

สารบัญ

	หน้า
กิตติกรรมประกาศ ค	
บทคัดย่อภาษาไทย ง	
บทคัดย่อภาษาอังกฤษ จ	
สารบัญตาราง ฉ	
สารบัญภาพ ญ	
บทที่ 1 บทนำ	1
1.1 หลักการและเหตุผล	1
1.2 วัตถุประสงค์ของการศึกษา	3
1.3 ประโยชน์ที่คาดว่าจะได้รับจากการศึกษา	3
1.4 แผนการดำเนินการ ขอบเขตและวิธีการศึกษา	3
1.4.1 แผนการดำเนินงาน	3
1.4.2 ขอบเขตการวิจัย	3
1.4.3 วิธีการวิจัย	4
1.5 อุปกรณ์ที่ใช้ในการวิจัย	4
1.5.1 ฮาร์ดแวร์ (Hardware)	4
1.5.1 ซอฟต์แวร์ (Software)	4
1.6 สถานที่ที่ใช้ในการดำเนินการวิจัยและรวบรวมข้อมูล	4
1.7 นิยามศัพท์	5
บทที่ 2 งานวิจัยและทฤษฎีที่เกี่ยวข้อง	6
2.1 งานวิจัยที่เกี่ยวข้องเว็บสไปเดอร์ (Web spider)	6
2.2 งานวิจัยที่เกี่ยวข้องการจัดกลุ่มเอกสารข้อความ (Text Classification)	10
2.3 งานวิจัยที่เกี่ยวข้องซัพพอร์ตเวกเตอร์แมชชีน (Support Vector machines)	12
2.4 ทฤษฎีที่เกี่ยวข้องเว็บสไปเดอร์ (Web spider)	22
2.5 การเปลี่ยนทิศทางใหม่ของยูอาร์แอล (URL Redirect)	28
2.6 คำวนมน้ำหนักของคำหลักของเอกสารบนเว็บ	29

สารบัญ (ต่อ)

	หน้า
2.7 ทฤษฎีที่เกี่ยวข้องกับการจัดกลุ่มเอกสารข้อความ (Text Classification)	31
2.8 ทฤษฎีที่เกี่ยวข้องการตัดคำและการสกัดคำหลัก (Word Segmentation)	31
2.9 ทฤษฎีที่เกี่ยวข้องการสร้างตัวแทนเวกเตอร์หรือการหาหน้าหนักของคำ	32
2.10 ทฤษฎีที่เกี่ยวข้องลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน	33
2.11 การวัดประสิทธิภาพของการจัดกลุ่มเอกสาร (Measurement)	37
บทที่ 3 แนวคิดในการออกแบบของเว็บสไปเดอร์	39
3.1 แนวคิดในการออกแบบของเว็บสไปเดอร์	39
3.1.1 การวิเคราะห์โดเมน (Domain) ของ www.cnn.com	40
3.1.2 เว็บสไปเดอร์ หรือ ครอว์เลอร์	42
3.2 การเตรียมเอกสาร	44
บทที่ 4 การจัดกลุ่มเอกสารด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน	47
4.1 การจัดกลุ่มเอกสาร โดยใช้ซัพพอร์ตเวกเตอร์แมชชีน	47
4.1.1 ขั้นตอนก่อนการประมวลผล (Pre-processing)	48
4.1.2 การจัดกลุ่มเอกสาร โดยจะใช้อัลกอริทึม Support Vector Machines	56
4.1.3 การสร้างโมเดลสำหรับการจัดกลุ่มเอกสาร	59
บทที่ 5 การประเมินผลการทดลอง	60
บทที่ 6 สรุปผลและข้อเสนอแนะของการค้นคว้าแบบอิสระ	66
6.1 บทสรุป	66
6.2 ข้อเสนอแนะสำหรับการค้นคว้าแบบอิสระ	67
6.3 แนวทางการพัฒนาต่อในอนาคต	68
บรรณานุกรม	69

สารบัญ (ต่อ)

	หน้า
ภาคผนวก	
ภาคผนวก ก โปรแกรมเว็บสไปเดอร์	74
ภาคผนวก ข โปรแกรมภาษาอาร์	83
ประวัติผู้เขียน	86

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
 Copyright© by Chiang Mai University
 All rights reserved

สารบัญตาราง

ตาราง	หน้า	
2.1 ตัวอย่างการแยกส่วนประกอบของยูอาร์แอล		25
2.2 แท็กในภาษาเอชทีเอ็มแอลที่กำกับยูอาร์แอล		26
2.3 การจัดระดับความสำคัญของข้อมูลในเอกสารบนเว็บ		29
3.1 การทดสอบการใช้งานเว็บไซต์		43
4.1 ผลลัพธ์ที่ได้จากการหาความถี่ของคำที่อยู่ในเอกสารทั้งหมดและส่วนกลับของเอกสาร		53
4.2 ผลลัพธ์ที่ได้จากการหาหน้าหนึ่งของคำในเอกสาร		54
4.3 การหาความถี่ของคำ		57
4.4 ตัวอย่างเอกสารสำหรับการสร้างโมเดลสำหรับการจัดกลุ่มเอกสาร		59
5.1 ผลการทดสอบโมเดลการจัดกลุ่มเอกสาร โดยใช้ซอฟต์แวร์เวกเตอร์แมชชีน		61
5.2 วัดประสิทธิภาพการจัดกลุ่มเอกสาร โดยใช้ซอฟต์แวร์เวกเตอร์แมชชีน		62

สารบัญภาพ

รูป	หน้า
1.1 การทำงานของเว็บสไปเดอร์	1
2.1 การจัดโครงสร้างแบบลำดับชั้นและการคำนวณตัวประมาณค่าโดยใช้เทคนิคการรวมกันของตัวประมาณค่าความน่าจะเป็นของคำ	11
2.2 แบบจำลองพื้นฐานของครอว์เลอร์	24
2.3 กลไกการตรวจสอบยูอาร์แอลก่อนนำไปใส่ในยูอาร์แอลคิว	25
2.4 กลไกการร้องขอเว็บเพจจากเซิร์ฟเวอร์	26
2.5 แบบจำลองครอว์เลอร์แบบขนาน	27
2.6 การเปลี่ยนทิศทางใหม่ของยูอาร์แอลโดยใช้ Meta tag ของ HTML	28
2.7 การเปลี่ยนทิศทางใหม่ของยูอาร์แอลโดยการหน่วงเวลา	29
2.8 ตัวอย่างชื่อหัวข้อของเอกสาร (Tag Title) ที่อ่านมาจากเอกสาร HTML	30
2.9 แสดง Tag Meta ที่คัดลอกมาจากเอกสาร HTML	30
2.10 แสดง Tag Anchor ที่คัดลอกมาจากเอกสาร HTML	30
2.11 แสดง Tag Body ที่คัดลอกมาจากเอกสาร HTML	30
2.12 การแบ่งข้อมูลโดยซัพพอร์ตเวกเตอร์แมชชีน	34
2.13 พื้นที่ข้อมูลที่แบ่งแยกไม่ได้ด้วยวิธี OAO SVM	36
3.1 สถาปัตยกรรมของระบบ	39
3.2 ตัวอย่างโดเมน “www.cnn.com”	41
3.3 วิธีการเดินทางเพื่อเก็บรวบรวมข้อมูลของสไปเดอร์หรือครอว์เลอร์	41
3.4 เว็บสไปเดอร์	42
3.5 ตัวอย่างการเก็บข้อมูลในลักษณะ HTML	43
3.6 ตัวอย่างก่อนการวิเคราะห์คำในประโยคของเอกสาร	45
3.7 ตัวอย่างหลังการวิเคราะห์คำในประโยคของเอกสาร	46
4.1 ขั้นตอนการของงานวิจัย	47
4.2 ขั้นตอนก่อนการประมวลผล	48
4.3 ตัวอย่างหลังการวิเคราะห์คำในประโยคของเอกสารจากบทที่ 3	49

สารบัญภาพ (ต่อ)

รูป	หน้า
4.4 ผลที่ได้จากการกำจัดเครื่องหมายต่างๆ และเปลี่ยนอักษรตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็ก	49
4.5 ผลที่ได้จากการกำจัดคำศัพท์	49
4.6 ผลที่ได้จากการหารากศัพท์	50
4.7 ผลที่ได้จากการสร้างคำสำคัญ-เอกสาร	50
4.8 การแบ่งกลุ่มด้วย Maximum Margin Hyperplane	58
5.1 เอกสารกลุ่มข่าวกีฬา	62
5.2 เอกสารกลุ่มข่าวสุขภาพ	63
5.3 เอกสารกลุ่มข่าวคอมพิวเตอร์	64
5.4 ประสิทธิภาพของระบบ	65