Web Rings

Sebastian M. Maurer*, Bernardo A. Huberman and Eytan Adar

Xerox Palo Alto Research Center

Palo Alto, CA 94304

November 20, 2000

Abstract

We present a theory of surfing in web rings, common link structures observed on the World Wide Web. In particular, we determine an optimum ring size for surfing, and show how to organize rings in a hierarchy so as to maximize the probability that a user will find the information he seeks. Given the surfing characteristics of users, which can be obtained from usage data, we then determine how to organize web rings in a hierarchy.

*SMM thanks the Fannie and John Hertz Foundation for financial support.

1 Introduction

Hyperlinks are one of the most important features of the World Wide Web. They help organize collections of pages, create directories, or form communities. Many sites link to each other voluntarily [1], and monitoring the traffic that flows from one site to another is the basis for both affiliate programs and advertising banners [2]. Sites also frequently pay for incoming traffic, either simply for the number of click-throughs, or by distributing commissions on sales resulting from such visits. Finally, the existence of links between pages (reciprocal or not) is used as a basis for ranking web pages in search results [3, 4, 5].

One particular link structure that has become very widespread is the ring structure. In its most ideal form, sites link to a "next" and a "previous" site, in such a way that the member sites together form a ring. Most web rings are accessed through ring portal sites [6] which organize hundreds of thousands of rings into a hierarchy, and allow to search web rings by topic. In reality these rings are obviously much more complicated. Their quality often depends on the efforts of a ring administrator. Many rings provide shortcuts, some have broken links, and even when all the links exist some can be hidden and be more or less obvious to the user. Very often, each member site of a web ring will link to an index page which itself links to all the member sites of the ring. Some suggest using this spoke structure to define and detect web rings automatically [7], even though this criterion will of course miss the most general web rings.

Web rings turn out be very useful for casual browsing, since they provide

Figure 1: Snapshots of pages belonging to a web ring at www.bomis.com. The links from one member site to another appear in a separate frame. Thus individual sites do not have to include any html in their own pages as is required by certain other web rings

an efficient way to discover and explore smaller sites about a particular topic. Conventional search engines tend to be rather ineffective in this case. For example, it is often very difficult to come up with an accurate search query that captures the nuances of the pages that one is trying to discover. On the other hand, web ring portal sites make it easy to find a ring covering a given generic topic. It is then very easy to quickly follow the links in the ring to sample the main page on a few dozen related sites. This makes it possible to find great sources of information that one would otherwise not

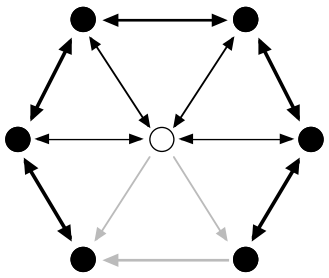

Figure 2: Diagram of a web ring with one directional links in gray.

even have known to search for. Note that for this purpose web rings have a slight advantage over directories in that it is possible to surf from one site to the next without returning to the directory every time.

Surprisingly, competing sites such as online stores or online auction sites often form web rings as well. While one might think that a user leaving one site for another is the first site's loss and the second site's gain, members of web rings recognize that helping a visitor find what he needs may ultimately bring him back at a later stage, especially if another member of the web ring links back. Thus, by belonging to a web ring a site can increase traffic and sales. And if users are going to leave a site anyhow, it is best if they leave it by going to a partner site [11].

One question asked by web ring administrators as new sites get added to a web ring, is how large such a ring should become. Obviously, a web ring that is very large might as well simply be a list that doesn't connect back to itself – a line instead of a loop. However, we will argue that a different consideration should limit the size of web rings. Since web rings are typically organized in a hierarchy, the size of a ring should be determined

by maximizing the probability that web surfers find the information they are looking for. This in turn determines whether one should choose one large ring over two or more smaller rings. In what follows, we first review the Law of Surfing [9] and its applicability to web rings. We then determine the optimum ring size by maximizing the probability that a page is found, given that a surfer must first find the correct ring, and then browse far enough in the ring to succeed.

2 The Law of Surfing

Web rings that are allowed to grow freely as people add their sites to the ring soon become very large. Some rings can have in excess of a thousand sites. The question then becomes when to split these rings up. Is it better to have one large ring, or several rings of smaller size?

A simple answer is given by the Law of Surfing [8], which tells us how far a user is likely to surf before he stops. We briefly summarize this model here. It is assumed that the pages encountered while browsing have a value or utility. In particular, the utility of the nth page is x_n. These values follow a random walk, in which the increments z_n are independent and identically distributed. Thus the value of the nth page is given by

$$x_n = x_{n-1} + z_n. \tag{1}$$

Intuitively, this means that if the current page is useful to the surfer, it is likely that the following page will also be useful. We also assume that the value of future pages is discounted at a rate ρ.

At each page, a surfer has the option of either continuing or stopping. We thus have an optimal stopping problem, whose solution can be determined by dynamic programming. Working backwards from the end, we have the recursion relations

$$\begin{aligned} V_N(x_N) &= x_N \\ V_n(x_n) &= x_n + max[0, \frac{1}{1+\rho} E[V_{n+1}(x_n + z_{n+1})]] \end{aligned}$$

where the expectation is taken over the random variable z_{n+1}. This means that the surfer will continue browsing provided that

$$E[V_{n+1}(x_n + z_{n+1})] > 0$$

Put another way, there is a threshold value x_n^* given by

$$E[V_{n+1}(x_n^* + z_{n+1})] = 0$$

such that it is optimal to continue only if $x_n > x_n^*$

With a finite (non-zero) discount rate $\rho > 0$, and a very large and sufficiently connected web, the threshold will be independent of n, so that $x_n^* = x^*$ is a constant. The number of pages visited before $x_n < x^*$ is then known, since in the continuous limit of the random walk specified by Eq. (1) one obtains true Brownian motion. In that limit the first passage times are distributed according to the inverse Gaussian distribution [10]. That is, the probability $p(x)$ that a surfer surfs x links is

$$p(x) \quad = \quad \sqrt{\frac{\lambda}{2\pi}} x^{-3/2} e^{-\frac{\lambda}{2\mu^2 x}(x-\mu)^2}.$$

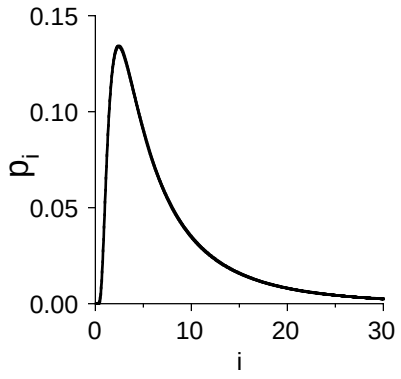

Figure 3: Inverse Gaussian distribution for $\lambda = \mu = 8$.

Huberman et. al. have matched this theoretical prediction with empirical usage logs obtained by tracking the number of pages surfed by visitors to AOL and several other web sites [9]. Thus, the two parameters of the inverse Gaussian distribution, λ and μ, can be determined experimentally.

Note that for large x, the inverse Gaussian distribution has an exponential tail, and for simplicity we will at first make the approximation that $p(x) = \frac{1}{\alpha} \exp(-\frac{x}{\alpha})$.

3 Optimum ring sizes

A fair comparison between two different ring sizes must consider the difference between one large ring with $2n$ sites, and two smaller rings with n sites each. Which is better? In order to answer this we assume there is one

particular site among these $2n$ sites that a surfer needs to find.

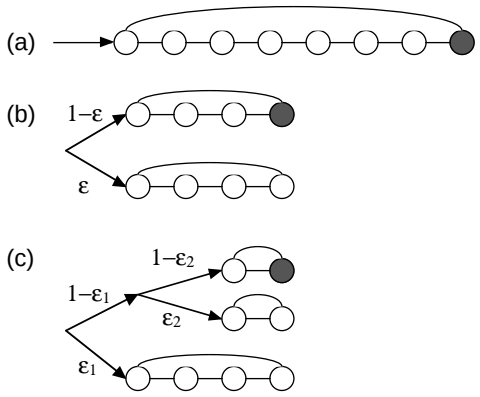

Figure 4: Three ways to organize eight web sites: either in a single ring, as two rings, or as three rings. The probability of finding a particular page is averaged over all possible locations. The best structure is then determined by the Law of Surfing and the probability of error ϵ at each branching point.

If the site is in a single large ring of $2n$ sites, and if we are given the expression for the Law of Surfing $p(x)$, we can compute the probability that the site will be found as a function of the position of the target site with respect to the starting point. The overall probability of finding the site is obtained by averaging over all possible positions for the site. For a ring with n sites, and in the continuous limit, this probability is given by

$$\frac{1}{n}\int_0^n \mathrm{d}y \int_y^\infty p(x)\mathrm{d}x$$

The second option is to have two rings of half the size. In this case, in order to find the desired site one must first chose one of the rings to browse. We assume that the two rings are somehow be organized by topic, such that

the correct ring will be chosen with probability $1-\epsilon$. Thus ϵ quantifies the error. Once we are in the correct ring, the Law of Surfing applies as before. Now, the probability that we find the desired page is given by

$$(1-\epsilon)\frac{1}{n}\int_0^n \mathrm{d}y \int_y^\infty p(x)\mathrm{d}x$$

Finally, even if we chose the wrong ring, it is still possible that after exhausting the first ring we will keep on searching the second ring. The probability that this happens is

$$\epsilon\frac{1}{n}\int_n^{2n} \mathrm{d}y \int_y^\infty p(x)\mathrm{d}x.$$

Putting everything together, we see that splitting one ring into two equally sized rings will be optimal (i.e. lead to a higher probability of finding the target site) provided the expected probability of finding the page in two rings is larger than the expected probability of finding it in one large ring:

$$\begin{aligned}(1-\epsilon)\frac{1}{n}\int_0^n \mathrm{d}y \int_y^\infty p(x)\mathrm{d}x & \\ +\epsilon\frac{1}{n}\int_n^{2n} \mathrm{d}y \int_y^\infty p(x)\mathrm{d}x \quad &> \quad \frac{1}{2n}\int_0^{2n} \mathrm{d}y \int_y^\infty p(x)\mathrm{d}x\end{aligned}$$

Substituting $p(x) = \frac{1}{\alpha}\exp(-\frac{x}{\alpha})$ and solving for n, we finally obtain

$$n \quad > \quad \alpha \log \frac{1-\epsilon}{1-2\epsilon}.$$

Thus, if the probability of picking the wrong ring is zero, we are always better off splitting the ring in half. In fact, in this case the sites are best

organized as a hierarchical tree structure, rather than in a ring. If, on the other hand, the probability of chosing the correct ring is $\frac{1}{2}$ or less, it is always better to keep the ring intact, since there is no way to satisfy the inequality.

Note that we obtain the same result if we solve the problem for the discrete case by explicitly summing the sums instead of approximating them by integrals.

What happens if we assume a different form for the Law of Surfing? In particular, there is a regime in which the inverse Gaussian distribution is approximated by a power-law with exponent $-3/2$. In this case, if proper care is taken in the integral approximation in order to avoid the divergence at zero, we get that a large ring should always be split provided that $\epsilon < \frac{1}{2}$. That is, as long as the sites can be divided into two meaningful groups, it will be advantageous to split the ring. This result is true for any power-law distribution. It may at first be surprising that this result is independent of the size of the original ring. However, this must be true since a power-law distribution does not provide a natural length scale that is necessary for a dependence on ring size.

We note that small modifications to the model may change the threshold probability of error that we allow until we split the ring. For example, if we initially chose the incorrect ring, and then assume that we can not return to the correct ring once we are done surfing our first choice, we should only split the ring if $\epsilon < 1 - \frac{1}{\sqrt{2}}$.

These results can be applied recursively. Any ring can be replaced by two rings and a decision of which ring to follow. Thus, as long as there is a

meaningful criterion to split the ring into two, such that the probability of choosing the correct ring is large enough, we will want to do so. The result will be a large collection of rings organized into a tree structure, very much like the hierarchy observed at ring portal sites [6].

4 Conclusion

The model presented in the last section suggests that there is an optimal tree structure to organize web rings of similar topic. The fact that web rings have a circular structure is not important – what matters is the ease with which one navigates from one site to the next by following the web ring links. It is then only a question of deciding how to optimally organize sites into rings, and how to fit rings into a hierarchy.

The parameters in the model that helps to construct such a hierarchy can in principle be measured, even though in practice there are a number of challenges. It should then be possible to make quantitative comparisons between this model and real usage data from a portal site. Combining the ideas presented here with machine learning and clustering algorithms may then lead to automated ways of optimally organizing an index to web sites [12].

References

[1] http://interviews.strategyweek.com/rs.nsf/Interviews/Hollis+Thomases

[2] http://www.wilsonweb.com/research/associate.htm

[3] L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank citation ranking: Bringing order to the Web," submitted for publication.

[4] http://www.cyber-robotics.com/info/

[5] http://www.almaden.ibm.com/cs/k53/clever.html

[6] http://webring.yahoo.com and http://bomis.com for example.

[7] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "Extracting large-scale knowledge bases from the web", Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999.

[8] Rajan M. Lukose and B. A. Huberman, "Surfing as a Real Option", Computational Economics Symposium, Cambridge, England. June 1998.

[9] Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow and Rajan M. Lukose, "Strong regularities in World Wide Web surfing", Science **280**, 95-97 (1998).

[10] Raj S. Chhikara and J. Leroy Folks, *The Inverse Gaussian Distribution*, Marcel Dekker, Inc., 1989.

[11] http://www.nytimes.com/library/tech/99/09/biztech/technology/22kilg.html

[12] S. Chakrabarti, B. Dom, R. Agrawal, P. Raghavan. "Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies". VLDB Journal **7**, 163-178 (1998) .