Web Page Classification Based on Web Page Size and Hyperlinksand Website Hyperlink Structure

Denis L. Nkweteyim

School of Information Sciences, 135 N. Bellefield Avenue, University of Pittsburgh, Pittsburgh, PA 15260

Abstract: This paper presents a new metric, Page Rank × Inverse Links-to-word count Ratio (PR × ILW), used in classifying web pages as content or navigation. The metric combines web page size and number of hyperlinks on a page, and the page rank metric based on website topology, to compute the new metric. We present a theoretical basis for the new metric, and the results of a web page classification study, which show that the new metric, when combined with the links-to-word count ratio of web pages, accurately classifies the pages into the two categories.

1 Introduction

Web page classification is required in several situations. For example, in *web content mining* and *web usage mining* [1, 2], web pages are grouped based on their content (web content mining), or classified as content or navigation pages (web usage mining). In hypermedia personalization systems, it is often necessary for the system to implicitly deduce items of interest to the user without explicitly requiring the user to rate items, that way reducing the cognitive load placed on the user. Automatic classification of web pages is useful in such systems.

There has been much research on parameters that can be used to automatically deduce items of interest to users in personalization systems. Reading time, for example, has been extensively studied as a factor that determines interesting pages in hypermedia (see for example, [3-7]). Chen, Park, and Yu [8] proposed *Maximal Forward Reference* (MFR) as an implicit indicator of user interests in web pages as they navigate a web site. This approach is based on tracking users' forward references (pages not in the set of pages already seen) and backward references (pages already in the set of visited pages), as they navigate a web site. Other implicit interests indicators that have been researched include: *edit wear and read wear* [9], and e*xamination, retention and reference* [10, 11].

Our research involves the design of hyperlink recommender systems based on past user navigation behaviors. The work involves mining of frequent user access patterns from web user access logs, and making recommendations based on these patterns. We use a number of heuristics to minimize errors as individual user sessions are extracted from the web logs (see [2, 8, 12-14] for a discussion of heuristics used), and the MFR heuristic [8] to extract transactions within the session data. We aim to

L. Nkweteyim D. (2005).
Web Page Classification Based on Web Page Size and Hyperlinksand Website Hyperlink Structure.
In *Proceedings of the 5th International Workshop on Pattern Recognition in Information Systems*, pages 225-233
DOI: 10.5220/0002577102250233

discover transactions of lengths between 2 and 5 URLs and mine these transactions for association rules that show the correlation relationships between user navigation behaviors and the pages they find interesting. The choice of desirable transaction lengths was motivated by research [15], that suggests that most users follow between two and five hyperlinks before reaching a page of interest.

Using MFR alone however, we found out that there were several transactions much longer than 5 URLs. We assumed that these long transactions contained *hidden* content (or recommendable) pages, which the MFR heuristic was not able to find. The challenge therefore, was to find a heuristic that could be used to make plausible guesses on what these hidden content pages were. Our approach combined the ratio links count:term count of web pages and the topology of the web site to compute a number that determined whether a page within a very long transaction should be classified as content or navigation, and hence determine new transaction boundaries.

2 Parameters Used to Determine Content and Navigation Pages

Based on Web Page Characteristics. A content page is a page with information that could be of interest to a user as she browses a web site. In general, web page contents could be text, graphic, or multimedia. For the purpose of this research though, only text was treated as content because text can be much more easily manipulated. The size of a web page can be considered equal to the number of terms found on the page. The larger the page size, the more likely it is that the page is a content page. In a hyperlink recommender system, the objective may be to recommend interesting content-rich pages to the user. A web page also typically has one or more hyperlinks. In general, the larger the number of hyperlinks on a web page, the more likely it is that the page is a navigation page.

Because a web page usually comprises text and hyperlinks, most web pages exhibit both navigation and content properties, and one or more metrics other than page size and the number of hyperlinks present, are required to correctly classify them. Besides, other factors may influence the classification of web pages (for example, the links to word count ratio, LW). Intuitively, the larger the value of LW, the more likely it is that the page is a navigation page, and vice versa for content pages. It should be noted though, that some web pages with high LW values could be content rich; likewise, some pages with low LW values may have too little content.

Based on Web Site Topology. In a navigation task a user follows hyperlinks the labeling of which suggest to the user that she is moving to pages that will meet her current information needs. Hyperlink labels are conceived by the web site designer to serve as local cues that users process in making judgments on which hyperlinks to follow. In foraging theory [16-19], these cues are called information scent[1], and could serve as an additional page classification parameter. In this research, we measured information scent using a variation of Google's *PageRank* algorithm.

---

[1] Information scent characterizes how users—like organisms that use scent to determine where go next—evaluate the utility of their actions to lead them to target information. The greater the scent, the more likely is the user to access the information.

PageRank *PR*, [20, 21] is a number that Google uses to determine the importance of a web page, and is one of several parameters used to determine the ranking of pages returned by the Google search engine. The PR algorithm assumes that a hyperlink from one page to another is a vote from the former to the latter. The importance of a page is determined by the number of votes it receives, and the relative importance of the pages casting the votes. A page also loses some of its page rank based on the number of outgoing links on the page. The following equation is used to determine the page rank of *Page A*:

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + ... + PR(T_n)/C(T_n)) \quad \textbf{(1)}$$

where, $PR(T_I)$ is the rank of page *I*; *C(TI)* is the number of hyperlinks on page *I*; *PR(TI)/C(TI)* is the contribution of page *I* to the rank of page *A*, if there is a link from page *I* to page *A*; and *d* is a damping factor, usually set to 0.85. In this work, PR was computed through an iterative process, with the page ranks on the right hand side of the equation initialized to *1/N*, where *N* is the number of web pages in the web site, and *d* set to 0.85.

In a navigation task, PR can be considered to be reflective of the web site designer's view of the way the site should be navigated: web pages with many incoming links are deemed more important than web pages with few incoming links. The problem with page rank in the domain of web navigation is that it is based solely on the network of links between web pages; it does not predict correctly how real users navigate the web site, and it says nothing about the content or size of pages. Hence, a page with several links pointing to it is likely going to be judged important, even if its content is minimal (and hence, not recommendable), while a page with several outgoing links is likely to lose much of its page rank, even if its content is important.

Combining PR and LW. Working with the intuition that a web page with a high page rank, but with little content is not as worth recommending, as a page with a low page rank, but with much content, we sought to combine PR and LW to get a single metric to use to classify web pages.

Let ILW represent the *inverse links-to-word count* ratio (i.e., 1/LW). Table 1 suggests that PR × ILW is a better metric to use to determine the content rating (how likely it is that a page is a content page) of a page than PR alone. For example, the third row of the table represents pages with low LW values (i.e., high ILW values, and hence high content rating), and large incoming:outgoing links ratio (i.e., high PR). The corresponding PR × ILW is equivalent to a large increase in the PR score since both PR and ILW are large. On the other hand, for the fourth row, which corresponds to web pages with high PR and low ILW, there is only a small increase in the PR × ILW score when compared to PR. A similar argument can be made for all the other rows of the table.

3 Experiment

The objective of this experiment was to determine how useful the LW and PR × ILW scores of web pages can be in the classification of the pages as content or navigation.

Subjects. The subjects comprised three Information Sciences graduate students who were very familiar with the web site of the School of Information Sciences, University of Pittsburgh (SIS): two fifth year and one third year.

Method. The web site of SIS (http://www.sis.pitt.edu) was used to test the new metric. A crawler was used to search for the hyperlinks found on the site's HTML pages, and the corresponding pages parsed to remove tokens related to page presentation, leaving only the contents. For each page, the SMART stop list [22] was used to remove commonly used words and the Porter stemming algorithm [23] to stem each term. High and low frequency terms (occurrence frequency respectively greater than 25%, and less than 1% of the web pages) were also removed. The web site topology was then constructed using an adjacency-lists graph representation.

Table 1. Illustrating the moderating effect of ILW on PR in the PR × ILW metric

Outgoing Links	Incoming	LW	Content	PageRank	PR × ILW
L	L	L	H	M	M_{++}
L	L	H	L	M	M_{+}
L	H	L	H	H	H_{++}
L	H	H	L	H	H_{+}
H	L	L	H	L	L_{++}
H	L	H	L	L	L_{+}
H	H	L	H	M	M_{++}
H	H	H	L	M	M_{+}
L = Low ; H = High; M = Moderate; + = small increase; ++ = large increase					

Fig. 1. Screenshot of one of the web pages that subjects classified. The top panel contains directions to subjects, the middle panel the page currently being classified, and the bottom panel page URLs and their ratings

Next, the page contents and site topology were used to determine PR and PR × ILW for each HTML page. The URLs of 4 web pages were then randomly assigned without replacement to each of the following 9 groups: PR_{high} PR_{med}, and PR_{low} (PR of the page is among the top, middle, and low third respectively of page ranks); ILW_{high}, ILW_{med}, ILW_{low}, for the ILW metric; and PR × ILW_{high}, PR × ILW_{med}, PR × ILW_{low}, for the PR × ILW metric. Finally, all 36 URLs were presented as hyperlinks on a web page to the subjects, who were instructed to view and rate the corresponding web pages on a scale of 0 to 10 using three different scales: a content scale, a navigation scale, and a dual content/navigation scale. Subjects' responses were collected in a web-based form containing their ratings data. Figure 1 shows a screenshot of one of the rating screens presented them.

4 Results and Discussion

Figure 2 shows the page ratings on the three scales by the subjects (S1, S2, and S3). As can be seen, the ratings were largely consistent, except for some ratings by S2, which we believed were errors (see the comments in the last column of the figure). For these erroneous cases, the relevant ratings were not used in computing mean ratings.

Figure 3 shows the variation of average user ratings with LW. As expected, web pages were generally classified as content (high content and dual scale ratings) for low values of LW (left of graph), and as navigation (high navigation scale ratings) for high LW (right of graph). Also as expected, navigation ratings were strongly negatively correlated with content ratings, while the content and dual scale ratings were strongly positively correlated, and so a single scale would be sufficient. Finally, Figure 3 suggests that LW alone is not sufficient to separate content from navigation pages. For example, there are a number of pages with relatively high LW ratio that subjects classified as content pages, but that using the LW parameter alone, would be classified as navigation pages.

Table 2 shows subjects' mean content scale ratings, and the page LW and PR × ILW values arranged first in order of LW (Columns 1–4), and then in order of PR × ILW (Columns 5–8). From the table, it can be seen that in general, web pages with a LW value below 0.05 or a PR × ILW value above 1.8 can be classified as content pages. We notice that using LW alone, content pages 35, 8, 7, 1, 2, and 4 would normally be classified as navigation pages because their LW values are above the minimum threshold of 0.05. However, because the corresponding PR × ILW values are above the 1.8 threshold for classifying pages as content pages, these pages are correctly classified.

	Content Scale			Navigation Scale			Dual Scale			Comments	
Item No	S1	S2	S3	S1	S2	S3	S1	S2	S3		
1	5	5	5	5	5	3	5	10	6	**d**	S2: dual scale rating inconsistent
2	8	3	4	2	7	4	7	2	5		
3	10	10	10	0	0	0	10	10	10		
4	9	7	7	1	2	7	9	6	5		
6	2	10	2	8	0	8	2	10	2	**a**	S2: Must have missed long list of hyperlinks at bottom of page
7	9	8	7	2	1	2	8	8	7		
8	8	8	4	3	1	6	7	7	4		
9	1	8	1	6	1	10	3	8	1	**a**	S2: Page comprises almost entirely of hyperlinks
10	9	2	9	7	2	8	8	2	9	**c**	S2: Content rating too low; page comprises mainly plain text
11	1	1	1	10	9	10	0	1	0		
12	10	7	9	1	1	1	9	7	10		
13	10	8	9	3	1	3	8	8	9		
14	8	8	7	7	1	4	5	8	7		
15	10	10	10	0	0	0	10	10	10		
16	9	10	10	2	0	1	9	10	9		
17	9	10	10	1	0	1	9	10	10		
18	5	10	6	8	0	4	4	10	5	**a**	S2: All ratings inconsistent; large table with many links and moderate amount of text
19	9	9	6	2	1	3	9	9	7		
20	9	9	8	2	1	2	9	9	8		
21	8	10	7	3	0	7	8	10	5		
22	7	10	9	2	0	2	8	10	9		
23	10	9	5	0	0	1	9	9	9		
24	10	9	9	0	0	1	10	9	10		
25	1	10	1	9	0	9	1	10	0	**a**	S2: All ratings inconsistent; large table with many links and moderate amount of text
26	1	10	1	9	0	10	1	10	0	**a**	S2: All ratings inconsistent; large table with many links and moderate amount of text
27	9	10	8	1	0	2	9	10	9		
28	10	5	7	2	5	7	9	5	5		
30	9	5	2	2	5	8	8	5	1		
31	10	10	10	0	0	0	10	10	10		
32	7	6	7	3	4	3	6	6	7		
33	5	10	2	9	0	3	2	10	3		
34	9	8	10	1	1	2	9	8	10		
35	9	9	10	1	1	3	9	9	9		

a – all 3 of subject's ratings not used in computing average score; **d** – subject's dual scale rating not used in computing average score; **c** - subject's content scale rating not used in computing average score

Fig. 2. Subject ratings of the web pages presented in the experiment. The commented items refer to cases where there were large discrepancies in subject ratings. Results of 3 of the 36 web pages that were presented to users are omitted because these pages were no longer linked to by the time the results were analyzed

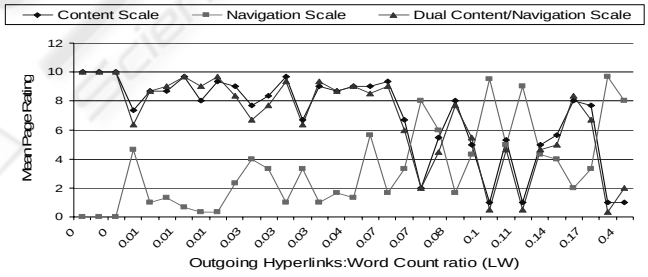

Fig. 3. Variation of page ratings with LW on the content, navigation and dual content/navigation ratings scales

Table 2. Combining LW and PR × ILW thresholds to classify web pages as content or navigation

Sorted by LW				Sorted by PR × ILW			
PageID	Mean Content Rating	LW	PR × ILW	PageID	Mean Content Rating	LW	PR × ILW
31	10.00	0.00	6.20	2	5.00	0.14	7.30
15	10.00	0.00	6.14	31	10.00	0.00	6.20
3	10.00	0.00	2.63	15	10.00	0.00	6.14
28	7.33	0.01	3.80	12	8.67	0.01	5.95
12	8.67	0.01	5.95	23	8.00	0.01	5.84
22	8.67	0.01	3.77	8	6.67	0.07	5.55
17	9.67	0.01	3.86	24	9.33	0.01	5.05
23	8.00	0.01	5.84	13	9.00	0.02	4.68
24	9.33	0.01	5.05	34	9.00	0.04	3.88
13	9.00	0.02	4.68	17	9.67	0.01	3.86
14	7.67	0.03	2.65	28	7.33	0.01	3.80
21	8.33	0.03	2.57	22	8.67	0.01	3.77
16	9.67	0.03	2.61	1	5.00	0.09	3.48
32	6.67	0.03	2.15	35	9.33	0.07	3.44
27	9.00	0.03	2.58	10	9.00	0.05	3.37
20	8.67	0.04	1.85	4	7.67	0.17	3.27
34	9.00	0.04	3.88	7	8.00	0.08	3.12
10	9.00	0.05	3.37	6	2.00	0.07	2.80
35	9.33	0.07	3.44	14	7.67	0.03	2.65
8	6.67	0.07	5.55	3	10.00	0.00	2.63
6	2.00	0.07	2.80	16	9.67	0.03	2.61
*18	5.50	0.08	1.36	27	9.00	0.03	2.58
7	8.00	0.08	3.12	21	8.33	0.03	2.57
1	5.00	0.09	3.48	32	6.67	0.03	2.15
26	1.00	0.10	1.28	20	8.67	0.04	1.85
*30	5.33	0.11	1.62	*30	5.33	0.11	1.62
25	1.00	0.11	1.21	11	1.00	0.34	1.51
2	5.00	0.14	7.30	*18	5.50	0.08	1.36
*33	5.67	0.14	1.08	26	1.00	0.10	1.28
*19	8.00	0.16	0.97	9	1.00	0.40	1.23
4	7.67	0.17	3.27	25	1.00	0.11	1.21
11	1.00	0.34	1.51	*33	5.67	0.14	1.08
9	1.00	0.40	1.23	*19	8.00	0.16	0.97

Content pages marked with an asterisk, "*" (18, 30, 33, and 19) were not correctly classified. We consider each of them in turn. Pages 18, 30 and 33 are borderline content pages (ratings close to 5), and they happen to have only 1, 2 and 1 incoming links respectively. Because these pages are linked to from so few pages, users have very few opportunities to reach them, and so not treating them as content or potentially recommendable pages is desirable.

Page 19 on the other hand was rated very highly as a content page, but both LW and PR × ILW failed to classify it as such. The reason for is that its PR, and thus PR × ILW, is too low to change its classification to content. The low PR resulted from the fact that the page is linked to from only one other page in the web site. Again, not classifying this page as a content page for recommendation purposes is desirable because users have very little opportunity to ever get to the page when they browse the web site.

Finally, navigation Page 6 which was correctly classified using LW is now incorrectly classified using PR × ILW.

5 Conclusion

We have shown that PR × ILW is useful in classifying web pages as content or navigation in applications where content pages are pages that a user browsing a web site may find interesting. This information can be used by a browsing agent that helps the user by observing her navigation behavior, comparing that behavior with those of past users of the web site, and recommending to her the content pages that the past users found interesting.

The PR × ILW metric works well in this regard because user browsing behaviors are usually constrained by the link structure of a web site (users typically navigate a site by following hyperlinks on pages on the site), and the metric exploits both this link structure and the properties of individual web pages.

It is worth noting that this classification scheme may not be as useful in other domains, for example where the main interest is in the contents, and not necessarily the connectedness of web pages.

References

1. Mobasher, B., R. Cooley, and J. Srivastava: Automatic Personalization Based on Web Usage Mining. In Communications of the ACM (2000) 142–151
2. Cooley, R., B. Mobasher, and J. Srivastava: Web Mining: Information And Pattern Discovery on the World Wide Web. In International Conference on Tools With Artificial Intelligence, Newport Beach, CA, (1997) 558–567
3. Morita, M. and Y. Shinoda: Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. In Seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval (1994) 272–281
4. Konstan, A..J., Miller, N., Maltz, D., Herlocker, L., Lee, R., Riedl, J.: GroupLens: Applying Collaborative Filtering to Usenet News. In Communications of the ACM, ACM Press, New York (1997) 77–87
5. Kim, J., W.D. Oard, and K. Romanik: User Modeling for Information Filtering Based on Implicit Feedback. In ISKO-France 2001, Nanterre, France (2001)
6. Claypool, M., Le, P., Waseda, M., Brown, D.: Implicit Interest Indicators. In Proceedings of the 6th International Conference on Intelligent User Interfaces, Santa Fe, NM (2001) 33–40
7. Kelly, D. and Belkin, N.: Reading Time, Scrolling and Interaction. In Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (2001) 408–409
8. Chen, M.S., Park, J.S., Yu, P.S.: Data Mining for Path Traversal Patterns in a Web Environment., in Proceedings of the 16th International Conference on Distributed Computing Systems (1996) 385–392
9. Hill, C.W., Hollan, D. J., Wroblewski, D., McCandless, T.: Edit Wear and Read Wear. In Proceedings of Conference on Human Factors and Computing Systems Monterey, CA (1992)
10. Nichols, D.: Implicit Rating and Filtering. In Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering Budapest, Hungary (1998) 31–36
11. Oard, W.D., Kim, J.: Implicit Feedback for Recommender Systems. In AAAI Workshop on Recommender Systems Madison, WI (1998)
12. Pitkow, J.: In Search of Reliable Usage Data on the WWW. In Proceedings of the Sixth International WWW Conference (1997) 1343–1355

13. Pirolli, P., Pitkow, J., Rao, R.: Silk from a Sow's Ear: Extracting usable structures from the Web. In Conference on Human Factors in Computing Systems (CHI-96) (1996)
14. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information Systems 1(1) (1999) 1999
15. Yan, T., Jacobsen, M., Garcia-Molina, H., Dayal, U.: From User Access Patterns to Dynamic Hypertext Linking. In 5th World Wide Web Conference. Paris (1996)
16. Larson, K. Czerwinski, M.: Web Page Design: Implications of Memory, Structure, and Scent for Information Retrieval. In CHI'98 Human Factors in Computing Systems. ACM Press (1998)
17. Pirolli, P.: Computational Models of Information Scent-Following in a Very Large Browsable Text Collection. In CHI ’97 Human Factors in Computing Systems Atlanta, GA. ACM Press (1997)
18. Pirolli, P., Fu, W.: SNIF-ACT: A Model of Information Foraging on the World Wide Web. In: Brusilovsky, P., Corbert, A., De Rosis, F. (eds.): User Modeling 2003 (2003)
19. Card, S.K., Pirolli, P., Van Der Wege, M., Morrison, J. B., Reeder, R.W., Schraedley, P.K., Boshart, J.: Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method For Web Usability, in Proceedings of the Conference on Human factors in Computing Systems, CHI '01. ACM Press, Seattle, WA (2001) 498–505
20. Rogers, I.: The Google PageRank Algorithm and How it Works. Retrieved September 5 2003 from http://www.iprcom.com/ papers/pagerank/
21. Craven, P.: Google's PageRank Explained and How to Make The Most of it. Retrieved September 5 2003 from http://www. webworkshop.net/ pagerank.html
22. Rocchio, J.: Relevance Feedback in Information Retrieval. In: Salton, G. (ed): The Smart Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Englewood, NJ.(1971) p. 313–323
23. Porter, M.: An Algorithm for Suffix Stripping. Program. 14 (1980) 130–137