งานวิจัยนี้ เป็นงานวิจัยที่เกี่ยวกับการพัฒนาโปรแกรมกรองคำหยาบในข้อความภาษาไทย โดยแบ่งขั้นตอนการทำงานของโปรแกรมออกเป็นสองขั้นตอนใหญ่ๆ คือ ขั้นตอนแรกเป็นการตัดคำ ขั้นตอนที่สองเป็นการค้นหาและแทนที่คำหยาบด้วยเครื่องหมายดอกจัน ซึ่งนิยามคำหยาบที่ใช้นำมาจากพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2542

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

งานวิจัยนี้ได้ทดสอบประสิทธิภาพ โดยเอกสารจากหนังสือพิมพ์ไทยรัฐ มติชน ข่าวสด และคมชัดลึก จำนวนหนึ่งร้อยคอลัมน์ และข้อความจากเว็บบอร์ดจำนวนเจ็ดร้อยเจ็ดข้อคิดเห็นในแหล่งข้อมูลอินเทอร์เนต พบว่าประสิทธิภาพการแทนที่คำหยาบจริงด้วยเครื่องหมายดอกจันมีความถูกต้อง 99.14%

อย่างไรก็ตามยังมีคำหยาบที่ไม่สามารถแทนที่ด้วยเครื่องหมายดอกจัน เนื่องจากการตัดคำในข้อความภาษาไทยยังไม่ถูกต้อง 100% จึงได้เสนอแนวทางปรับปรุงโดยการเพิ่มคำศัพท์ที่ไม่ปรากฏในพจนานุกรมและเพิ่มข้อความพิเศษที่ไม่ปรากฏในพจนานุกรมข้อความพิเศษ เพื่อเพิ่มประสิทธิภาพในการตัดคำและกรองคำหยาบให้สูงยิ่งขึ้น

181702

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.