Improved Focused Web Crawling Based on Incremental Q-Learning

Yunming Ye[1], Yan Li[1], Joshua Huang[2] and Xiaofei Xu[1]

[1] Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China
yym_sjtu@yahoo.com.cn
[2] E-Business Technology Institute, The University of Hong Kong, Hong Kong
jhuang@eti.hku.hk

Abstract. This paper presents *IQ-Learning*, a new focused crawling algorithm based on incremental Q-learning. The intuition is to extend previous reinforcement learning based focused crawler with the incremental learning mechanisms so that the system can start with a few initial samples and learns incrementally from the knowledge discovered online. First, a sample detector is used to distill new samples from the crawled Web pages, upon which the page relevance estimator can learn an updated estimation model. Secondly, the updated page relevance information is fed back to the Q value estimator constantly when new pages are crawled so that the Q value estimation model will be improved over time. In this way, the reinforcement learning in focused crawling becomes an incremental process and can be more self-adaptive to tackle the complex Web crawling environments. Comparison experiments have been carried out between the *IQ-Learning* algorithm and other two state-of-the-art focused crawling algorithms. The experimental results show that *IQ-Learning* achieves better performance in most of the target topics.

Keywords: Focused Crawler; Q-Learning; Incremental Learning; Web

1 Introduction

A Web crawler is an information gathering system that traverses the Web by following the hyperlinks from page to page and downloads Web pages that are interested. General Web crawlers visit the Web in an unselective mode. They aim at collecting Web pages as many as possible to build search engines. Different from the general crawler, a focused crawler [1] is an intelligent Web crawler that traverses the Web selectively to download pages in some predefined target topics. Given a search topic and a predefined maximum download number, the goal of a focused crawler is to collect as many relevant pages as possible while retrieving as fewer irrelevant pages as possible in the crawling process.

Focused crawlers are very useful in several Web applications [2], such as collecting Web pages with specific topics for domain-specific search engines, archiving specific page collections for a digital library, and gathering systematic information in some specific topics for market research or survey of literature on the

Advances in Natural Language Proccesing
Research in Computing Science 18, 2006, pp. 139-150