Punnawat Tadapak 2011: Machine Learning Based Thai Web Host Identification. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Assistant Professor Arnon Rungsawang, Ph.D. 79 pages.

Web pages serve as the largest knowledge bases for humans in all areas, and are continuously increased. Current limited resources, such as time, data storage, network bandwidth, make us unable to collect all web pages from the Internet. Therefore, a language specific web crawler (LSWC) is an alternative approach to help gathering the needed web pages written in a specific language. Some examples usage which apply the LSWC are; the search engine providers in the country that need to collect as most as web pages written in their specified native language to satisfy their own people's need of information, the national web archive project that needs to long term preserve their social heritage, cultural, way of life, for future generations.

This thesis proposes a prototype of the Thai language specific web site crawler. It composes of two main components; a language predictor and a web site crawler. The former employs the classifier emsemble based predictors that each different machine learning based classifier has been trained from the sample sets of host features, both positive and negative examples. This component then takes responsibility to identify the relevant web hosts in which web pages are written in Thai, including with some confident scores. The later, i.e. the web site crawler, chooses to collect the web pages from the prioritized list of relevant web hosts provided by the former. The experimental result from the real Internet data set shows that the crawling performance in term of the standard harvest rate of the proposed approach is better than those traditional LSWCs reported in the literature.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature