An Initial Investigation on the Relationship between Usage and Findability

Colin Wilkie and Leif Azzopardi

University of Glasgow, 18 Lilybank Gardens, G12 8QQ, Glasgow, UK
{colin.wilkie, leif.azzopardi}@glasgow.ac.uk

Abstract. Ensuring that information within a website is findable is particularly important. This is because visitors that cannot find what they are looking for are likely to leave the site or become very frustrated and switch to a competing site. While findability has been touted as important in web design, we wonder to what degree measures of findability are correlated to usage. To this end, we have conducted a preliminary study on three sub-domains across a number of measures of findability.

1 Introduction

Designing and organizing a website or any other information space is a complex process where the goal is to create structure around content that enables users to complete their tasks (such as finding the relevant information and performing associated transactions) in a seamless and efficient manner. Typically, Information Architecture techniques are applied in an attempt to improve a site's usability and the overall user experience by optimizing "the structural design of an information space to facilitate task completion and intuitive access to content" [5]. Although there are numerous principles and heuristics that have been developed, Information Architecture as a discipline, lacks formal models for evaluating or predicting whether such techniques will improve the usability of a website. One of the key challenges when optimizing a website is to ensure that the content within the website is findable [5]. The premise is that if something is not findable, then it will not be used. Or conversely the more findable something is, the more likely it is to be used. In this poster paper, we perform an initial study to investigate whether it is possible to predict or estimate the usage of webpages within a website using measures of findability. While there is no measure of Ambient Findability, there are a number of measures that approximate or measure part of the findability of a page given the two major modes of access: browsing and retrieval. In terms of browsing, a number of navigatability based measures have been developed [6, 4, 7, 8], and with respect to retrieval a number of retrievability based measures have been developed [1]. In this work, we examine the correlation between existing measures and usage data collected from our institution's website. If there is a strong correlation then it would be possible to evaluate the structure of a web site without the need for usage data.

2 Measures of Findability

For the purposes of this initial study we shall the consider the following measures of navigability and retrievabilty. Navigbility, relates to how a user traverses the link structure of a site when browsing a site looking for a page. Navigability, as discussed by Zhang *et al* [7] is a measure of the ease with which a user can traverse a website by clicking through the link structure. PageRank [6], and Hits [4] each provide an indication of how important a page is, which we assume is a proxy, for how likely the page is to be visited. In the case of PageRank, which models a random surfer[2], we would expect this measure to be reasonably correlated with how users browse through a site. Hits provides a slightly different view with respect to how authoritative a page is, analogous to PageRank, along with the hubness, which provides an indication of how central a page is (here the presumption is that page that is more central is more likely to be visited.) On the other hand, Retrievability measures how easily a document can be retrieved using a search engine[1]. Two types of retrievability measures have been proposed, cumulative based measures and gravity based measures. Cumulative measures simply count the number of times a page appears in the top c documents [1]. Gravity based retrievability scoring is a more sophisticated method where a documents rank in the rank list determines the score it accumulates if it appears before the cut off c. Here a document receives a weight that is exponentially proportional to the rank, which is determined by a discount factor (β). This measure of retrievability accounts for the likelihood of a user visiting a document at a given/particular rank, and is thus more realistic of user interaction.

3 Methodology

To examine the relationship between usage and the aforementioned findability measures, we first performed a crawl of three web sites for which we have access to the usage data. Thirty days of usage data were collected during January and February, 2012 for three subdomains of the University of Glasgow (A with 348 pages and 27810 views, B with 761 pages and 107625 views and C with 1048 pages and 57384 views).

To compute the navigability scores, we extracted all hyperlinks from the pages and built a web graph to represent the structure of the site. We implemented PageRank and HITS where we set the mixing parameter *alpha* equal 0.85 as done in [6]. The PageRank scores, Authority Scores and Hub Scores where then unit normalized. To compute the retrievability scores for each page we followed the methodology reported in [1]. Thus, we extracted bigram queries from the content of the pages by performing stopword removal and then selected the top 2000 most frequently appearing bigrams within each site. The sites were indexed using Terrier's PL2[1]. The set of queries were issued to the index and the results returned used to calculate the retrievability scores for cumulative based retrievability where the cutoff was set to 5 (C5) and 10 (C10) and a gravity based retrievability score where the cutoff was 10 and the discount was 0.5 (G10b0.5).

[1] `www.terrier.org.uk`

The usage data was split into two parts: 15 days (past usage) and 15 days (future usage). We then compared the navigability and retrievability scores against the 15 days labeled as future usage, and to show how well past usage predicts usage, we used the 15 days of past usage data. This provides a benchmark and gold standard to achieve. Note that the purposes of the findability measures (or idea behind them) is to be able predict the usage of pages, when usage data is not available (i.e. the cold start problem, which happens when a site is redesigned or different site designs are being mooted.)

To determine the relationship between measures we used Pearson's Correlation and Kendal's Tau B (which takes into account ranks and ties). Statistically significant correlations are denoted by an * when the p-value is less than 0.05.

	Retrievability			**Navigability**			**Usage**	
Site A								
	C5	**C10**	**G10b.5**	**Hub**	**Authority**	**PageRank**	**Past Us.**	**Fut. Us.**
C5	-	0.98*	0.88*	0.07	0.01	0.02	0.14*	0.19*
C10	0.89*	-	0.85*	0.09	0.00	0.00	0.14*	0.19*
G10b.5	0.80*	0.75*	-	0.05	0.01	0.03	0.16*	0.19*
Hubs	0.23*	0.21*	0.17*	-	-0.02	-0.09	-0.00	0.00
Authority	0.41*	0.41*	0.34*	0.28	-	0.96*	0.58*	0.58*
PageRank	0.09*	0.06	0.04	0.39*	0.12	-	0.68*	0.69*
Past Us.	0.08*	0.09*	0.10*	-0.20*	0.01	0.07	-	0.97*
Fut. Us.	0.09*	0.10*	0.11*	-0.23*	-0.08*	-0.01	0.64*	-
Site B								
C5	-	0.98*	0.92*	-0.10*	0.07*	0.08*	0.03	0.05
C10	0.72*	-	0.88*	-0.11*	0.08*	0.09*	0.04	0.06*
G10b.5	0.58*	0.49*	-	-0.07*	0.07*	0.06*	0.02	0.04
Hubs	-0.05*	-0.04	-0.14*	-	-0.12*	-0.20*	-0.14*	-0.16*
Authority	-0.06*	-0.10*	-0.09*	0.43*	-	0.91*	0.16*	0.19*
PageRank	0.09	0.08	0.18*	-0.75*	-0.34*	-	0.20*	0.24*
Past Us.	0.16*	0.11*	0.29*	-0.45*	-0.01	0.52*	-	0.96*
Fut. Us.	0.21*	0.17*	0.28*	-0.32*	0.08*	0.42*	0.81*	-
Site C								
C5	-*	0.97*	0.85*	0.08*	-0.03	-0.03	0.04	0.01
C10	0.88*	-	0.81*	0.09*	-0.04	-0.04	0.04	0.01
G10b.5	0.70*	0.65*	-	0.06	-0.01	0.00	0.07	0.02
Hubs	-0.24*	-0.27*	-0.11*	-	-0.10*	-0.18*	-0.01	-0.02
Authority	0.35*	0.36*	0.26	-0.35*	-	0.92*	0.48*	0.41*
PageRank	-0.16*	-0.17*	-0.16*	-0.03	0.19*	-	0.57*	0.51*
Past Us.	0.26*	0.24*	0.29*	0.10*	0.27*	0.03	-	0.97*
Fut. Us.	0.20*	0.17*	0.23*	0.19*	0.23*	0.08*	0.71*	-

Table 1. Correlation table for sites

4 Results

Table 1 provides the correlations between each of the measures used in our experiment for the three sites. The values reported in the top right hand triangle are Pearson's correlation tests and the bottom left hand triangle reports the Kendall Tau correlations. Examining these results it is clear that the retrievability score are all highly correlated with each other, and these correlations are statistically significant and falls in line with the correlations reported in [1]. For the navigability measures, Authority and PageRank score are also highly correlated across all sites in terms of Pearson's correlation (around 0.9) however they are less correlated according to Kendall Tau (but the relationship is significant).

Looking at past and future usage we see that across all sites there is a significant and strong correlation between them. This result is to be expected and demonstrates how effective usage data is at predicting future usage (providing an expensive but strong baseline to approach). For the retrievability measures that we employed we note that the correlation with usage is quite low (0.01-0.2 on Pearson's), however, often this correlation was significant. If rank ordering is important then the Kendall Tau correlations show that low to moderate correlations are possible (e.g. up to 0.29* on site C). For the navigability measures that we test, the results are far more promising. Moderate correlations were observed between PageRank and Authority with respect to usage (0.4*-0.6* on Pearson's). However, for the Hub scores the correlation tended to be very low or close to zero.

5 Discussion and Future Work

Developing objective measures of findabilty is an important direction of research which would help in assessing how easily documents/webpages can be found (without going to the expense of collecting usage data). In this work, we have analyzed some simple measures of findability. Given the size of the sites analyzed it is not too surprising that navigability measures provide the best indication of page usage. Future work will look at adopting more sophisticated navigability measures, such as MNav [8] and using proximal clues as in the work on information scent [3] to make better estimates of usage. Also, we shall explore how to combine measures of navigability and retrievability to create a measure of "Ambient Findabiltiy."

Acknowlegements: This work is supported by the EPSRC Project, *Models and Measures of Findability* (EP/K000330/1).

References

1. L. Azzopardi and V. Vinay. Retrievability: An evaluation measure for higher order information access tasks. *CIKM*, October 2008.
2. A. Blum, T.-H. hubert Chan, and M. R. Rwebangira. A random-surfer web-graph model. *Workshop on Algorithm Engineering and Experiments and the Third Workshop on Analytical Algorithmics and Combinatorics*, 8:238–246, 2006.
3. E. H. Chi, P. Pirolli, and J. Pitkow. The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a web site. In *Proc. of the SIGCHI conference*, CHI '00, pages 161–168, 2000.
4. J. M. Kleinberg. Authoratative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604 – 632, September 1999.
5. P. Morville. *Ambient Findability*. O'Reilly Media, 2005.
6. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, Nov. 1999.
7. Y. Zhang, H. Zhu, and S. Greenwood. Website complexity metrics for measuring navigability. *QSIC*, 4, 2004.
8. Y. Zhou, H. Leung, and P. Winoto. Mnav: A markov model-based web site navigability measure. *IEEE Transactions on Software Engineering*, 33:869–890, 2007.