A Probabilistic Model for Intelligent Web Crawlers

Ke Hu and Wing Shing Wong
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
khu8@ie.cuhk.edu.hk, wswong@ie.cuhk.edu.hk

Abstract

With the enormous growth of the World Wide Web in recent years, the issue of how to discover web pages efficiently has become an important challenge for web crawler designers. In this paper, we will outline a simple model to predict the distribution of the search depth in a breadth-first search to reach the first web pages relevant to a user query. We define this probability as the crawler confidence. Recent studies indicate that at a large scale the Web structure subscribes to power law distribution on several aspects [3][7]. However, our work tries to model a microscopic linkage structure of the Web from an intelligent crawler's point of view. With the information provided by crawler confidence, an intelligent crawler can adjust its crawling behavior to achieve a higher harvest rate.

1. Introduction

Web crawlers, also known as robots or spiders, are the essential components of all search engines and are becoming increasingly important in data mining and other indexing applications. A traditional search engine consists of three major parts: a crawler, an indexer, and a query processor. Crawlers traverse the WWW by following the hyperlinks to fetch web pages from the WWW for indexing.

A traditional search engine typically tries to index all the web pages of the WWW in a centralized database for later query processing. However, due to the exponential growth of the web, search engines cover only a small part of the WWW. According to a study released in October 2000 [8], the directly accessible "surface web" consists of about 2.5 billion pages, with a total of 19 terabyte data, while the "deep web" (dynamically generated web pages) consists of about 550 billion web pages, 95% of which are publicly accessible. The most popular search engine today is arguably Google. Its index contained 560 million full-text-indexed web pages in June 2000. In other words, Google covered only 0.1% of the publicly accessible web at that time. One can expect other major search engines also face similar limitations. Moreover, the size of WWW is increasing with a rate of 7.3 million pages per day [8]. The trend towards automated production of web pages from databases makes it likely that such a growth rate will continue, or even accelerate. The enormous number of web pages available becomes a challenge for today's crawling systems. Moreover, due to the dynamical nature of web pages, issues of how to ensure freshness of a page index or how to design re-visiting strategy pose interesting challenges to researchers.

Recently, new types of crawling techniques, such as focused crawling [4] and intelligent crawling [1], have attracted many attentions. Compared with traditional crawlers, these crawlers only crawl a small part of the Web starting from a few well-chosen web sites. Given a predefined topic, or some query words, they can find their ways to the community of the related web pages efficiently and automatically. In the intelligent crawling approach, no specific model for web linkage structure is assumed. On the contrary, the crawler learns the statistics of linkage structure along the crawling process.

Our overall aim is to propose a simple probability model to enable crawlers to estimate how far away they are from the nearest relevant web page. These estimates can serve as performance baseline for intelligent crawling. For example, if searching from a given web page does not yield any relevant pages after an expected number of search depth has been reached, it may be more beneficial to restart the search elsewhere rather than following the search tree further. Although web pages are well connected [3], it is also acceptable that web pages relevant to the same topic are connected more closely [4]. When a crawler has been misled to a wrong direction, we shall be able to suspend searching this direction after few trials. As expected, such an intelligent algorithm is complicated since the web structure is enormously complex. In this paper, we describe a key step in our approach that concerns with the question of how to predict the distribution of the search depth in a breadth-first search

Proceedings of the 27th Annual International Computer Software and Applications Conference (COMPSAC'03)
0730-3157/03 $ 17.00 © 2003 IEEE

to reach web pages relevant to a user query. We define this probability as the crawler confidence. In the next section, we describe a simple web model and define the concept of crawler confidence. Some basic properties of the crawler confidence are presented. Conclusion and directions for future work will be presented in section 3.

2. Crawler Confidence

As an intelligent crawler crawls the Web, information it gathers during the process can be exploited to adjust upcoming crawling behavior dynamically [1]. It is important to an intelligent crawler that it should maintain estimates for each candidate URL during the crawling. One such estimate is how far an intelligent crawler is away from the first nearest web page that is relevant to the given query and not yet visited. Consider a breadth-first search starting from a given page. The search history can be described by a tree with the given web page serving as the root. We define the probability of meeting the first relevant web page at a certain depth of the search tree as the crawler confidence of that web page. In this paper, we propose a simplistic probabilistic model to illustrate this concept.

Three heuristic assumptions about crawling are made in this study:

1. A web crawler visits only a small part of the whole Web.
2. The number of outgoing links of a web page is modeled by identical, independently distributed (i.i.d.) random variables.
3. The process of crawling can be modeled as a Markov chain.

To illustrate our web model, assume a user has entered a query and it is the objective of the crawler to dynamically locate a page on the Web relevant to the topic. If a page is relevant to the user-defined query, we use the symbol *relevant* to represent this event. If not, we use the symbol *irrelevant*. Consider now the simple scenario where there are two web pages, page A and page B, with page A pointing to page B through a hyperlink. Page A is the current page visited by the crawler and page B is a new page, not yet visited by the crawler. It follows that we may use the following four probabilities to represent the relationships between these two pages.

$$P_1 = p(B = relevent|A = relevant) \quad (1)$$

$$P_2 = p(B = irrelevent|A = relevant) \quad (2)$$

$$P_3 = p(B = relevent|A = irrelevant) \quad (3)$$

$$P_4 = p(B = irrelevent|A = irrelevant) \quad (4)$$

Note that, $P_1 + P_2 = P_3 + P_4 = 1$.

In order to illustrate the concept of crawler confidence, we model the web linkage structure by the following diagram:

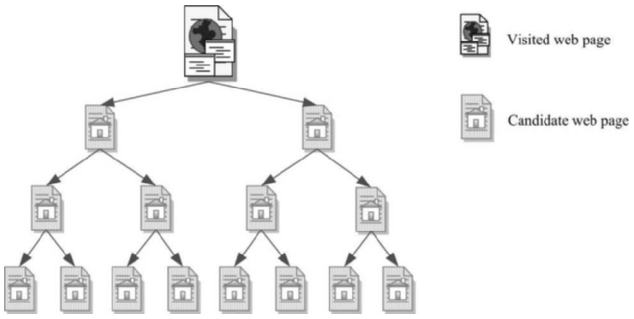

Figure 1. A binary search tree structure.

That is, we assume that all the web pages have exact two outgoing links to unvisited web pages, forming a binary tree structure. Therefore, if a crawler visits all the pages at a depth of 1, (one level below the starting page,) it will visit 2 new pages. Accordingly, the crawler will access 2^n new web pages by visiting all the pages at a depth of n below the starting page. Since the starting web page can be either relevant or irrelevant, we have the following two cases:

Figure 2. Two cases of the starting web page.

In this paper we derive the probability distribution for a crawler to meet the first relevant web page at a search depth i. Since the starting page could be either relevant or irrelevant, we have the following two cases.

1. If the starting page is a relevant page, then.

$$p(s = 1) = (1 - P_2^2) \quad (5)$$

$$p(s = i) = (1 - P_4^{2^i})P_2^2 P_4^{2^i - 4} \quad for \quad i = 2, 3, \ldots \quad (6)$$

2. If the starting page is an irrelevant page, then

$$p(s = i) = (1 - P_4^{2^i})P_4^{2^i - 2} \quad for \quad i = 1, 2, \ldots \quad (7)$$

The first factor of $p(s = i)$ in both (6) and (7), i.e. $(1 - P_4^{2^i})$, is the probability that a crawler meets at least

one relevant web page at level i. The second factors, i.e. $P_2^2 P_4^{2^i-4}$ and $P_4^{2^i-2}$, are the probability that the crawler cannot meet any relevant web pages in the previous $i-1$ levels. Moreover, one can easily verify that $\sum_{i=1}^{\infty} p(s = i) = 1$ in these two cases. We call the distribution $p(s)$ the crawler confidence of the starting web page. The expected number of depth is given by $E[S] = \sum_{i=1}^{\infty} s \cdot p(s = i)$. $E[S]$ provides an indication of how far the starting web page is from the nearest relevant web page.

Suppose for a given query, an intelligent crawler visits N web pages, L of the pages satisfy the user query. One can conclude that $p(A = relevant) = L/N$. Moreover, assuming that the links are completely random, it follows that $p(B = relevant, A = relevant) = (L/N)^2$ and $p(B = irrelevant, A = irrelevant) = (N-L)^2/(N)^2$. In reality, because of the short-range topic locality discussed in the focused crawling [4], the probability that both the source and the destination web pages are relevant to the predefined query is much higher than $(L/N)^2$ [1]. And due to the same reason, P_4 shall also be greater than $(N-L)^2/(N)^2$. For example, for a certain given query, an intelligent crawler may fetch 10,000 web pages and only 100 pages of them are relevant. So, we have $p(A = relevant) = 0.01$. Since web pages that relevant to the same topic are usually directly connected by hyperlinks, P_1 should be greater than $p(B = relevant|A = relevant) = \frac{p(B=relevant, A=relevant)}{p(A=relevant)} = 0.01$, and P_4 greater than $p(B = irrelevant|A = irrelevant) = \frac{p(B=irrelevant, A=irrelevant)}{p(A=irrelevant)} = 0.99$. In order to illustrate our model, we assume $P_1 = 0.1$ and $P_4 = 0.999$ in the Figure 3.

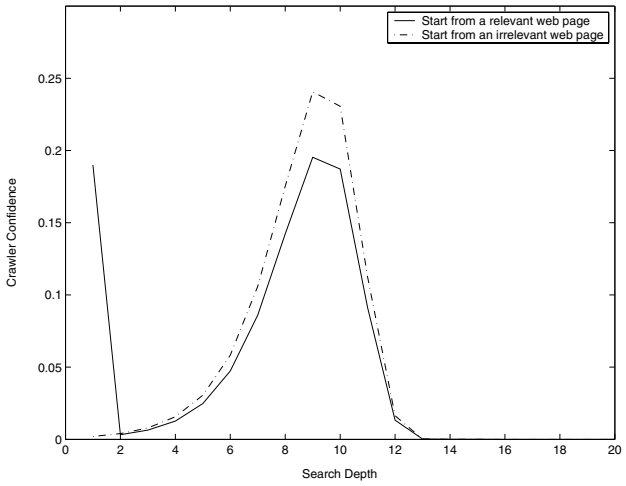

Figure 3. Distribution of crawler confidence.

In this example, the expected search depth of case 1 is 7.21 and for case 2 it is 8.65, i.e. there are 7.21 levels on the average between two closest relevant pages, and 8.65 levels between an irrelevant page and the nearest relevant page. Moreover, in this figure, we can observe that there is a unique local maximum in the range $2 < i < \infty$ for both cases.

In the previous example, if the search from a given web page does not yield any relevant pages after the expected depth, 7.21 or 8.65, has been reached, we may consider that the crawler has not been started in a fruitful direction, as it seems to be harder than expected to find a relevant web page. It may be more beneficial to restart the search elsewhere rather than following the search tree further. Intelligent searching is evaluated by the harvest rate [1], which is the ratio of the number of retrieved relevant web pages and the number of retrieved web pages. Therefore, suspending crawling in a wrong direction will benefit the crawling performance. Our model can provide a crawler a feasible method to estimate the minimal number of trials before we suspend.

In general, we can prove the following results.

Proposition 1: In the second scenario, if $P_4 > \sqrt{(\sqrt{5}-1)/2}$, there exists an integer $N > 1$, such that the crawler confidence is monotonic increasing until N and deceasing from $N+1$. If $P_4 < \sqrt{(\sqrt{5}-1)/2}$, the crawler confidence is monotonic decreasing. If $P_4 = \sqrt{(\sqrt{5}-1)/2}$, $p(s=1) = p(s=2)$ and the crawler confidence is monotonic decreasing for $i \geq 2$.

Proof: To prove this, consider the function:

$$f(n) = (1 - P_4^{2^n}) P_4^{2^n - 2} \quad for \quad i = 1, 2, \ldots \tag{8}$$

First of all, note that $f(n) = (1 - P_4^{2^n}) P_4^{2^n-2} > 0$. Therefore,

$$\frac{f(n+1)}{f(n)} = \frac{(1 - P_4^{2^{n+1}}) P_4^{2^{n+1}-2}}{(1 - P_4^{2^n}) P_4^{2^n-2}} = (1 + P_4^{2^n}) P_4^{2^n} \tag{9}$$

So whether $f(n)$ is increasing or not depends on $(1 + P_4^{2^n}) P_4^{2^n}$ whether is greater than 1 or not, which in turns depends on whether $P_4^{2^n}$ is greater than $(\sqrt{5}-1)/2$ or not. Note that the function $(1+x)x$ is monotonic increasing in x and $P_4^{2^n}$ is monotonic decreasing in n. The proposition follows by considering the different possibilities for $P_4^{2^n}$.

If $P_4 > \sqrt{(\sqrt{5}-1)/2}$, then at $N = \lceil \frac{ln(-ln((\sqrt{5}-1)/2) - ln(-lnP_4)}{ln2} \rceil > 1$, the distribution for the crawler confidence achieves its maximum.

If $P_4 < \sqrt{(\sqrt{5}-1)/2}$, at $i = 1$, the distribution for the crawler confidence achieves its maximum. In other words, the crawler confidence is monotonic decreasing for $i \geq 1$.

If $P_4 = \sqrt{(\sqrt{5}-1)/2}$, $p(s=1) = p(s=2)$ and the crawler confidence is monotonic decreasing for $i \geq 2$.

In reality, since the number of web pages that are both relevant to a certain user query must be a very small part of the whole World Wide Web, P_4 shall be always close to 1. $P_4 > \sqrt{(\sqrt{5}-1)/2} = 0.7862$ is the most likely case.

Proposition 2: In the first scenario, if $P_4 > \sqrt[4]{(\sqrt{5}-1)/2}$, there exists an integer $N > 2$, such that the crawler confidence is monotonic increasing until N and deceasing from $N+1$. If $P_4 < \sqrt[4]{(\sqrt{5}-1)/2}$, the crawler confidence is monotonic decreasing for $i \geq 2$. If $P_4 = \sqrt[4]{(\sqrt{5}-1)/2}$, $p(s=2) = p(s=3)$ and the crawler confidence is monotonic decreasing for $i \geq 3$.

Proof: To prove this, consider the function:

$$f(n) = P_2^2(1-P_4^{2^n})P_4^{2^n-2} \quad for \quad i = 1, 2, \ldots \quad (10)$$

Note that $f(n) > 0$. Therefore,

$$\frac{f(n+1)}{f(n)} = \frac{P_2^2(1-P_4^{2^{n+1}})P_4^{2^{n+1}-4}}{P_2^2(1-P_4^{2^n})P_4^{2^n-4}} = (1+P_4^{2^n})P_4^{2^n} \quad (11)$$

From equation (11), we find that $\frac{f(n+1)}{f(n)}$ is independent of P_2 when $n \geq 2$.

If $\frac{f(n+1)}{f(n)} = (1-P_4^{2^n})P_4^{2^n} > 1$, which is equivalent to $P_4 > \sqrt[4]{(\sqrt{5}-1)/2}$, we can find $N = \lceil \frac{ln(-ln((\sqrt{5}-1)/2)-ln(-lnP_4)}{ln2} \rceil > 2$, so that the crawler confidence is monotonic increasing from 2 to N and monotonic decreasing after N .

If $P_4 < \sqrt[4]{(\sqrt{5}-1)/2}$, then the crawler confidence is monotonic decreasing for $i \geq 2$. (Whether $p(s=1)$ is greater than $p(s=2)$ or not would depend on the value of P_2 and P_4.)

If $P_4 = \sqrt[4]{(\sqrt{5}-1)/2}$, $p(s=2) = p(s=3)$ and the crawler confidence is monotonic decreasing for $i \geq 3$.

Proposition 3: In the first scenario, fixing P_4 and assuming $P_4 > \sqrt[4]{(\sqrt{5}-1)/2}$, the maximum value of the crawler confidence distribution achieved in the domain $[2, \infty)$ is a monotonic increasing function of P_1.

Proof: If P_4 is fixed, the value N where $p(s=N)$ achieves the maximum value in the domain $[2, \infty)$ is also fixed. Since $P_2 = 1 - P_1$, P_2 is a monotonic decreasing function of P_1. Hence, $p(s=N) = P_2^2(1-P_4^{2^N})P_4^{2^N-4}$ is monotonic decreasing as a function of P_1.

Proposition 4: In the second scenario, if $P_4 > \sqrt{(\sqrt{5}-1)/2}$, as the probability P_4 increases, the value N that maximizes $p(s=N)$ is monotonic increasing.

Proof: From the proof of Proposition 1, we have $n = \frac{ln(-ln((\sqrt{5}-1)/2)-ln(-lnP_4)}{ln2}$. When P_4 decreases, n is monotonic decreasing. The relation between N and P_4 is shown in the Figure 4.

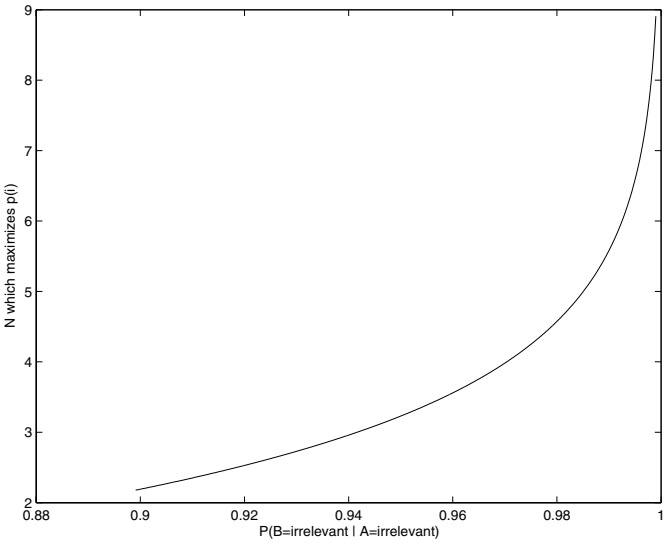

Figure 4. The relation between N and P_4.

In fact, the number of outgoing links can be more accurately modeled as identical, independently distributed (i.i.d.) random variables. Recent research indicates that outdegree distribution of web pages subscribes to the power law [3]. In future work we will address the crawler confidence for this case.

3. Conclusions and Future Work

In this paper, we have proposed a probabilistic model for estimating how far is the nearest relevant web page from a given web page as an intelligent crawler crawls along the URLs. Considering the limitation of bandwidth and the computational capability, intelligent crawlers need to choose the most promising way to follow to achieve a higher harvest rate. With this model, intelligent crawler could determine when it should go further and when it should terminate searching in a certain direction after few trials from a given starting web page.

For future work, we note that based on this model, other crawling techniques, such as URL token based estimation, content based estimation, and linkage structure based estimation, can be incorporated to adjust the crawler confidence of each direction. We also have to admit that the proposed simplistic model is just an illustration of the crawler confidence concept. The model can be improved by considering issues such as linking to visited pages and randomizing the number of outgoings links to make it more realistic and reliable.

References

[1] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. *Proceedings of the 10th International World Wide Web Conference*, May 2001.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search. *Proceedings of the 7th International World Wide Web Conference*, pages 107–117, April 1998.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. *Proceedings of the 9th international World Wide Web conference on Computer networks*, pages 309–320, June 2000.

[4] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic specific resource discovery. *Proceedings of the Eighth World Wide Web Conference*, pages 545–562, 1999.

[5] S. M. Cherry. Weaving a web of ideas. *IEEE Spectrum*, Sept 2002.

[6] J. Edwards, K. McCurley, and J. Tomlin. An adaptive model for optimizing performance of an incremental web crawler. *Proceedings of the 10th International World Wide Web Conference*, May 2001.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Proceedings of ACM SIGCOMM*, 1999.

[8] P. Lyman and H. R. Varian. How much information? *http://www.sims.berkeley.edu/research/projects/how-much-info/*, October 2000.

[9] M. Najork and J. L. Wiener. Breadth-first search crawling yields high-quality pages. *Proceedings of the 10th International World Wide Web Conference*, May 2001.