GENETIC ALGORITHM FOR EVALUATION METRICS IN TOPICAL WEB CRAWLING

T. Peng, W.L. Zuo and Y.L. Liu
College of Computer Science and Technology, Jilin University, Changchun 130012, P. R. China

Abstract  A topic driven crawler chooses the best URLs to pursue during web crawling. It is difficult to evaluate what URLs downloaded are the best. This paper presents some important metrics and an evaluation function for ranking URLs about pages relevance. We also discuss an approach to evaluate the function based on GA. The best combination of the metrics’ weights can be discovered by GA evolving process. The experiment shows that the performance is exciting, especially about a popular topic.

Keywords:  topic driven crawler, genetic algorithm, real-coded, relevant pages.

1.  INTRODUCTION

A crawler is a program that retrieves web pages for a search engine, which is widely used today. The WWW information is distributed, also the information environments become complex. Because of limited computing resources and limited time, topic driven crawler (also called focussed crawler, retrieve web pages relevant a topic) has been developed. Topic driven crawler carefully decides which URLs to scan and in what order to pursue based on previously downloaded pages information. Some evaluation methods for choosing URLs [1] and several special crawlers, Naive Best-First crawler and DOM crawler [2] do not have satisfying adaptability. In this paper, we present an approach to evaluate a function about pages’ relevance based on genetic algorithm (GA). We use GA to evolve some weights of the metrics. GAs are general purpose search algorithms which use principles inspired by natural genetic populations to evolve solutions to problems [3, 4]. In our approach, not as usual, an individual is a combination of the real-coded metrics’ weight, and it’s more natural to represent the optimization problem in the continuous domain.

G. R. Liu et al. (eds.), Computational Methods, 1203–1208.
2. THE EVALUATION FUNCTION (METRICS)

Not all pages which crawler observed are ‘relevant’ during crawling. For instance, if a crawler builds a specialized database on a particular topic, then pages referring to that topic are more important, and should be visited as early as possible. Similarly, if a page points to lots of authority pages, then the page is a high hub score page [5]. If the crawler cannot visit all the pages, then it is better to visit those ‘important’ pages, since this will give the end-user higher ranking results. We define the evaluation function $I(u)$, $u$ is the URL that the crawler will be pursued, $p$ is the parent page of $u$. The evaluation function is a weighted combination of followed metrics:

1. $sim(p, q)$ (similarity of page $p$ to topic $q$): A topic $q$ drives the crawling process, and $sim(p, q)$ is defined to be the textual similarity between $p$ and $q$, this is shown in the work by the authors [6].

$$sim(p, q) = \frac{\sum_{j=1}^{r} q_j \times l_j \times \omega_j}{(|\omega| \times |p|)}$$  \hspace{1cm} (1)

where $\omega' = (\omega_1 \times l_1, \ldots, \omega_r \times l_r)$; $\omega_j$: the weight of the $j$th word; $l_j$: the inverse document frequency (idf) of the $j$th word;

2. $hub(p)$ (the evaluation of hubs property): Hub pages are defined to be web pages which point to lots of ‘important’ pages relevant a topic.

$$hub(p) = \frac{|L_p|}{\sum_{i=1}^{N} |L_i|}$$  \hspace{1cm} (2)

where $|L_p|$: the number of outlinks of page $p$; $\sum_{i=1}^{N} |L_i|$: the average number of outlinks of the pages that are already downloaded.

3. $bc(p)$ (backlink count): the number of links to $p$

$$bc(p) = \frac{|P_p|}{M}$$  \hspace{1cm} (3)

where $M$ is a parameter provided by user.

4. $uts(u, q)$ (similarity of URLs text to topic $p$):

$$uts(u, q) = sim(u, q) + thesaurus(u, q)$$  \hspace{1cm} (4)

where $thesaurus(u, q)$ is uses the thesaurus dictionary of topic $q$, this experimentation does not take the metric (the future work I will do).