

b148902

การรอกคำหยาบในข้อความภาษาไทยโดยไฟไนท์สเตทแมชชีน

วุฒิชัย วิเชียรไชย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

วิทยาศาสตร์มหาบัณฑิต (สถิติประยุกต์)

คณะสถิติประยุกต์

สถาบันบัณฑิตพัฒนบริหารศาสตร์

2549

## บทคัดย่อ

ชื่อวิทยานิพนธ์	การกรองคำหยาบในข้อความภาษาไทยโดยไฟไนต์สเตทแมชชีน
ชื่อผู้เขียน	วุฒิชัย วิเชียรไชย
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
ปีการศึกษา	2549

งานวิจัยนี้ เป็นงานวิจัยที่เกี่ยวกับการพัฒนาโปรแกรมกรองคำหยาบในข้อความภาษาไทย โดยแบ่งขั้นตอนการทำงานของโปรแกรมออกเป็นสองขั้นตอนใหญ่ๆ คือ ขั้นตอนแรกเป็นการตัดคำ ขั้นตอนที่สองเป็นการค้นหาและแทนที่คำหยาบด้วยเครื่องหมายดอกจัน ซึ่งนิยามคำหยาบที่ใช้ นำมาจากพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2542

การตัดคำในข้อความภาษาไทย ในงานวิจัยนี้ใช้วิธีตัดคำด้วยพจนานุกรม แบบเทียบข้อความ (String Matching) และเทียบคำที่ยาวที่สุด (Longest Matching) และใช้ตัวแบบทางคณิตศาสตร์ไฟไนต์สเตทแมชชีนด้วยข้อมูลออก (Finite State Machines with Outputs) ในการพัฒนาโปรแกรมค้นหาและแทนที่คำหยาบในข้อความภาษาไทยด้วยเครื่องหมายดอกจัน

งานวิจัยนี้ได้ทดสอบประสิทธิภาพ โดยเอกสารจากหนังสือพิมพ์ไทยรัฐ มติชน ข่าวสด และคมชัดลึก จำนวนหนึ่งร้อยคอลัมน์ และข้อความจากเว็บอร์ดจำนวนเจ็ดร้อยเจ็ดข้อคิดเห็นในแหล่งข้อมูลอินเทอร์เน็ต พบว่าประสิทธิภาพการแทนที่คำหยาบจริงด้วยเครื่องหมายดอกจันมีความถูกต้อง 99.14%

อย่างไรก็ตามยังมีคำหยาบที่ไม่สามารถแทนที่ด้วยเครื่องหมายดอกจัน เนื่องจากการตัดคำในข้อความภาษาไทยยังไม่ถูกต้อง 100% จึงได้เสนอแนวทางปรับปรุงโดยการเพิ่มคำศัพท์ที่ไม่ปรากฏในพจนานุกรมและเพิ่มข้อความพิเศษที่ไม่ปรากฏในพจนานุกรมข้อความพิเศษ เพื่อเพิ่มประสิทธิภาพในการตัดคำและกรองคำหยาบให้สูงยิ่งขึ้น

## ABSTRACT

<b>Title of Thesis</b>	Impolite Word Filtering in Thai Text by Finite State Machines
<b>Author</b>	Mr. Vuttichai Vichanchai
<b>Degree</b>	Master of Science (Applied Statistics)
<b>Year</b>	2006

---

In this research, we present an implementation of impolite word filtering in Thai text program. The program mainly works two steps; word segmentation and searching and replacing impolite words by asterisk. The word segmentation step is completed by using string matching and longest matching method and gives the output as single word collection. Then the process of searching and replacing impolite words works by using finite state machines with outputs. The collection of impolite words used in the program refers to Royal Thai Dictionary of Buddhist Era 2542.

The testing data are taken from 100 articles of Thai newspaper; Thairath, Matichon, Khaosod and Komchadluek and 707 comments from Thai webboards. The accuracy of the program is 99.14%

However, there are some impolite words that are not properly detected and replaced owing to word segmentation phase that could not deliver 100% of correctness. The technique of word segmentation, therefore, can be potentially improved by supplementing some of undefined words and combination words that possibly exist in Thai text to the dictionary.

## กิตติกรรมประกาศ

วิทยานิพนธ์เรื่องการรongsคำหยาบในข้อความภาษาไทยโดยไฟไนท์สเตทแมชชีน สำเร็จได้เนื่องจากบุคคลหลายท่านได้กรุณาช่วยเหลือให้ข้อมูล ข้อเสนอแนะ คำปรึกษาแนะนำ ความคิดเห็นและกำลังใจ

ผู้เขียนขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. ระวีวรรณ เอื้อพันธ์วิริยะกุล รองศาสตราจารย์ ดร. วิจิต หล่อจิระชุนท์กุล และผู้ช่วยศาสตราจารย์ ดร. โอม ศรีนิล ที่ได้ให้คำชี้แนะแนวทางและตรวจสอบวิทยานิพนธ์ทุกชั้นตอน

ขอขอบพระคุณอาจารย์ทุกท่านของคณะสถิติประยุกต์ ที่ได้ถ่ายทอดความรู้ให้แก่ผู้ศึกษา และขอขอบคุณเจ้าหน้าที่คณะสถิติประยุกต์ ที่ได้ให้ความช่วยเหลือประสานงานการติดต่อเป็นอย่างดีด้วยอัธยาศัยไมตรีที่อบอุ่นเป็นกันเอง

ขอขอบคุณนักศึกษาร่วมรุ่นที่ได้ให้ความช่วยเหลือ และประสานงานตลอดช่วงเวลาที่ได้ศึกษาอยู่ที่คณะสถิติประยุกต์

ท้ายสุด ขอกราบขอบพระคุณบิดา มารดา และขอบคุณพี่น้องที่ได้ช่วยส่งเสริมสนับสนุน และเป็นกำลังใจตลอดมาให้แก่ผู้จัดทำวิทยานิพนธ์

วุฒิชัย วิเชียรไชย

เมษายน 2549