchors (Haas and Grams, 2000); informetric measures (Almind & Ingwersen, 1997); links to e-journals and their articles (Harter and Ford, 2000); the feasibility of using implicit structure in identifying differences automatically (Rahardjo and Yap, 2001); rates of change over time (Koehler, 2002). Little investigation has been done into descriptions in meta tags. Turner and Brackbill (1998) have, however, reported results of a small experiment that showed that addition of a description did not improve retrievability of Web pages on Infoseek and Altavista; similar results have been reported for these two search engines and five others by Henshaw and Valauskas (2001); Drott (2002) has noted the extent to which descriptions and keyword meta tags are used in the sites of 60 Fortune Global 500 companies.

Another article by the author (Craven, 2001) has reviewed advice given in both printed and Web-based sources on the function, content, structure, and style of meta tag descriptions. Some of the main recommendations found in that review are that the description can be used for an abstract; that it should be no more than about 200 characters (about 30 words or one sentence, the typical maximum length for an abstract of a letter to the editor); that (like an abstract) it should be concise and not be the same as the title.

The author's studies to date have revealed from 27% to 38% of Web pages containing meta tag descriptions, depending on the method of sampling. Some of the descriptions greatly exceed typical length guidelines. From 7% to 10% duplicate exactly phrasing found in the text of the body and most repeat some words and phrases. Meta tag keywords are slightly more likely to appear nearer the beginning of a description than nearer the end. Unlike standard abstracts, descriptions often use noun phrases instead of complete sentences.

A question raised in the course of these studies related to the specifics of repetition of wording from the text of the page body. Were there particular parts of the body which were more likely to be used than others? If so, suitable cues might be employed to extract such parts automatically from pages without descriptions and so assist in the writing of descriptions for those pages. Another question was the extent to which passages so extracted might be better or worse candidates for extraction than the page titles or any keywords included in meta tags on the page.

Since the aim was to develop generalized tools and the subject matter and language of Web pages various greatly, it was decided to look for cues in the HTML tags, which tend to occur over and over again independently of the specific topic or level of the page.

Methodology

Selection of Tags for Consideration

It was thought to be more effective to concentrate first on tags that were actually more common in Web pages, since this would both make it easier to obtain statistically significant results and yield information that would be helpful in composing descriptions for a larger proportion of pages.

Four sample sets of URLs were employed for this purpose. All had been found a year earlier to correspond to pages containing meta-tagged descriptions, though it was not expected that this would still be the case. The first two sets (1a and 1b) had originally been obtained directly using the Yahoo! random

page service (this was referred to as level one); each member of the next set (2) had been obtained by following a random link from a page returned by the Yahoo! service (this was referred to as level two); and each member of the last set (3) had been obtained by following a random link from a page returned by the Yahoo! service and then following a random link from that page in turn (this was referred to as level three). The continued division into levels, rather than combining into a single set, was considered potentially worthwhile because earlier studies had shown some significant differences between levels, with level 1 apparently containing, for example, a much higher proportion of home pages.

Specially written software (using the NEWT HTML ActiveX control) was used to attempt to access the page corresponding to each URL and to record what tags occurred in the page. Success ranged between 79% and 85%, with 272 pages being accessed in set 1a, 344 in 1b, 169 in 2, and 184 in 3.

Results were broadly similar regardless of set. Table 1 shows the frequencies of the most common tags (occurring in at least 30% of pages in at least one set). It may be noted that some other well-known tags, including "ul", "ol", "li", and the "h" series, did not meet the minimum requirement set.

a	70%-78%
body	87%-95%
br	61%-71%
center	46%-51%
div	27%-34%
font	59%-69%
head	88%-97%
img	75%-78%
meta	80%-88%
p	69%-78%
script	28%-36%
table	61%-69%
td	60%-67%
title	88%-96%
tr	60%-67%

Table 1: Frequencies of most common tags

Pages in set 1b tended to use the most common tags least; this tendency seems to be connected with the greater use of frames in that set.

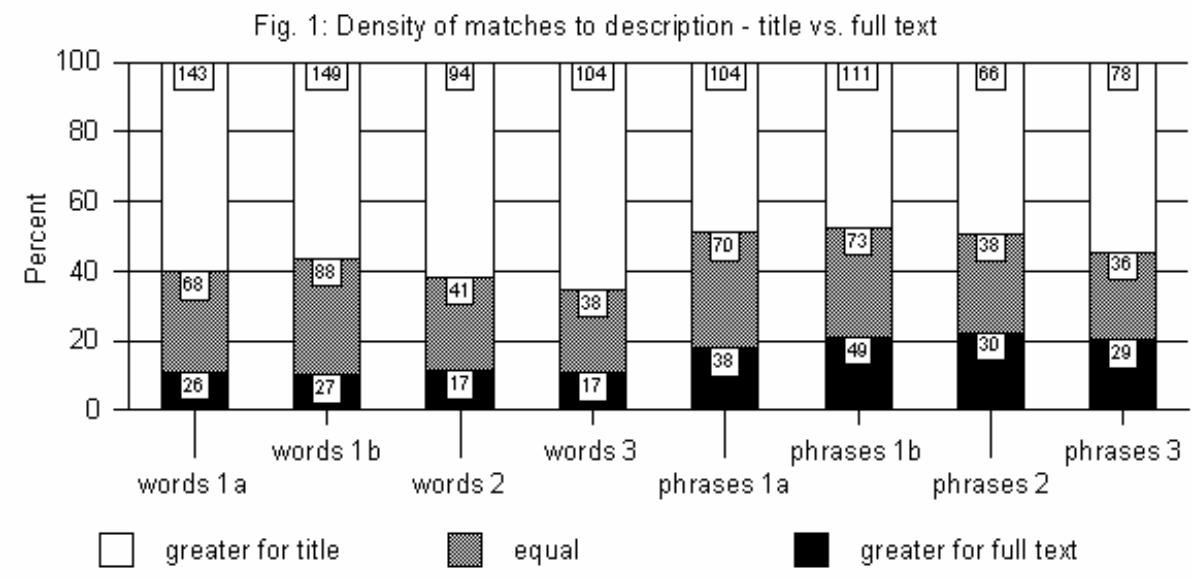
It was decided to restrict the initial investigation further to tags that would have well-defined portions of text with which they were associated. For the most part, this meant tags that are normally used with corresponding end tags (for example, "" with ""); an exception was the "meta" tag with the "name" attribute equal to "keywords", where the value of the "content" attribute could be considered the associated text. The decision to restrict to tags with well-defined associated text eliminated "br", "img","p", "td", and "tr"; in addition, "script" was eliminated because it does not normally have ordinary language text associated with it. That left only "a", "body", "center", "div", "font", "head", "meta" (for keywords), "table", and "title". From the point of view of enclosing visible text, "head" and "title" were considered virtually identical; so, only "title" was retained.

Since the "<body>" and "</body>" tags would normally enclose all of the visible text except the title, the "<body>" tag would not be analyzed in the same way as the other tags. Instead, the first 200 characters of "body"-tagged text would be treated as a special case.

Analysis of Description Word and Phrase Density

To estimate the value of the various tags as cues for extracting text suitable for writing descriptions or other summaries of the corresponding Web pages, two measures were used: the density of description words in the tagged text and the density of two-word phrases from the description in the tagged text. Obviously, these measures could be computed only for pages which actually contained descriptions; pages where descriptions were no longer present were therefore disregarded. For purposes of the latter measure, a two-word phrase was defined simply as any sequence of two words, where a word was any sequence of characters, outside an HTML tag, of the 26-letter alphabet and separated by one or more characters not in the alphabet. The use of such a simplistic definition was considered adequate, since applying a stop list in other research in the series had yielded no significant results.

As a standard of comparison, the density of description words and phrases in the visible text as a whole (minus the title) was employed. The aim was to see how often text tagged in a given way would perform

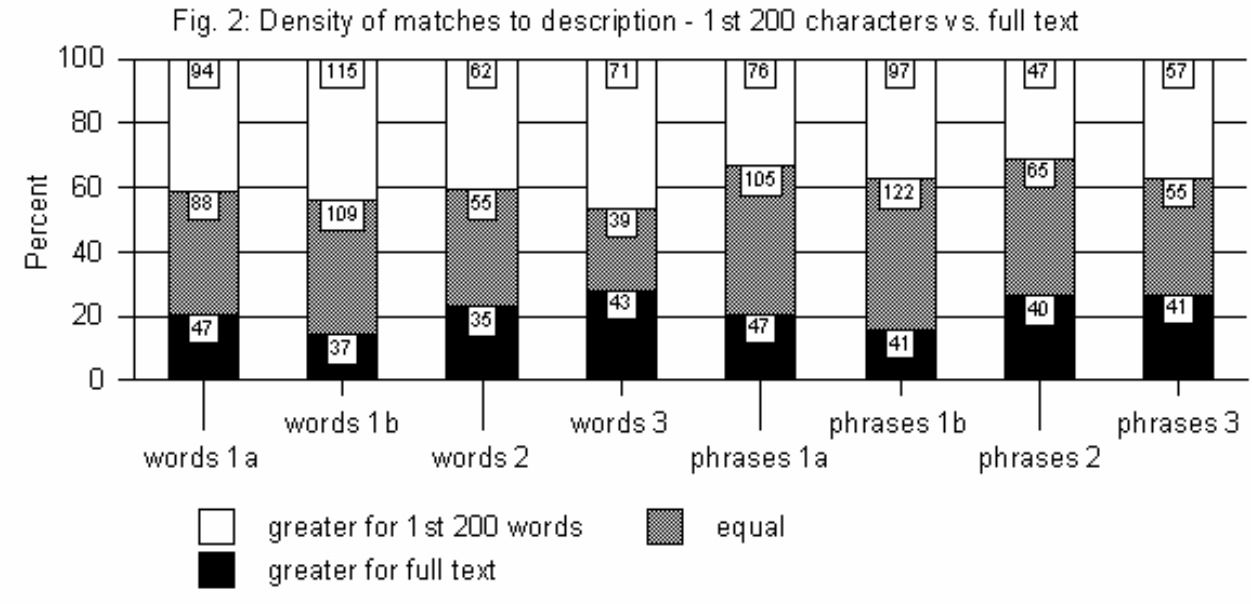


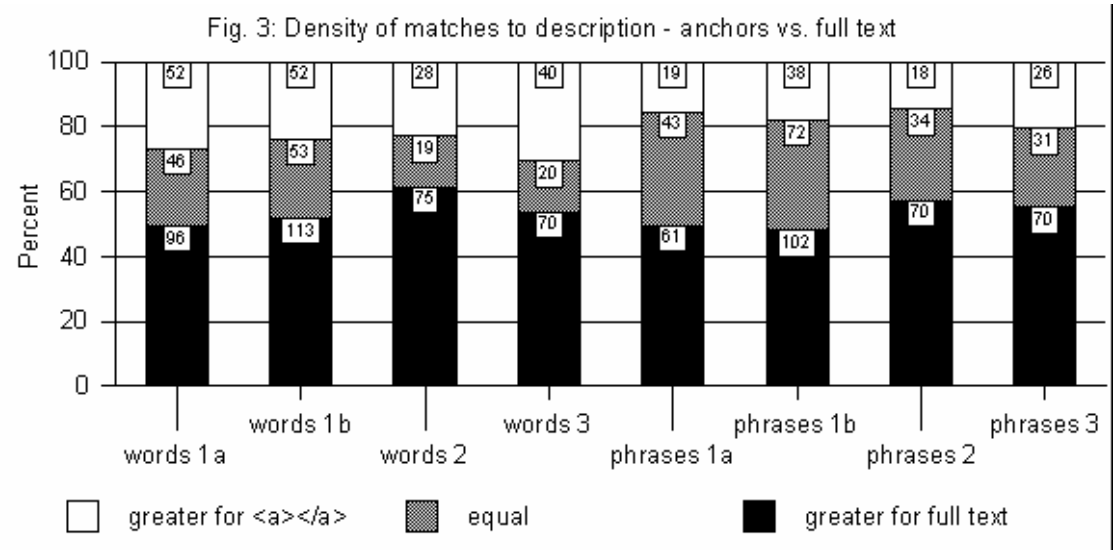
better than, as well as, or worse than, this "full" text in providing a rich source of wording and phrasing for the description. It was anticipated that tags more commonly associated with text with description word and phrase densities less than that of the full text should be assigned negative weights in any automatic extraction, while those more commonly associated with text with higher densities should be assigned positive weights.

Word and Phrase Matching Results

Titles performed best relative to the rest of the visible text in all four sample sets (Figure 1): about 60% better, 30% equal, and 10% worse for words ($p=0.0000$ for all sets using a binomial test of better versus worse) and about 50% better, 30% equal, and 20% worse for phrases ($p<0.0005$ for all sets).

The first 200 characters of the body also generally performed significantly better (Figure 2) ($p<0.0500$ at all levels for words and $p<0.0001$ for sets 1a and 1b, but $p>0.0500$ for phrases for sets 2 and 3). The "center" tag showed marginally significantly better performance for words in set 1a ($p=0.0215$) and set 3 ($p=0.0489$) and for phrases in set 1a ($p=0.0488$). Keywords showed a significant advantage for words in





all sets (p=0.0000), and for phrases in all (p<0.0500) except set 1a. The "font" tag showed a marginally significant advantage for phrases in sets 1a (p=0.0440), 1b (p=0.0235), and 3 (p=0.0184); for words, the advantage was significant only in set 3 (p=0.0086).

Text tagged with "a" performed significantly worse than the non-title text as a whole throughout (Figure 3) (p<0.0100 for words and p=0.0000 for phrases).

Tags that generally failed to show significant results in comparison with the non-title text as a whole were "b", "div" (p<0.0500 only for phrases in set 1b), and "table" (p<0.0500 only for phrases in set 1b).

A couple of examples will suffice here to illustrate patterns of matching (words also found in the description are capitalized).

(1) The longest centered (and table) text completely matching description phrasing in set 1a was

> RENTZ AGENCY INC IS A FULL SERVICE INSURANCE REAL ESTATE AND TRAVEL AGENCY SERVING MORRIS MINNESOTA AND THE SURROUNDING AREA SINCE WE HAVE BUILT OUR REPUTATION ON SERVING OUR CUSTOMERS FIRST IN SERVICE IS OUR GOAL

where the description is

> Rentz Agency, Inc. is a full service Insurance, Real Estate and Travel Agency serving Morris, Minnesota and the surrounding area since 1954. We have built our reputation on serving our customers. "First In Service" is our goal.

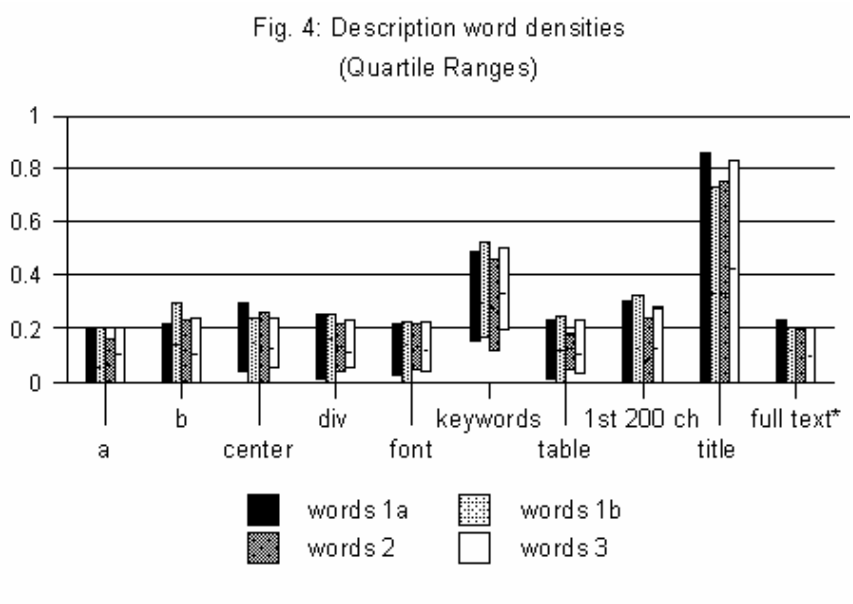
(2) The best phrase matching for anchor-tagged text in set 1a was 56%, for

> WEBMAIL netaus ONE OF AUSTRALIA S FASTEST GROWING INTERNET SERVICE COMPANIES DELIVERING QUALITY HIGH SPEED INTERNET CONNECTIONS AND RELATED SERVICES TO HOME AND BUSINESS CUSTOMERS click TO enter

where the description reads

> Webmail Internet Solutions are one of Australia's fastest growing Internet Service companies, delivering quality, high-speed Internet connections and related services to home and business customers

Figure 4 gives a general picture of how density of description words varied over each set of pages in the various types of text.



As is alluded to in the qualification "(Quartile Ranges)", each bar represents the range of values of the middle 50% of the cases by density (this method of display, omitting the bottom and top quarters, is intended to give a better picture of typical values than would be obtained by simply showing the entire range from lowest to highest). It can been seen, for example, that, on the one hand, titles were generally the densest source of description words, but, on the other hand, at least one quarter of all titles for every set contained no description words; likewise, it can be seen that, although keywords did not perform quite as well as titles, they showed less variation and so might be deemed more reliable where present.

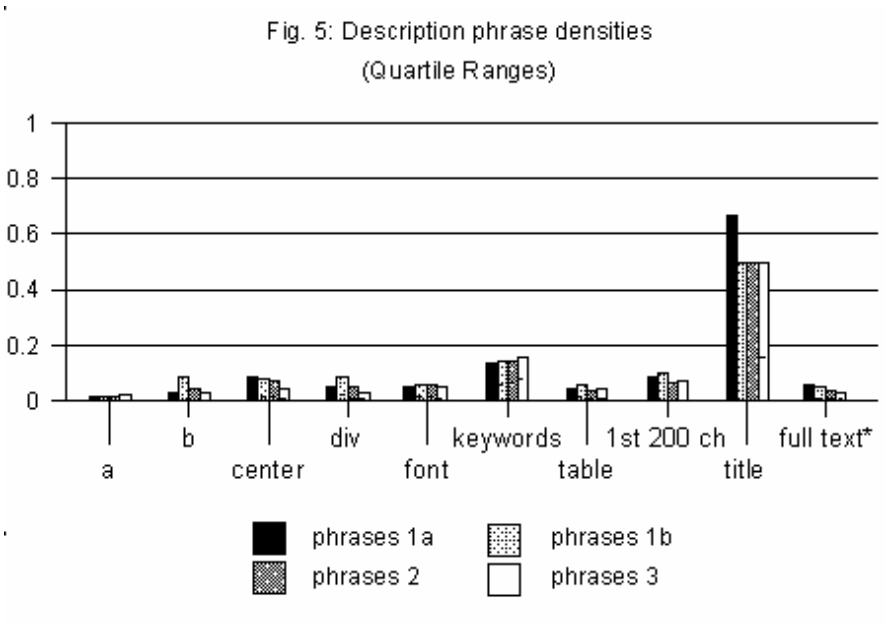


Figure 5 gives a similar general picture of the variation of density of description phrases.

Most notable is the fact that description phrase density is generally very low, except in titles; even the meta tag keywords, which perform quite well for individual description words, are not much better than other parts of pages as a dense source of description phrases.

*The actual set of pages successfully retrieved varied slightly from tag to tag, depending on server response. Values for full text in Figure 4 and Figure 5 are taken from the pages successfully retrieved in comparing titles with descriptions.

Clumping of Description Words in Full Text

The use of tags to identify potential description wording in the main text of Web pages had proved relatively unsuccessful. An obvious question arising was whether this result might have been improved upon by the use of more suitable cues. Before embarking on a search for such cues, however, it was de-

cided to take a step back and ask whether description wording in fact did tend to concentrate in any parts of the main text.

To answer this question, a measure of clumping of description words was needed. Inspiration was taken from the work on clumping measures of Bookstein and Raita (2001), but the specific measure adopted for the present study was somewhat different. The measurement process began by computing the mean distance, in words, between description-word matches in the text. This computation turned out to be simpler than it first seemed: it did not actually involve determining the distance between each pair of matches; instead, running totals of words, match positions, and number of matches so far could be used to achieve the desired result. The mean match distance value had then to be compared to the mean distance between words in the text. Analysis showed that the mean distance between pairs of words in a text of n words is equal to $(n+1)/3$, where $n>1$. In order to have a positive measure of clumpiness, the ratio of the mean match distance value to the mean word distance was subtracted from one, giving the formula

$$\text{clumpiness} = 1 - 3m / (n+1)$$

where m is the mean match distance and n is the number of words in the text.

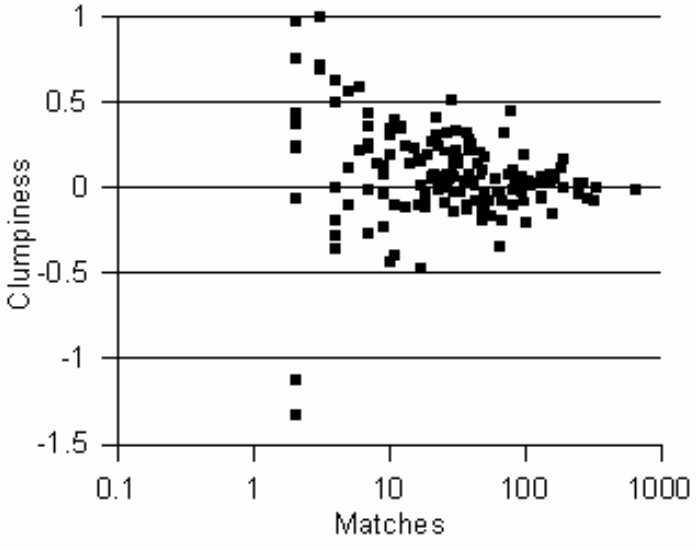
Some general properties of this measure may be noted. First, a clumpiness of zero indicates a random expected proximity of matches. Positive values indicate various degrees of clumping; negative values indicate that matches tend to be separated more than would be the case on average by random placement. At its maximum, the clumpiness measure approaches one; this is typically true for a very long text with very few matches, all concentrated in one place. A text with many matches, even if these are all together, will show a more moderate clumpiness; for example, a text of 20 words in which the first ten consist only of matches and the second ten consist only of non-matches will have a clumpiness of 0.48. The measure lacks symmetry in that its minimum is not -1: it can have lower values for long texts with very few matches that are unusually widely separated; for example, for a text of 100 words of which only the first and last match, the value would be -1.94. Finally, the measure will show greater random variation for smaller numbers of matches than for larger numbers of matches.

Figure 6 shows a plot of clumpiness against number of matches for description words in pages from set 1a. A number of the general features of the measure mentioned above can be observed. Somewhat more subtle is the fact that significantly more of the data points fall above the zero line than below it (97 to 54, p=0.0006), indicating that description word matches in the texts do in fact tend to clump.

If it is assumed that all negative clumpiness values are the result of random variation and that a corresponding random variation will account for an equal number of positive clumpiness values, about 43 (=97-54) pages, or 28%, remain in which clumping can be attributed to non-random factors.

The concentration of key information in certain parts of the text may certainly be one such non-random factor, suggesting that a further search for cues for identifying such parts of page texts may still be of value, at least for a substantial minority of pages. Other non-random factors may be at play, however. For example, the two words "of" and "the" tend somewhat to occur together

Fig. 6: Clumping of description words in text



in English text; a positive clumpiness value due to the occurrence of the expression "of the" in both the description and the main text would thus be the result of a non-random factor, but not of one necessarily relevant to methods for selection of key passages.

The tendency of description words to clump was in fact somewhat weaker and not statistically significant for the other sets taken individually. Set 1b showed 87 data points above zero to 77 below (p=0.4823), suggesting about 6% non-random clumping; set 2, 55 to 46 (p=0.4262), about 9% non-random; and set 3, 64 to 45 (p=0.0842), about 17% non-random.

TexNet32 HTML Filters

An immediate practical application of the findings of the present study with regard to HTML tags lies in the default HTML filter settings for the TexNet32 experimental abstractor's assistance package. These filters are normally applied to HTML documents imported by the user and specify various options for each tag: what beginning and ending text should be substituted, what weight should be assigned to the tagged text, whether a paragraph break should be inserted, whether the tagged text should be discarded, whether line breaks should be translated to paragraph breaks, and the text style (bold, italic, underline, strikeout). Of these options, it is that of the weight assigned that is significant to the present discussion.

Table 2 shows the default weight assignments for tags assigned weights in previous releases of TexNet32.

It should be noted that weights in TexNet32 are always assigned to whole paragraphs and not to smaller units of text. Thus, the weight of an HTML tag that applies to only a part of a paragraph is prorated over the paragraph. Some additional caution is therefore necessary in translating results from the present study to application in TexNet32.

The great majority of the tags in the table above did not occur sufficiently frequently in pages to be considered in the present study and may therefore be disregarded at this point: "blockquote", "code", "em", "h1", "h2", "h3", "h4", "kbd", "pre", "samp", "strong", "sub", "sup", and "u".

Results of the present study clearly validate the decision to assign the highest weight of 15 to the "title" tag, although, as already noted, the extent to which the title matches description wording is quite variable. The moderately positive weight of 3 given to the "center" tag would seem to be supported by the findings of the present study. The "font" tag, given a weight of zero in previous versions of TexNet32, might be given a slight positive weight, perhaps of 1.

The indecisive results for the "b" tag in the present study suggest that giving it either a significant positive weight or a significant negative weight would not be appropriate. The previous weight of 2 is likely too far on the positive side.

The "div" tag is a relatively newer addition to HTML and as such was not included in the original filter list. Results from the present study suggest that its omission (effectively giving it a weight of zero) would not in fact create difficulties in terms of weighting material imported from HTML format. The same conclusion would also apply to the relatively more established "table" tag.

Text tagged with the "a" (anchor) tag appears relatively less useful for automatic extraction. It is possible, therefore, that this tag should

A	1
B	2
BLOCKQUOTE	-1
CENTER	3
CODE	-1
EM	1
H1	10
H2	9
H3	8
H4	7
KBD	-1
PRE	-1
SAMP	-1
STRONG	2
SUB	-1
SUP	-1
TITLE	15
U	1

Table 2: Default weight assignments in previous releases of TexNet32

be assigned a negative weight in the TexNet32 filters. Some caution may be called for here, however: the "a" tag is typically applied to words or relatively short phrases, which may occur within a longer, highly indicative, paragraph. At least not assigning a positive weight may nevertheless be advisable, since the "a" tag is often applied to list items or other passages that would be translated into separate paragraphs in TexNet32. Since the "a" tag is perhaps most commonly used for links to other documents, text so marked may be more relevant to the construction of descriptions of those documents; the appropriate processing for this purpose is not covered by the current version of TexNet32.

The existing TexNet32 filter system takes no account of tag attributes. This means, among other things, that any meta tag keywords assigned to a page will be discarded upon import. (Meta tag descriptions can be extracted in TexNet32 by a separate mechanism.) In addition to the "meta" tag, other tags where recognition of attributes might prove of value include "font" (specifically, for the "size" attribute), "div" (which may have a "center" attribute), and "a" (for example, for whether the "href" attribute points to a location within the same page or directory).

Conclusion

Summary

In the pages examined, the following HTML tags were the most common: "a", the most common way of indicating hypertext links; "body", normally required in good HTML style unless a frameset is used; "br", or line break; "center", one way of centering text or other material; "div", the most common tag for marking layers; "font", the traditional tag for setting text appearance; "head", normally required in good HTML style; "img", the inline image tag; "meta", used for specifying keywords, descriptions, and various other meta data; "p", marking a new paragraph; "script", marking material in a scripting language; "table", "td", and "tr", marking a tabular structure and its main parts; and "title".

Text marked as "title" generally showed the best match in wording to the meta tag description, with a typical median value of 33% of words; but there was considerable variation, with many titles showing no matching at all. Meta tag keywords show similarly good matching of description words, with a typical median value of 30% of words, and less variability. Parts of the main text generally performed much less well. The first 200 characters, with a typical median of 10% of words, generally performed somewhat better than the main text as a whole. The "center" and "font" tags also showed some promise as positive cues, while the "a" tag appeared to be a candidate for a weak negative cue.

As general recommendations for automatic extraction to assist in description writing, the results might be taken as suggesting the following: extract the title, which is almost always available and of a reasonable length; extract also any keywords that do not duplicate title wording up to a reasonable limit; if the title and keywords are short, or at the user's request, extract also the first 200 characters of the body; if using other extraction methods, such as word frequency, as well, give a small positive weight to text with the "font" and "center" tags and a small negative weight to text with the "a" tag.

Although automatic extracts may be found useful by some writers, it should not be assumed that all description authors will wish to employ them. Instead, it is likely that different tools will suit different types of users. It is expected that individuals will use quite different approaches in writing descriptions, just as they have been reported to do in writing abstracts of scholarly articles (Endres-Niggemeyer *et al.*, 1991).

For those who prefer not to use automatic extracts, it may be worth noting that the advice given above may be transferred fairly easily to the manual realm: when constructing a description, look at the title; look at meta tagged keywords up to a reasonable limit; if desired, look at the top bit of text on the page; pay somewhat more attention to centered text and passages emphasized by different fonts, and some-

what less attention to passages with many links to other pages. None of these suggestions, of course, appears particularly surprising.

Further Research

The present study used the internal page descriptions, stored in meta tags, as the standard of comparison for estimating importance of parts of web pages for summarization. Many pages lack such descriptions. Nor is it clear that, even where present, such descriptions are particularly good; they are certainly not particular consistent in content or form. Inconsistencies and other defects have also been demonstrated in published author abstracts (Pitkin et al. 1999).

A parallel study, therefore (Craven, 2002), has made use of external descriptions, found in some pages announcing themselves as lists of "links" and "resources" (most such pages do not contain descriptions as such, but only titles or very brief phrases with associated links). Results are broadly similar, but there are some interesting differences. For example, a number of external descriptions include evaluative comments, and, in some sets, almost all contain URLs. Some properties of external Web-based descriptions have also been studied by Wheatley and Armstrong (1997). Amitay (2001) developed a tool called SnipIt to extract descriptive passages with URLs from Web pages and another tool called InCommonSense to select from among these the "best" descriptive passage for each URL.

The method used has focused only on exact duplication of words or phrases. Further research might look at the effect of applying stemming algorithms or thesaurus lookup. Some preliminary experimentation with the latter has, however, proved somewhat unpromising, on account of the great variety and specificity of the vocabulary of Web pages combined with the relative brevity of the descriptions.

A final point that may be noted is the fact that Web page coding practices remain a moving target. Certain techniques, such as the use of frames, may wax and wane. Coding certainly seems to have become more complex over time, as more and more work is done via powerful page editing software packages such as FrontPage, rather than by direct text editing. Insertion of passages containing non-HTML code, is now not uncommon. It may therefore be of interest to revisit the results of the present research in a few years time to see whether they are still valid. For instance, one might wish to discover whether the World Wide Web Consortium's long-standing deprecation of such handy tags as "font" and "center" in favour of more complex alternatives causes the former to fall out of use in the long run. To take another example, if application of XML restrictions to HTML should catch on, elements such as "li" would be required to have corresponding end tags and could thus be more readily included in the type of study described here.

Acknowledgments

Research reported in this article was supported in part by the University of Western Ontario Office of Research Services with funds provided by the Natural Sciences and Engineering Research Council of Canada.

The extensive assistance of research assistant Emmett Macfarlane in data gathering is also acknowledged.

References

Almind, T.C.. & Ingwersen, P. (1997). "Informetric analyses on the World Wide Web: methodological approaches to 'Webmetrics'". *Journal of Documentation*, 53 (4), 404-426.

Amitay, E. (2001). *What lays in the layout*. Retrieved from http://www.ics.mq.edu.au/~einat/thesis/

Bookstein, A., & Raita, T. (2001). Discovering term occurrence structure in text. *Journal of the American Society for Information Science and Technology*, 52 (6), 476-486.

Craven, T.C. (1988). Text network display editing with special reference to the production of customized abstracts. *Canadian Journal of Information Science*, 13 (1/2), 59-68.

Craven, T.C. (1991). Algorithms for graphic display of sentence dependency structures. *Information Processing and Management*, 27 (6), 603-613.

Craven, T.C. (1993). A computer-aided abstracting tool kit. *Canadian Journal of Information Science*, 18 (2), 19-31.

Craven, T.C. (1996). An experiment in the use of tools for computer-assisted abstracting. In S. Hardin (Ed.), *ASIS '96: Proceedings of the 59th ASIS Annual Meeting 1996 (Volume 33), Baltimore, Maryland, October 21-24, 1996* (pp. 203-208). Medford, New Jersey: Information Today.

Craven, T.C. (1998a). Human creation of abstracts with selected computer-assistance tools. *Information Research*, 3 (4), paper 47. Retrieved from http://InformationR.net/ir/3-4/paper47.html

Craven, T. (1998b). TexNet32 - WWW filters. In *Texnet32*. Retreived from http://instruct.uwo.ca/gplis/677/texnet32/wwwnet32.htm

Craven, T.C. (2000). Features of DESCRIPTION META tags in public home pages. *Journal of Information Science*, 26 (5), 303-311.

Craven, T.C. (2001). 'DESCRIPTION' META tags in locally linked Web pages. *Aslib Proceedings*, 53 (6), 203-216.

Craven, T.C. (2002). External descriptions of Web pages: their features and their relationships to Web page elements. *Libri*, 52 (1), 36-47.

Drott, M.C. (2002). Indexing aids at corporate websites: the use of robots.txt and META tags. *Information Processing and Management*, 38 (2), 209-219.

Endres-Niggemeyer, B. (1998). *Summarizing information*. Berlin: Springer.

Endres-Niggemeyer, B.; Waumans, W.; Yamashita, H. (1991). Modelling summary writing by introspection: a small-scale demonstrative study. *Text*, 11 (4), 523-552.

Haas, S.W., & Grams, E.S. (2000). Readers, authors, and page structure: a discussion of four questions arising from a content analysis of Web pages. *Journal of the American Society for Information Science*, 51 (2), 181-192.

Harter, S.P., & Ford, C.E. (2000). Web-based analyses of e-journal impact: approaches, problems, and issues. *Journal of the American Society for Information Science*, 51 (13), 1159-1176.

Henshaw, R, & Valauskas, E.J. (2001). Metadata as a catalyst: experiments with metadata and search engines in the Internet journal, First Monday. *Libri*, 51 (2), 86-101.

King, D.L. (1998). Library home page design: a comparison of page layout for front-ends to ARL library Web sites. *College and Research Libraries*, 59 (5), 458-465.

Koehler, W. (2002). Web page change and persistence - a four-year longitudinal study. *Journal of the American Society for Information Science & Technology*, 53 (2), 162-171.

Paice, C. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26 (1), 171-186.

Paice, C.D. (1994). Automatic abstracting. In A. Kent & C.M. Hall (Eds.), *Encyclopedia of Library and Information Science*, volume 53 (supplement 16) (pp. 16-27). New York: Dekker.

Pinto, M.; Galvez, C. (1999). Paradigms for abstracting systems. *Journal of Information Science*, 25 (5), 365-380.

Pitkin, R.M., Branagan, M.A., & Burmeister, L.F.(1999). Accuracy of data in abstracts of published research articles. *JAMA*, 281 (12), 1110-1111.

Rahardjo, B., & Yap, R.H.C. (2001). Automatic information extraction from Web pages. In *SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA* (pp. 430-431). New York: ACM.

Turner, T.P., & Brackbill, L. (1998). Rising to the top: evaluating the use of the HTML meta tag to improve retrieval of World Wide Web documents through Internet search engines. *Library Resources and Technical Services*, 42 (4), 258-271.

Wheatley, A., & Armstrong, C.J. (1997). Metadata, recall, and abstracts: can abstracts ever be reliable indicators of document value? *Aslib Proceedings*, 49 (8): 206-213.

Biography

Timothy C. Craven is a Professor in the Faculty of Information and Media Studies, The University of Western Ontario. He has published some 50 articles since 1976 in the areas of computer-assisted indexing, summarization, and thesaurus construction and is the author of String Indexing (New York: Academic Press: 1986). He currently teaches courses in the Graduate Library and Information Science program in information systems and technology, Internet information services, and subject analysis and thesaurus construction.