Graduate School of Fundamental Science and Engineering
Waseda University

博 士 論 文 概 要
Doctoral Thesis Synopsis

論 文 題 目
Thesis Theme

A Study on Efficient Topical Focused Website Segment Crawler

申 請 者
(Applicant Name)

Tanaphol	SUEBCHUA
スブチュア	タナポール

Department of Computer Science and Communications Engineering,
Research on Parallel and Distributed Architecture

April, 2018

Topic-specific web pages have become essential data for vertical search engines and NLP researches. To acquire these web pages, many researchers proposed to use a specialized web crawler, called the focused crawler. The general problem in the focused crawling research is to collect the web pages related to a specific topic with minimal computing resources, e.g., bandwidth and disk space. Thus, focused crawlers adopt some prediction mechanisms, such as machine learning predictor, to predict the relevancy of unvisited web pages before selectively downloading them. Although several efficient machine-learning-based focused crawlers have been proposed, there are still several problems that might affect their crawling efficiency as follows:

- **Feature dependencies problem —** Some prediction features that are utilized for predicting the topic of web pages, such as anchor text terms, are dependent on the relevancy of the download source web pages. These dependent features affect the prediction of the predictor. Suppose that the target topic is "Computer" and there is an unvisited web page that is cited by anchor text "mouse". It is well known that the term "mouse" can represent either a computer device or an animal. Thus, the destination should be relevant only if the contents in the source web pages are related to "Computer" topic. Otherwise, it is likely to be irrelevant.
- **Sources' relevancy problem —** The relevancy of the source web pages is one of the most common prediction features that has been utilized in several focused crawlers for predicting the unvisited destination one. However, the limitation of this feature is that the relevancy of the unvisited destination cannot be estimated correctly if the number of sources is small.
- **Repetitive visiting problem —** There are many situations that many irrelevant web pages are retrieved from a single website, repeatedly. For instance, the feature vectors of the unvisited irrelevant web pages are similar to the relevant ones in the training dataset. Consequently, the crawling efficiency will be decreased after collecting these irrelevant web pages.

These aforementioned problems directly affect the crawling efficiency of the crawler. Our previously proposed focused website segment crawling approach (FSC) is also affected by these problems. In our FSC crawler, a machine-learning-based predictor is trained to predict a group of web pages located in the same directory path, called website segment, whether it is a group of relevant web pages or not. All the web pages belonging to the most probable relevant website segment will be downloaded at once. Therefore, if the crawler predicts the irrelevant website segment incorrectly as relevant, many irrelevant web pages will be downloaded at once. Consequently, the crawling efficiency will be dropped, rapidly. On the contrary, the crawler will miss relevant web pages if the relevant segment is predicted as irrelevant. In this study, our main goal is to address these problems in the FSC crawler to improve its crawling efficiency**.** To achieve this goal, the following enhancement approaches are proposed:

1. **Twin predictors and noisy-aware updating scheme —** To address the **feature dependencies** problem, a new prediction approach, called the **"twin predictors"**, is proposed. In our approach, we propose to prepare two sub-predictors. The first and the second sub-predictors are learned by feature vectors (patterns) of web pages that are extracted from relevant and irrelevant sources, respectively. Depending on the relevancy of the sources, only one predictor will be selected for prediction. However, in general, the number of samples, i.e., feature vectors, that can be extracted from the fixed web dataset is limited. Thus, the number of samples for training each sub-predictor might not be enough for building the accurate prediction model. Therefore, we here also propose the **"noisy-aware updating scheme"** for updating the predictor. In our updating scheme, some noisy feature vectors of the web pages in the template-based website are eliminated by the proposed heuristic rule. The remaining are later used to update the predictor, periodically.

2. **Neighborhood feature —** To solve the **sources' relevancy problem**, a new prediction features, named the **"neighborhood feature"**, is proposed. Instead of relying only the relevancy of the downloaded web pages that directly cite to the destination, the proposed feature enables the adoption of additional already-downloaded web pages to estimate the priority of a target web page. The additionally adopted web pages consist of web pages located in the same environment as the target web pages, e.g., web pages belonging to the same website or the same directory path as the target one.

3. **History feature —** To mitigate the effect from the **repetitive visiting problems**, we propose another prediction features, named the **"history feature"**. The idea behind this feature is to track the changes of the number of relevant and irrelevant web pages retrieved from the website, which hosts the unvisited web pages, within the recent crawling attempts. Our assumption is that the unvisited web page belonging the website in which many relevant web pages have been repeatedly found should be given more priority than the one that many irrelevant web pages have been found.

According to our experiments, the twin predictors and noisy-aware updating scheme help the FSC crawler to gain more 12% of improvement in the crawling efficiency, at best. When the neighborhood feature is adopted, the enhanced FSC crawler could gain another 10% of improvement. As for the history feature, the FSC crawler enhanced by history feature obtains higher harvest rate than FSC crawler by maximum of approximately 5.2%. In addition, the results also suggest that some of the proposed approaches can also be used to enhance the efficiency of other web-page-based focused crawlers, such as Best-First or HMM crawler, as well.

This thesis is organized as follows:

- Chapter 1 provides details of the background, the problems, and the contribution of this thesis.
- Chapter 2 gives the overview details of web crawlers, related work in focused crawling researches, the relevant web page classification method used in our study, and the FSC crawler.
- Chapter 3 introduces the topics evaluated in this study. This chapter also includes the details on how to build the training dataset of the relevant web page classifier and the predictor in the FSC crawler.
- Chapter 4 details the twin predictors and noisy-aware updating scheme approach.
- Chapter 5 gives the details on the proposed neighborhood feature and the focused crawlers enhanced by the neighborhood feature.
- Chapter 6 presents the details of the proposed history feature and history-enhanced focused crawlers.
- Chapter 7 concludes this thesis and discuss the future work.

早稲田大学　博士（工学）　学位申請　研究業績書

(List of research achievements for application of doctorate (Dr. of Engineering), Waseda University)

氏 名(Full Name)Tanaphol SUEBCHUA　印(seal or signature　　)

（As of March, 2018）

種 類 別 (By Type)	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）(theme, journal name, date & year of publication, name of authors inc. yourself)
Journal	○**Tanaphol Suebchua**, Bundit Manaskasemsak, Arnon Rungsawang, and Hayato Yamana. Efficient topical focused crawling through neighborhood feature. New Generation Computing, Volume 36, Issue 2, pages 95-108, April 2018. doi:10.1007/s00354-017-0029-8
International Conference (refereed)	○**Tanaphol Suebchua**, Arnon Rungsawang, and Hayato Yamana. Adaptive focused website segment crawler. In Proceedings of the 19th International Conference on Network-Based Information Systems, pages 181–187. IEEE, September 2016. doi: 10.1109/NBiS.2016.5 ○**Tanaphol Suebchua**, Bundit Manaskasemsak, Arnon Rungsawang, and Hayato Yamana. History-enhanced focused website segment crawler. In Proceedings of the 32nd International Conference on Information Networking, pages 80-85. IEEE, January 2018. doi: 10.1109/ICOIN.2018.8343090