ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
www.computerscijournal.org

ISSN: 0974-6471
December 2011,
Vol. 4, No. (2):
Pgs. 411-415

A Site Rank-Based Swarming Ordering Approach

MAYA RAM ATAL*, ROOHI ALI, RAM KUMAR and RAJENDRA KUMAR MALVIYA

Department of Computer Application, Government Geetanjali Girls PG College, Bhopal (India).
*Corresponding author: E-mail: atalprof2002@gmail.com

(Received: August 12, 2011; Accepted: September 04, 2011)

ABSTRACT

Search engines are in performance a major essential role in discovering information nowadays. Due to limitations of network bandwidth and hardware, search engines cannot obtain the entire information of the web and have to download the most essential pages first.

In these paper, we propose a swarming ordering strategy, which have based on SiteRank, and compare it with several swarming ordering strategies. All the four strategies make an optimization for the naive swarming more or less. At the beginning of the swarming process, all the strategies can crawl the pages with high PageRank. When downloading 48% of the pages, the sum of PageRank is over 58% even for the worst one. At the later phase of swarming, the sum of PageRank varies slowly and reaches to unique finally. The objective of these strategies is to download the most essential pages early during the crawl. Experimental results indicate that SiteRank-based strategy can work Efficiently in discovering essential pages under the PageRank evaluation of page quality.

Key words: Web crawler, Swarming ordering strategy, Web page importance, SiteRank.

INTRODUCTION

Web search engines are becoming a predominant tool for information discovery and retrieval on the infinite web. But, as the huge volume of web pages and limitations of hardware, a search engine can only fetch a fraction of the web. Therefore, which pages should be downloaded first is significantly crucial. No doubt that the quality of the downloaded page set is one of the most essential factors.

When a web crawler prepares to download a page set, it starts with an initial seed URL. Then all the URLs that are parsed from the seed page are added into the URL waiting list for further visit. A main problem here is how to choose the pages that should be visited next. This is so-called Swarming Ordering Strategy.

There are several swarming ordering strategies such as Breadth First-Search[2,3], Backlink-Count[3], PageRank[2] and LargerSites-First[2].

Although some search engines have used all of these strategies, they share the same problem. They do not take effective page quality metric. So we propose a strategy based on SiteRank, which prefers the pages from sites with high SiteRank scores. The reminder of this paper is structured as follows. Section 2 gives an overview of the related work. Section 3 introduces our algorithm in detail.

Section 4 shows our experimental results running on 27K pages. We draw our conclusions in the last section.

Relative Work

Given the current size of the web and its constant change, even large search engines cover only a fraction of publicly available Internet. Since a crawler always downloads a portion of web pages, it is highly desirable that the downloaded pages have high quality. Thus, here come the challenges: what does high quality means and how to select the page for next visit during swarming process. Breadth-First-Search selects page according to the order of the URL in the list[2,3]. These method is easy to implement and can have work effectively in most situations. When using PageRank [6] as the page quality metric, ithave been found that BreadthFirst-Search downloaded the hot pages first and the average quality of pages decreased over the duration of the swarming process[5].

Backlink-Count regards a page with the most backlink count as the candidate of the next visit, that is to say, the number of backlink is the metric of page quality. An experiment on 199,000 pages of stanford.edu domain shows that Backlink-Count had poorer performance than BreadthFirst-Search on the page quality metric of PageRank[3].

PageRank uses the web graph of pages that have been downloaded so far to compute the PageRank [6] values of pages in the list, and then chooses the page with the highest score for further visit. Obviously this is a time-consuming process which makes this strategy quite inefficient. At the same time, the experimental results[2] suggested that the PageRank was less effective than BreadthFirst-Search.

The LargerSites-First strategy is mainly based on the assumption that a large scale site may have a high possibility of high quality [2]. So this strategy prefers large sites to small sites. Experimental results indicate that this strategy outperforms the three methods above. However, it has a great bias against small sites and how to define “large” is still a subtle problem.

SiteRank-Based Swarming Ordering Strategy

We first introduce SiteRank and then illustrate how to employ it to direct the swarming process. At least, we give a crawler simulator model.

SiteRank

There are two key problems we care about. The first is how to define the concept of site; the second is how to compute SiteRank score. Site can be defined according to either the URL or the host address of each site. When we use site URL to define a site, we regard the URL before the first slash as an independence site. For example, http://www.vit.edu and http://www.lib.vit.edu are considered as two different sites. When using host address, we just consider different server IPs as different sites. In our experiment, we use site URL to divide sites. This way is not only reasonable but also easy to operate. Further, it is fit for the concept of sites in common sense.

SiteRank is an algorithm, which is similar to PageRank. It exploits the relation of citations among sites. For a Site Graph G < V, E > which consists of two parts: a set V of vertices where each vertex represents a web site and a set E of edges where each edge represents a direct link between two sites. Note that the links between two sites' pages are accumulated, i.e., if there are three sites S1, S2, S3, where 100 direct links between S1 and S2, 10 between S1 and S3. Obviously, the recommendations of S2 and S3 by S1 are not equivalent. So we weight the link count when compute the SiteRank. When using random surf model, we can gain the transition matrix M as 3.1 [7]:

$$M(i,j)=\begin{cases} C_{ij}/C_i & \text{if } S_i \text{ points to } S_j \text{ and } C_i \neq 0 \\ 0 & \text{if } S_i \text{ does not points to } S_j \\ 1/N & \text{if } C_i = 0 \end{cases} \quad ...(3.1.1)$$

In this definition, C_{ij} represents the direct link count that site S_i points to site S_j and C_i represents outlinks of site S_i. N is the total number of web sites. So we can get a formula for SiteRank as in 3.2:

$$M = \alpha M + \frac{(1-\alpha)}{N} * I \quad ...(3.1.2)$$

where the decay factor á is usually set to 0.85, I is the matrix whose size is the same as that of M and all elements of it have the value of 1, and N is the total number of web sites.

Advantages of SiteRank

The SiteRank-Based swarming ordering strategy has an assumption that the pages included in high SiteRank sites are always good in quality [8]. This assumption is reasonable and this strategy has some advantages in discovering high quality pages.

Stableness

Compared within PageRank, stableness is the main advantage. As we know, SiteRank ignores the intra-site links and it only considers the inter-site links. The statistics[7] on the web links suggests that 76% of the web links comes from intra-site. The web is varying all the time; the architecture of a site is changing as well. The links among pages are not stable enough. Thus, PageRank has lack of stableness. However, the links between sites are varying less frequently. So the SiteRank can be stable for a longer period.

Fairness

The Largest Sites-First strategy chooses the sites having the most pending pages for the next visit [2]. These inevitably leads to the injustice to small sites, which are good in quality. Also, this strategy has the limitation of introducing more sites to the user. The SiteRank prefers the quality to the size of a site to some extent. SiteRank can give all the sites equivalent opportunities.

Anti-Spamming

Link spamming is one of the common spamming techniques [4]. Spammers often build link farms on their own sites to boost their target pages. When we use links to analyze the quality of the pages or sites, inter-site links are more significant. So SiteRank can greatly weaken the action of link farms. Thereby, the SiteRank-Based strategy can improve the quality of the crawled pages.

Experimental Evaluation

In our experiment, we do not crawl web pages of the real web. Instead, we use a simulator to imitate the swarming process on a static dataset. The reason to use the dataset is that the real web is changing all the time. To ensure the justness to all the swarming ordering strategies, we must make sure that all the algorithms are running on the same dataset. Also, we use the crawler simulator because we know that a real crawler needs to take many factors into consideration, such as time and network bandwidth. While in our experiment, we only focus on the ability of discovering high quality pages.

We run our crawler on a 27K-pages dataset and compare our proposed strategy, SiteRank, with the other strategies. The experimental result is presented in this section.

Dataset

We implement the web crawler program to create the dataset. This dataset is then used by crawler simulator to compare result of applying BreadthFirst-Search, LargerSites-First, BackLink-Count and also SiteRank algorithms. Our dataset contains 37K pages with about 1.8M links on 4399 sites. Table 1 gives some statistical information about this dataset.

Table 1: Statistic of the Dataset

Sites	4399
Pages	37 K
Links	1.8 M

Seed Selection

When the dataset is created, we should select a seed URL to start the swarming process with it. Seed selection is essential in this experiment. The quality of the seed will affect the experimental results. If the seed is not properly selected, the crawled part may poorly represent the whole dataset and the analysis on the results is useless and makes no sense. In order to guarantee the propagation and coverage of the seed, we manually choose a high quality page as the initial seed. We have two principles for the seed. First, the site to which the seed belongs should have high Rank to ensure the richness in links. The second is that the seed should

result in finding links that are domain dispersed which can avoid the bias of the swarming pages. Based on these two principles we can get a good coverage of the dataset.

Swarming Process

We implement BreadthFirst-Search, Backlink-Count, LargerSites-First as well as our SiteRank-Based strategy. We do not make any optimization for any of the algorithms. We just use the naive algorithms because we do not care about running time.

To perform the swarming, first we apply the PageRank algorithm on the dataset and get a PageRank value for every page in it. Then, we run the simulator using BreadthFirst-Search algorithm. The result of this crawl is a set of pages that are retrieved according to this strategy. Finally, the value from PageRank will be associated to the links generated by PageRank. We use the same process for the remaining algorithms and then we will have a set of links and values as a result of swarming process.

Performance Evaluation

We give a performance evaluation model to evaluate the performance of all the algorithms. In this model we evaluate a strategy based on the quality of the pages retrieved. Therefore, a high performance strategy should find high quality pages as early as possible. Based on this rule, we can get a generally qualitative analysis for each strategy. The detail of this evaluation metric is discussed below.

Cumulative PageRank

This metric gives us a prospect of the process of discovering high quality pages. We equally set ten detecting points in our experiment. At each point, we calculate the sum of PageRank of the crawled pages. On an ideal occasion, the distribution of cumulative PageRank has a rapid increase at the beginning of the swarming process and varies to flat in the middle of the process.

Result of Experiment

Fig 1 shows the performance of each swarming ordering strategy on discovering high quality pages. We use PageRank as the metric of page quality. We set ten detecting points at each 10% increment of pages and compute the sum of PageRank of the crawled pages.

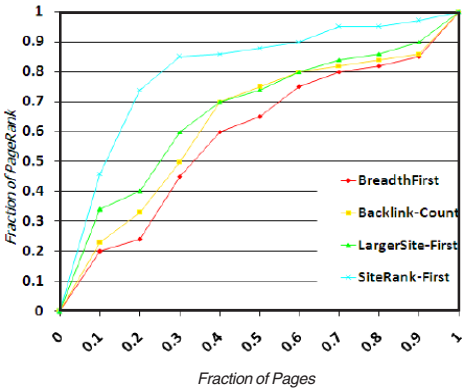

Fig. 1: Performance of Discovering High Quality Pages

CONCLUSION

The explosion of the web information and many restrictions of the crawler, forces the crawler to download the very high quality pages first. In our experiment on a 37K-pages dataset, we implement four strategies, which are considered to be effective. At the same time we propose an ordering strategy, which is based on SiteRank, since SiteRank has advantages of stableness and fairness.

In order to evaluate the performance of all the strategies, we use the cumulative PageRank as the evaluation metric. The experimental result shows that the SiteRank-Based strategy has the best performance in discovering high quality pages.

From the fig 1, we can have draw some conclusions

- All the four strategies make an optimization for the naive swarming more or less. At the beginning of the swarming process, all the strategies can crawl the pages with high PageRank. When downloading 40% of the pages, the sum of PageRank is over 50% even for the worst one. At the later phase of swarming, the sum of PageRank varies slowly and reaches to 1 finally.
- As we can see figure 4.5.1, LargerSites-first,

Backlink-Count and also BreadthFirst-Search perform almost similar to each other. However, we can say that result of applying LargerSites-First, is slightly better than Backlink-Count.

´ Our strategy makes a notable improvement and works effectively in discovering high quality pages. From this figure we can find that at the first 20% fraction, the sum of PageRank can reach to 75%. It is almost two times higher than the other three. At the half of the swarming process, the sum of PageRank is near to 90%.

ACKNOWLEDGEMENTS

We thank to Prof. M. L. Dhore, head of computer department in our institute that provided the required facilities and all of the staffs that helped us to complete this work.

REFERENCES

1. A. Gulli and A. Singnorini. The indexable web is more than 11.5 billion pages. In proceedings of the 14th international conference on World Wide Web.
2. R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez. Swarming a country: Better strategies than breadth-first for web page ordering. In Proceedings of the 14th international conference on World Wide Web (2005).
3. J. Cho, H. Garc´ýa-Molina, and L. Page. Efficient swarming through URL ordering. Computer Networks and ISDN Systems, **30**(1-7): 161–172, (1998).
4. Z. Gyongyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web (2005).
5. M. Najork and J. L. Wiener. Breadth-First Swarming Yields High-Quality Pages. In Proceedings of the 10th International World Wide Web Conference, Hong Kong, May 2001.
6. L. Page, S. Brin, R.Motwani, and T.Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998).
7. J. Wu and K. Aberer. Using siterank for p2p web retrieval. Technical Report IC/2004/31, Swiss Federal Institute of Technology, Lausanne (2004).
8. Qiancheng Jiang, Yan Zhang. SiteRank-Based Swarming Ordering Strategy for Search Engines. In proceedings of Seventh International Conference on Computer and Information Technology.