HTML FOR THAI LANGUAGE

THEERA DURONGROENGRIT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
1999

ISBN 974-663-308-2

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 WONSAWANG, 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 "กก", "กล" or "ลม" will get results in "โลกกลม" (the round world or the world is round) which does not contain any words that means "กก" (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.

3737105 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)

ธีระ ดุรงค์เริงฤทธิ์ : HTML สำหรับภาษาไทย (HTML FOR THAI LANGUAGE). คณะกรรมการควบคุมวิทยานิพนธ์ : ศุภชัย ตั้งวงศ์ศานต์, Ph.D., ดำรัส วงศ์สว่าง, Ph.D. 79 หน้า. ISBN 974-663-308-2

การค้นหาคำในภาษาไทยด้วยโปรแกรมจำพวก Text Editor หรือ Word Processing ต่าง ๆ ที่มีอยู่ ยังไม่สามารถค้นหาคำได้ถูกต้องตามความหมายที่ต้องการ เช่น เมื่อต้องการค้นหาคำ 'กก' หรือ 'กล' หรือ 'ลม' คำตอบที่ได้รับสามารถปรากฏอยู่ในคำว่า 'โลกกลม' (โลก-กลม = The world is round) ได้ ซึ่งในคำว่า 'โลกกลม' นี้ไม่ได้มีความหมายของคำ 'กก' หรือ 'กล' หรือ 'ลม' อยู่เลย ต้นเหตุของปัญหาเกิดจากลักษณะภาษาไทยที่การเขียนประโยค คำแต่ละคำในประโยคจะเขียนเรียงติดกันไป ไม่มีช่องว่างคั่นระหว่างคำเพื่อบอกขอบเขตของคำ (Word Boundary) เหมือนกับภาษาอังกฤษ ทำให้โปรแกรมเหล่านั้นไม่สามารถแยกแยะคำในภาษาไทยให้อย่างถูกต้องได้

ปัจจุบันมีการใช้ Tag '<WBR>' คั่นระหว่างคำภาษาไทยในเอกสาร HTML เพื่อช่วยแสดงผลในการจัดย่อหน้าให้สวยงามเท่านั้น ยังไม่สามารถบอกขอบเขตของคำได้ เมื่อทำการค้นคำ คำตอบที่ได้ก็อาจไม่ถูกความหมายที่ต้องการอยู่ ซึ่งความจริงการแยกคำในระดับพยางค์เป็นสิ่งที่เหมาะสมกว่าการแยกคำตามแบบที่ใช้ '<WBR>' เพื่อให้สามารถแสดงความสัมพันธ์ที่ขาดหายไปในคำภาษาไทยได้ งานวิจัยนี้ได้นำเสนอ Tags ซึ่งเป็นชุดคำสั่งที่ใช้ใน HTML เพิ่มขึ้นมาอีกจำนวนหนึ่ง เพื่อทำให้คำแต่ละคำในประโยคภาษาไทยมี Link เพื่อบอกความสัมพันธ์ระหว่างพยางค์หรือหน่วยคำที่อยู่ติดกันว่า เป็นกลุ่มของคำที่มีความหมายหนึ่งได้หรือไม่

งานวิจัยได้สร้างแบบจำลอง (Prototype) ขึ้นมาเพื่อทดสอบข้อมูลภาษาไทยที่เป็นประโยคและเป็นบทความ โดยมีรูปแบบตามเอกสาร HTML และมีการกำหนด Tags ที่สร้างขึ้นใหม่ลงไปด้วย การทดสอบจะทดสอบการค้นหาในรูปแบบต่าง ๆ ได้แก่ ค้นหาแบบ Semantic ค้นหาแบบ Substring ค้นหาแบบ String และค้นหาแบบ Wildcard ผลการทดสอบทำให้เกิดความมั่นใจที่จะนำไปขยายผลเพื่อให้สามารถทำการวิเคราะห์ภาษาไทย หรือการแปลภาษาไทยต่อไป