Tanaphol Suebchua 2014: Thai-related Foreign-language Specific Web Crawler. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Arnon Rungsawang, Ph.D. 41 pages.

National web archive that preserves national knowledge for generations to come has been successfully made available through a domain-specific web crawler for years. However, that kind of crawler still misses many foreign language web pages which are also related to the nation. This thesis proposes new machine learning based focused crawling approach to collect national related web pages written in a foreign language, especially the English web pages that related to Thailand. We first propose a notion of website segment which groups the related web pages from their same longest directory paths. Thus, rather than looking for a target web page as proposed in many traditional focused crawling approaches, an ensemble classifier is trained with several features to predict the relevancy of the website segments. The most relevant website segments in the crawling frontier are then enqueued to download. The preliminary experiments on the real web space show that our approach can provide better promising harvest results than the Breadth-First and Best-First baselines for the Thai-tourism, Thai-estate and Thai-diving topics which are written in English language.

_____ _____ __ / __ / __

Student's signature Thesis Advisor's signature