https://doi.org/10.48009/2_iis_2006_233-240

HOW RELIABLE ARE WEBSITE RANKINGS? IMPLICATIONS FOR E-BUSINESS ADVERTISING AND INTERNET SEARCH

Bruce W.N. Lo, University of Wisconsin-Eau Claire, lobw@uwec.edu
Rosy Sharma Sedhain, University of Wisconsin-Eau Claire, sharmar@uwec.edu

ABSTRACT

This study examines the similarities and differences among several publicly available website ranking lists to determine how reliable the lists are. Several metrics were used to measure the concordance and discordance of these lists. The effect of list size was investigated. Practical implications for e-business advertising costs and the review of search engine results are discussed.

Keywords: Website Ranking, Website Traffic, E-Business Advertising, Web Search, Search Engine Optimization (SEO)

INTRODUCTION

Comparing and ranking public websites have attracted much attention since the early days of Internet advertising [6]. Advertisers, advertising agencies, academic researchers, and consumers all have devoted a great deal of interest to website ranking [1, 9]. The reason for site ranking may be viewed from two angles. From the providers' perspective, the concern is to ensure that they can reach the intended audience by advertising on websites that are visited by the largest pool of potential customers. From the consumers' perspective, the concern is to get what they are looking for from the most reputable sources—presumably that coincides with the most highly ranked providers on the Web.

Thus e-businesses want to drive web traffic to their sites by advertising in the top ranking websites, while advertising agencies want to drive web traffic to their sites so that they can charge top advertising dollars. It is generally believed that those who rank high in a "search list" from a search engine or a portal site are more likely to be visited by information seekers and that high site traffic is the pre-condition to generating revenue-producing transactions [7]. So e-businesses aspire to have their websites listed towards the front of these lists. This is achieved by manipulating their website contents so that the sites become more "crawler-friendly," or by paying actual advertising dollars to the Web gateways to improve their exposure [11]. On the other hand, the consumers themselves and web users are also interested in being able to intelligently discriminate among the entries in the "search results" from search engines.

In practice one finds many different ranking lists on the Web. List providers use a variety of methods to rank the websites (all claim to be the most reliable!). Below is a partial list of the ranking providers and their URLs:

- *Alexa Internet, Inc. http://www.alexa.com;*
- *100BestWebsites, http://www.100bestwebsites.org;*
- *BtoBOnline NetMarketing 100, http://www.btobonline.com/netMarketing200/2003;*
- *ComScore Media Metrix, http://www.comScore.com;*
- *Nielsen, http://www.nielsen-netratings.com*
- *PC Magazine Top Websites, http://www.pcmag.com/article2/0,1759,1554010,00.asp*
- *Ranking.com http://www.ranking.com;*
- *Time's, http://www.time.com/time/2005/websites;*
- *Websearch, http://www.websearch.com;*
- *Web100, http://www.web100.com;*
- *World Hottest Sites, http://www.worldhot.com.*

How reliable are these rankings? Are there similarities between them? How much can we trust the ranks when it comes to charging advertising dollars? From the point of view of web search users, what precaution should we take to ensure that we do not leave out important sites on our search returns? Do we just examine the top 10 search results (as is the habit of many searchers) or should we go further?

In this paper, we shall consider these questions and attempt to discover what can be learned about justifying web advertising costs and user search expectations.

WEBSITE RANKING METHODS

Internet or e-business authorities employed many different methods to rank websites [2]. These methods may be grouped into 3 categories: activity-based criteria, reference-based criteria, and opinion-based criteria.

Activity-based criteria, also known as traffic-based ranking, are the best known among the three. It is

usually regarded as the most objective method. Here, websites are ranked according to the amount of activities that take place on the site. The site that attracts the most traffic or has the highest usage would rank at the top. Examples of this approach include Alexa, comScore, Nielsen, Ranking.com, and Websearch.

Measuring website traffic is by no means a simple task. It depends on what is being measured as well as how it is measured. With respect to site traffic alone, we can distinguish at least three different concepts: how many people visit a site, how long they stay on the site, and how frequently they return to the same site. All three are valid measure of website traffic, depending on what is the purpose of measurement. In web advertising literature, these three concepts are formally defined as follows [9]:

- Reach: percentage of unique visitors who visited a website at least once during a measurement period. This indicates the breadth of audience coverage.
- Frequency: average number times that a visitor (those who visited at least once) visits a website during a measurement period. This indicates the likelihood of repeat visits.
- Duration: average time (say in minutes) that visitors spend on a website. This indicates the "stickiness" of the website.

A major challenge is how to differentiate genuine pageviews from bogus visits generated by automated programs. Chu. [3] suggested using 3 indices to overcome this problem. Lee [10] pointed out that other factors, e.g., measurement periods, need to be considered when assessing web traffic.

There are two ways that traffic data may be collected: site-centric, where the unit of analysis is carried out at the web server site (hits, sessions, visits, and impressions); and user-centric, where data such as cookies, online registration, or transactions, are collected on the client's browser. Site-centric statistics are susceptible to manipulation by webmasters who are anxious to raise their own ranking. On the other hand, user-centric approach raises the issue of how representative are the sample statistics collected from a selected panel of users. All traffic statistics are subject to sampling bias. For example, in the Alexa and Websearch methods, users must first download a tool bar. Therefore, statistics are generated by a self-selected sample.

Reference-based criteria rank a website according to how frequently that site was cited by another in relation to a given search topic. Presumably the more frequently a site is cited, particularly by other sites that are regarded as subject matter authorities, the more important that link will be weighted. Citation links may be classified as
Self links – referenced by itself (recursive links);
Foreign links – referenced by another site;
Weighted foreign links – weighs the links according to the importance of the citing sites.

Google was the first to incorporate this link-popularity concept to its page rank (PR) algorithm [4]. Unfortunately this method is also susceptible to manipulation and abuse by Search Engine Optimization (SEO) [8].

Opinion-based criteria use the opinion of a panel of judges to rank the list of websites. The resulting rankings reflect the subjective judgment of the panel members. In a sense all ranking methods, including the traffic-based ones, rely on a panel of judges. But opinion-based criteria are more explicitly dependent on the subjective opinions of the judges, with little regards for objective data. Examples of opinion-based ranking lists include Time's 50 coolest websites, PC Magazine's top 100 websites, 100 Best websites, World's hottest, and Web100. Opinions may come from three sources:

- Opinion from a panel of experts
- Impression of potential customers
- Buying experience of actual customers (e.g. Bizrate.com or BBB.com)

To ensure that their ranking lists are accepted by web users, many of the ranking providers first establish their credibility with the intended audience by some other means. For example, Time and PC magazines are already acknowledged authorities in their respective fields. Naturally one can create new methods by combining two or more of the above into one, e.g., that proposed by Chen [2].

RESEARCH QUESTIONS AND METHODOLOGY

In this paper we shall report on the findings on 6 publicly available ranking lists. Three (Alexa, Ranking.com, & Websearch) use traffic-based ranking, while the other three (100 Best websites, Web100, and World's Hottest websites) use opinion-based ranking. The top 10 sites in each list are shown in Table 1. In our computation, the top 100 websites in each list were actually used.

The predominance of Yahoo is clear. It was ranked 1st in 4 of the lists, and 2nd in one. This is followed in second place by MSN, which was ranked 1st by one

and 2nd by 3 lists. Less obvious in 3rd position is Google. After these Big-Three the situation becomes less clear. In fact the opposite is true - a website which is ranked high in one does not guarantee that it will make it in the other top 10 lists. This prompts the rather curious question, "What does it really mean to be among the top ranking websites?" If a site ranks high in one list, how far down the list in the others, before it can be found?

Table 1. The Six Ranking Lists Used in This Study (only the top 10 sites are shown)

Alexa.com	Ranking.com	Websearch.com	100Bestwebsites	World Hottest	Web 100
yahoo.com	msn.com	yahoo.com	yahoo.com	yahoo.com	cnet.com
msn.com	yahoo.com	msn.com	google.com	msn.com	sutterfly.com
google.com	google.com	passport.com	amazon.com	aol.com	espn.com
ebay.com	passport.com	adwave.com	about.com	altavista.com	nationalgeographic.com
passport.net	passport.net	myspace.com	bartleby.com	lycos.com	evite.com
myspace.com	microsoft.com	google.com	groups.google.com	excite.com	amazon.com
amazon.com	yieldmanager.com	websearch.com	news.google.com	go.com	zdnet.com
microsoft.com	whenu.com	passport.net	cnn.com	xoom.com	ebay.com
google.co.uk	aol.com	yieldmanager.com	ebay.com	amazon.com	cnn.com
bbc.co.uk	ebay.com	ebay.com	download.com	cnet.com	etrade.com

Based on these preliminary observations, we formulated the following questions:

1. Are there similarities between these lists? Can we measure their degree of similarity? How different are the lists? Can we measure their differences?
2. How reliable are these rankings? If two lists claim to be based on similar ranking criteria, can we expect the lists themselves to be similar too?
3. If these lists are similar and are telling us more or less the same piece of information, why do we need so many different lists?
4. How much can we believe in the claims of a single ranking provider? Can we assess advertising charges according to the site rank? If a single site rank is unreliable, should advertising costs be associated with a range of ranks rather than a single ordinal number?
5. In using search engines to find information, how can we ensure that important information is not left out in our search returns? Would examining just the top 10 search results (as is the common habit of many) be sufficient? How far should we go down the list in our search results?

To find the answers to these questions, we conducted the following analyses:

Analysis 1: Membership Concordance of Ranking Lists. The first question to be asked is, to what degree do two given lists have common membership? Here we determine how many sites are common to both lists of size n (without considering the ordinal rank of each individual site). Our focus here is on the commonality of membership rather than the agreement on rank order. We shall compute this percentage for list size of n = 1 to 100. So for each pair of lists, we need to compute 100 values. Membership concordance is expressed as the percentage of commonality.

Analysis 2: Rank Correlation of Common Sub-lists. The next question is, do the ordinal ranks in the two given lists agree with each other? Are they correlated? To answer this, we first extract the subset of sites that are common to both lists. We then compute standard correlation coefficients, Kandall's Tau, τ, and Spearman's Rho, ρ [5], for each pair. Tau and Rho give us an indication whether the two rank lists are significantly correlated. We do this for list of size n = 1 to k, where k is the number of common sites in each pair of the top 100 sites in our data set.

Table 2. Values of k

Alexa–Ranking	47	Alexa–100Best	18
Alexa–Websearch	47	Alexa–Worldhottest	18
Ranking-Websearch	64	Alexa–Web100	7

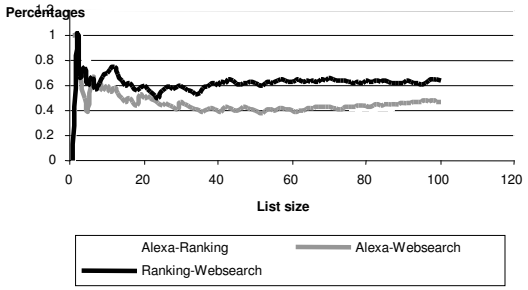
Analysis 3: Degree of Discordance in Rank Order. The final question is, to what degree do the 2 "common" sublists disagree with each other? In this investigation, we go beyond asking whether the two ranked lists have different order, but actually measure how much do the *original* ranks differ from each other between the two lists. Thus the metric here is the degree of discordance of rank order rather than the coefficient of rank correlation as in the previous analysis.

RESULTS AND DISCUSSION

Analysis 1: Membership Concordance of Ranking Lists. The results of membership concordance are presented in Figures 1 and 2. Figure 1 reports the pair-wise comparison of the three traffic-based ranking lists. Figure 2 depicts the pair-wise

comparison of the three opinion-based ranking lists. The X-axis shows size, n=1 to 100. The Y-axis is the percentage of commonality. The higher the percentage, the more they have in common.

Figure 1. Concordance of Traffic-ranked Lists



In Figure 1, we note that for small *n*, the membership concordance can be quite high (>60%). But as the sublist size increases, the membership concordance drops. For the pairs, Alexa-Ranking and Alexa-Websearch, the membership concordance hovers around 40% for large n. This means while the agreements among the top few sites are reasonable, these lists differ significantly when *n* increases. As *n* approaches 100, the two pairs differ by close to 60%.

Figure 2: Concordance of Alexz vs Opinion-based Lists

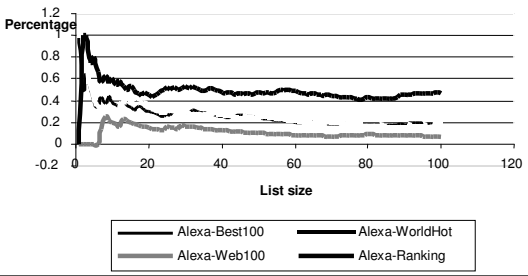


Figure 2 is even more interesting. We have chosen Alexa for our comparison with the opinion-based ranks. The obvious trend here is that membership concordance with the opinion-based ranks is a lot lower compared to the traffic-based pairs. When n=10, they all move below 40% for the pairs: Alexa-Best100, and Alexa-WorldHottest. The pair Alexa-Web100 gets even lower to less than 10%, indicating that Web100 is probably quite different from the other two.

Analysis 2: Rank Correlation of Common Sublists. Next we considered the correlation among the common sublists. To do this, we took two at a time. For Alexa & Ranking.com, we found k=47, i.e., there are 47 websites common to both top-100 lists. We want to determine whether there is any significant correlation between the two sets of ranks—not just common membership. We computed both the Kendall's Tau, τ and the Spearman's Rho, ρ. Figure 3 shows the τ values for the three: Alexa-Ranking (AR), Alexa-Websearch (AW), and Ranking-Websearch (RW). We select only these three from the traffic-based lists, because they have sufficient commonality for a meaningful analysis. Again, we did this computation for sublists of size n = 1 to k, (k=47 for AR & AW, and k=64 for RW). By way of contrast, we also plotted the critical values of τ^* at the p=0.005 level. A stringent p value was chosen because we wish to minimize Type 1 error.

The most interesting feature is that for small value of n, the correlation is not significant. In fact, fluctuations in the τ values indicate that, not until after n=16, can we say with some degree of certainty that the two lists are significantly correlated. This means we cannot make a general claim that the top 10 websites in one ranking list bear significant correlation to another list, even though both follow similar ranking methods. The discrepancy probably arises from the different sample sets they used to collect their data (different groups of users installed the tracking toolbar). This statement does not contradict the earlier observation. The τ statistics is a more stringent measure than membership test, because we now take into consideration the rank order of the sites and not just the commonality of the two lists.

We also computed Spearman's ρ for the same 3 lists with similar results. They have not been ported here due to space limitation.

There is a subtle point to be made here: even when the correlation is significant, these tests do not tell us whether the correlation is *strong* enough to determine advertising costs or *reliable* enough to include in a search return. The τ & ρ were calculated based on new ordinal positions in the common list. The rank differences in the *original* ranking lists were not considered. Two sites originally separated by a large rank distance, may end up right next to each other in the new common list. To overcome this, we proceed to the next analysis.

Analysis 3: Degree of Discordance in Rank Order. We are not aware of any standard statistics for this metric, so we developed our own. Suppose we want to compute the degree of discordance between the top-n websites of List A with the top-n websites of List B. Out of the *n* sites in each list, *k* of them are common to both. The degree of discordance, denoted by delta, Δ, is proportional to *1-k/n*, which is the

fraction of disjoint sites. We add 1 to avoid getting a zero when $k=n$:

$$\Delta = K_1 (2 - k/n)$$

where K_1 is a proportionality constant. This means that, the larger the proportion of non-overlapping websites, the large the discordance. Next consider the j^{th} common website in the list. Let r_{ja} be the rank of this website in the original List A, while r_{jb} be the rank of this website in the original List B. The j^{th} site discordance is proportional to the square of the distance between this two ranks, $(r_{ja} - r_{jb})^2$. These "individual" site discordances are then summed over all k to get the total discordance delta, Δ. Thus

$$\Delta = K_2 \Sigma_{j=1..k} (r_{ja} - r_{jb})^2$$

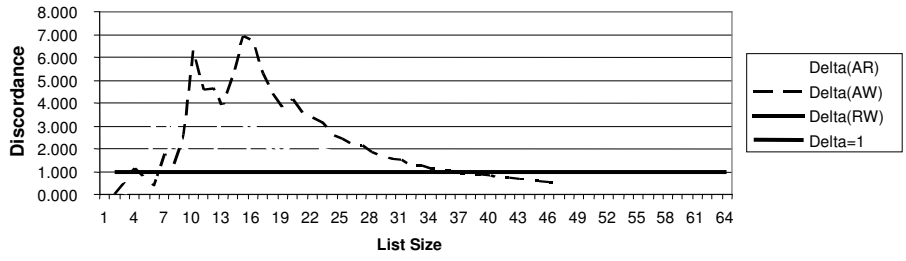
K_2 is another proportionality constant. We combine the two to create a final equation for delta, Δ,

$$\Delta = K (2 - k/n) \Sigma_{j=1..k} (r_{ja} - r_{jb})^2 / [n(n^2-1)]$$

where K is a new constant absorbing the other two. We introduced the normalizing factor $n(n^2-1)$ to emulate the Spearman formula. However, it is important to realize that this formula is not the same Spearman formula because the ranks are the original ranks and not the re-assigned ranks. With this formula, we computed delta values for the three pairs of web rankings: AR, AW, and RW, and plotted them as a function of the sublist size, n. The results are shown in Figure 4. Since we do not know the exact value of constant K, Δ/K rather than Δ itself, was plotted. Even though we do not know the exact critical value of delta, we can get around this by comparing the three series.

Figure 3: Kendal's Tau for Common Sublists

Figure 4: Degree of Discordance in Rank Order (Delta)



Firstly we note that delta fluctuates a lot for lower values of n, i.e. when the common list size is small. But as n gets larger, all 3 delta curves decrease monotonically to zero, i.e. the discordance asymptotically approaches 0. What it means is that the discordance becomes rather small when list size exceeds about 30. Although the decreasing trend was evidenced earlier at around n=22, but the behavior of the 3 curves appears to be rather unpredictable and varied wildly from each other for $7<n<22$. Thus 30 is a safer choice. This seems to indicate that we should be very careful if we want to claim that we have produced a list of "Top 10" or "Top 20" websites. It would appear that in this range of list size, the rank order is extremely sensitive to ranking methods used. It would be difficulty to claim with any degree of certainty that one ranking list is any more reliable than another list in this size range.

CONCLUSION AND IMPLICATIONS

We recognize that the results presented here are preliminary and that more analyses are needed to generalize these findings. But let us summarize what can be said thus far with respect to each of the research questions:

1. Yes, there are clear similarities and differences among the different ranking lists. We can certainly quantify and measure in some ways, the extent of their similarities and differences.
2. Using the metrics described in this paper, we can discern that ranking lists based on similar ranking methods, do exhibit a certain degree of similarity. However, the question "How reliable are the ranking lists?" is more difficult. Generally speaking, most lists agree on who are the top 3 sites, and also show a fair degree of agreement for lists of size 30 or more. But it is very difficult to demonstrate the reliability of ranking lists of size in the range of 10s or 20s.
3. Because the ranking lists are so sensitive to the method of ranking, it is essential that we obtain rank lists from a variety of sources so that they can be used to validate each other. Having ranking lists from different sources is valuable.
4. One should always be skeptical about the claims of a single ranking provider. If the advertising cost on a website is tied to a single rank position of that website, unless the site is one of those indisputable Web giants, one should be cautious about the advertising charges. In particular ranking positions in the teens and 20s are rather unpredictable and should always be questioned.
5. When looking for critical information on search results from search engines, one should go beyond the conventional first 10 entries to ensure no important information is omitted. In fact, based on the computations in this study, it is recommended that one should comb beyond the top 30 entries in the search results. This is because ranking lists of size 10 to the mid-20s are extremely sensitive to the ranking algorithm used and are thus unlikely to be reliable.

Based on this discussion we may derive two important pieces of practical advice.

- As an advertiser on the web, you should be critical of any website who wants to charge their advertising costs based on a single site ranking index. A more reasonable approach is to base web advertising costs on a range of ranks. Both web advertisers and web advertising agents should be aware that rankings lists of sizes in the 10s and 20s have the greatest margin for errors.
- When searching information on the Web using a search engine, one should review the list of search results with sufficient depth. Merely glancing through the first 10 entries is not sufficient. As a practical guide, one should go probably beyond the 30th entry.

REFERENCES

1. Berthon, P., Pitt, L. & Watson, R.T. (1996). The WWW as an advertising medium: Toward an understanding of conversion efficiency, *Journal of Advertising Research*, *36* (1), 43-54.
2. Chen, M., Tang, B. & Cheng, S. (2005). An Index System for Quality Synthesis Evaluation of B2C Business Website, *Proceedings of the 7th International Conference on Electronic Commerce ICEC'05*, August 15-17, Xian, China,.75-77.
3. Chu, K.K., Shen, T.C., & Hsia, Y.T. (2004). Measuring website popularity & raising designers' effort, *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics*, pp.4095-4099. Available http://ieeexplore.ieee.org/iel5/9622/30426/01401172.pdf?arnumber=1401172.
4. Craven, P. (2005). Google's PageRank explained—how to make the most of it? http://www.webworkshop.net/pagerank.html.
5. Daniel, W.W. (1978). *Applied Nonparametric Statistics*, Houghton Mifflin Company: Boston,
6. Hoffman, D.L. & Novak, T.P. (1996). Marketing in hypermedia computer-mediated environments, *Journal of Marketing*, *60* (3), 50-58.
7. Ilfeld, J.S. & Winer, R.S. (2002). Generating website traffic, *Journal of Advertising Research*, Sep-Oct, 49-61.
8. Jones, R. (2004). The murky world of search, from http://www.romjon.com/briefing/murky-search.php.
9. Lee, S. & Leckenby, J.D. (1998). An investigation of website ranking methods, *Proceedings of American Academy of Advertising*, 1998. Available http://www.ciadvertising.org/studies/reports/measurement/3ASuckkee.html.
10. Lee, S. & Leckenby, J.D. (1999). Impact of measurement periods on website rankings and traffic estimation: A user-centric approach, *Journal of Current Issues and Research in Advertising*, *21* (2), 1-10.
11. Weideman, M. (2003). Payment for increasing website exposure in search engine results, *Proceedings of The 5th annual Conference on WWW Applications*. Retrieved 3/13/2006 from http://general.rau.ac.za/infosci/www2003/Papers/Weideman,%20M%20Payment%20for%20increasing%20website%20exposure%20in%20sear.pdf.