

3737105 SCCS/M : MAJOR : COMPUTER SCIENCE; M.Sc.(COMPUTER SCIENCE)

KEY WORDS : THAI HTML / THAI HYPERTEXT / THAI HYPERTEXT MARKUP LANGUAGE / HTML FOR THAI LANGUAGE

THEERA DURONGROENGRIT : HTML FOR THAI LANGUAGE. THESIS  
ADVISORS: SUPACHAI TANGWONGSAN, Ph.D., DAMRAS WONGSAWANG,  
Ph.D. 79P. ISBN 974-663-308-2

Searching text in Thai passages using a typical text editor or word processor nowadays cannot find the right word. For example, searching for “นก”, “กล” or “ลม” will get results in “โลกกลม” (the round world or the world is round) which does not contain any words that means “นก” (reed) or “กล” (trick) or even “ลม” (wind) at all. The problems stem from the writing system in Thai language where a sentence or phrase is written without word boundary marks like spaces in English, making it difficult for a computer program to separate Thai words correctly.

Now, Thai hypertext uses word break tags ‘<WBR>’ to separate each word in a sentence, that is for paragraph wrapping only when displayed on a screen rather than for representing word boundaries. If we try to search for Thai words in a Thai hypertext document, we still found meaningless results. In fact, the word break tag <WBR> in Hypertext Markup Language-HTML should be better regarded as syllable separator, instead of word separator. Therefore, there should be some missing ‘parts’ or links if we attempt to blend the HTML for Thai words, or better known as morphemes. This research attempts to design a tag set for Thai HTML, each morpheme in a sentence can be identified at a morphological relationship to inflect aggregate semantic understanding of morphemes.

The research includes developing a prototype for testing sample text data: both sentences and articles. The sample texts are marked up conforming to HTML standards and including new designed tags. Experiments are conducted on various types of searching, namely semantic searching, substring searching, string searching and wildcard searching. The results are satisfied and assured to be able to help develop further studying Thai text analysis or translation.