Ekkasit Srisukha 2009: Knowledge Based Thai Language-Specific Web Crawling. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Assistant Professor Arnon Rungsawang, Ph.D. 90 pages.

Language-specific web crawler is a tool used for gathering web pages that is written in a specific language. In case of gathering the Thai web pages, the simplest way is to setup a web crawler to follow only web pages according to a national domain name restriction, e.g. by specifying the ".th" domain. However, the main problem of this method is that more than half of Thai web pages are outside the ".th" domain. Therefore, the web crawler still misses many Thai web pages. This thesis proposed a language-specific web page finding method which calculates the probability scores of next links that are likely to find Thai's web pages, by using a Naïve Bayes theory to compute with the knowledge base that has been built from the first crawling. The output probability scores are then used to reorder the URLs in the crawler's priority queue, so that, the web crawler will collect the highest probability scores web pages first. According to the evaluation results, the proposed strategy achieves better harvest rate and crawling coverage than the other approaches proposed in the literature.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature