



สำนักพิมพ์มหาวิทยาลัยราชภัฏนครราชสีมา

ฉบับพิมพ์ครั้งที่ 2
แก้ไขเพิ่มเติม

การค้นคืนสารสนเทศ Information Retrieval

ดร.วิรัช ศรีเลิศล้ำวานิช

การค้นคืนสารสนเทศ

Information Retrieval



ถ่ายเอกสารแทนการใช้หนังสือ
คือการทำลายภูมิปัญญาสร้างสรรค์

การค้นคืนสารสนเทศ Information Retrieval

ฉบับพิมพ์ครั้งที่ 2 แก้ไขเพิ่มเติม

ดร.วิรัช ศรีเลิศล้ำวาณิช
สถาบันเทคโนโลยีนานาชาติสิรินธร
มหาวิทยาลัยธรรมศาสตร์

สำนักพิมพ์มหาวิทยาลัยธรรมศาสตร์

2561

วิรัช ศรีเลิศล้ำวานิช.

การค้นคืนสารสนเทศ = Information Retrieval.

1. การค้นสารสนเทศ. 2. ระบบการจัดเก็บและค้นสารสนเทศ.
3. การสืบค้นข้อมูล.

Z667

ISBN 978-616-314-403-4

ISBN (E-BOOK) 978-616-314-416-4

ลิขสิทธิ์ของ**ดร.วิรัช ศรีเลิศล้ำวานิช**

สงวนลิขสิทธิ์

ฉบับพิมพ์ครั้งที่ 2 เดือนสิงหาคม 2561

จำนวน 200 เล่ม

ฉบับอิเล็กทรอนิกส์ (e-book) กันยายน 2561

จัดพิมพ์และจำหน่ายโดย**สำนักพิมพ์มหาวิทยาลัยธรรมศาสตร์**

ท่าพระจันทร์: อาคารธรรมศาสตร์ 60 ปี ชั้น U1 มหาวิทยาลัยธรรมศาสตร์

ถนนพระจันทร์ กรุงเทพฯ 10200 โทร. 0-2223-9232

ศูนย์รังสิต: อาคารโคมบริหาร ชั้น 3 ห้อง 317 มหาวิทยาลัยธรรมศาสตร์

ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12121

โทร. 0-2564-2859-60 โทรสาร 0-2564-2860

<http://www.thammasatpress.tu.ac.th>, e-mail: unipress@tu.ac.th

พิมพ์ที่**ห้างหุ้นส่วนจำกัด เอ เอ เอ เซอร์วิส**

นายอาทิตย์ พงษ์ภัทรวิทย์ ผู้พิมพ์ผู้โฆษณา

พิมพ์ครั้งที่ 1 เดือนกุมภาพันธ์ 2560 จำนวน 100 เล่ม

พิมพ์ครั้งที่ 1 เดือนธันวาคม 2560 จำนวน 70 เล่ม (ฉบับพิมพ์เพิ่ม)

พิมพ์ครั้งที่ 2 เดือนสิงหาคม 2561 จำนวน 200 เล่ม

ราคาเล่มละ 140.- บาท

สารบัญ

คำนำ	(11)
บทที่ 1 การค้นคืนแบบบูลีน (Boolean Retrieval)	1
1.1 ปัญหาของการค้นคืน	5
1.2 การสร้างบัญชีดัชนีเอกสาร (Postings list) หรือบัญชีผกผัน (Inverted list)	10
1.3 การคำนวณแบบบูลีน	13
1.4 การลำดับผลการค้นคืน	15
1.5 คำถามท้ายบท	16
บทที่ 2 ชุดคำศัพท์และบัญชีดัชนีเอกสาร (Term Vocabulary and Postings Lists)	17
2.1 การแจงเอกสาร	18
2.2 ตัดคำ (Word segmentation, Tokenization)	20
2.3 สร้างดัชนีเอกสาร	26
2.4 การสร้างดัชนีตำแหน่งกับคำค้นวลี (Positional indexes and phrase queries)	31
2.4.1 การสร้างดัชนีคู่คำ (Biword index)	31
2.4.2 การสร้างดัชนีตำแหน่ง (Positional indexes)	32
2.5 คำถามท้ายบท	33

บทที่ 3	การค้นคืนอย่างครอบคลุม (Tolerant Retrieval)	35
3.1	รูปแบบการจัดเก็บดัชนีกับการค้นคืน	36
3.1.1	ต้นไม้แบบทวิภาค (Binary tree)	37
3.1.2	B-Tree	40
3.2	ข้อความด้วยอักขระตัวแทน (Wildcard queries)	43
3.2.1	การสร้างดัชนีสำหรับการค้นคืนด้วยอักขระตัวแทน (General wildcard queries)	43
3.2.2	การสร้างดัชนีแบบ K-gram (K-gram indexes for wildcard queries)	44
3.3	การแก้ไขตัวสะกด (Spelling Correction)	45
3.3.1	ระยะแก้ไข (Edit Distance)	45
3.3.2	การสร้างดัชนีแบบ K-gram สำหรับงานการแก้ไขปัญหาตัวสะกด	47
3.3.3	การเข้ารหัสคำที่มีเสียงคล้าย (Soundex)	48
3.4	คำถามท้ายบท	51
บทที่ 4	การสร้างดัชนีและการบีบอัดดัชนี (Index Construction and Index Compression)	53
4.1	การจัดเรียงผลงาน (Merge Sort)	54
4.2	แมพ-รีดิวส์ (MapReduce)	57
4.3	การบีบอัดดัชนี (Index Compression)	59
4.3.1	กฎของฮีปส์ (Heaps' Law)	60
4.3.2	กฎของซิปฟ์ (Zipf's Law)	61
4.4	การบีบอัดไฟล์บัญชีดัชนีเอกสาร (Postings list file compression)	62
4.4.1	รหัสไบต์ผันแปร (Variable byte codes)	63
4.4.2	รหัสแกมมา (Gamma code)	64
4.5	คำถามท้ายบท	66

บทที่ 5 การกำหนดคะแนนหรือน้ำหนักสำหรับคำสำคัญ

(Scoring and Term Weighting)	67
5.1 ค่าสัมประสิทธิ์แจ็กการ์ด (Jaccard coefficient)	68
5.2 คำสำคัญที่เป็นตัวแทนของเอกสาร	69
5.2.1 การกำหนดคะแนนตามตำแหน่งที่ปรากฏในเอกสาร (Weighted zone scoring)	69
5.2.2 ค่าความถี่และน้ำหนักของคำ (Term frequency)	70
5.2.3 ค่าความถี่เอกสารผกผัน (Inverse document frequency)	73
5.2.4 คำนน้ำหนัก Tf-idf	74
5.3 แบบจำลองปริภูมิเวกเตอร์ (Vector space model)	74
5.3.1 ความคล้ายคลึงโคไซน์ (Cosine similarity)	75
5.3.2 รูปแบบการกำหนดน้ำหนักความสัมพันธ์ ระหว่างเอกสารกับคำค้น	77
5.4 คำถามท้ายบท	81

บทที่ 6 การประเมินความสามารถของการค้นคืน

(Retrieval Evaluation Measure)	83
6.1 การประเมินผลระบบค้นคืน	85
6.2 การประเมินผลการค้นคืนจากผลที่ไม่มีการจัดลำดับ	87
6.3 การประเมินผลการค้นคืนจากผลที่มีการจัดลำดับ	92
6.3.1 ความแม่นยำประมาณการเฉลี่ย (Interpolated average precision)	94
6.3.2 ความแม่นยำประมาณการเฉลี่ย 11 จุด (11-point Interpolated average precision)	94
6.3.3 ค่าเฉลี่ยของค่าเฉลี่ยความแม่นยำ (Mean Average Precision, MAP)	96
6.4 การประเมินความคล่องจองของเอกสารที่เกี่ยวข้อง	96
6.5 คำถามท้ายบท	98

บทที่ 7	การตอบสนองของความเกี่ยวเนื่องและการขยายคำค้น	99
	(Relevance Feedback and Query Expansion)	
7.1	เวกเตอร์ของเอกสาร (Document Vector)	100
7.2	การตอบสนองของความเกี่ยวเนื่องกับการตอบสนองของความเกี่ยวเนื่องเทียม (Relevance Feedback and Pseudo Relevance Feedback)	102
7.2.1	อัลกอริทึมร็อกคิโอ (Rocchio algorithm) สำหรับการตอบสนอง ของความเกี่ยวเนื่อง (Relevance Feedback)	103
7.2.2	การตอบสนองของความเกี่ยวเนื่องเทียม (Pseudo Relevance Feedback)	105
7.3	การขยายคำค้น (Query Expansion)	106
7.3.1	การจัดลำดับผลลัพธ์ที่ได้จากการขยายคำค้น	113
7.4	คำถามท้ายบท	116
บทที่ 8	การจำแนกเอกสาร (Text Classification)	117
8.1	ทฤษฎีของเบส์ (Bayes' theorem)	118
8.2	การจำแนกเอกสารแบบ Naïve Bayes	119
8.2.1	Multinomial Naïve Bayes	120
8.2.2	Bernoulli Naïve Bayes	123
8.3	การคัดเลือกคุณลักษณะ (Feature Selection)	125
8.3.1	การทดสอบไคสแควร์ (χ^2 หรือ Chi Square)	126
8.3.2	มิวชัลอินโฟเมชัน (Mutual Information)	129
8.3.3	การคัดเลือกคุณลักษณะตามค่าความถี่	131
8.4	คำถามท้ายบท	132

บทที่ 9	การจำแนกปริภูมิเวกเตอร์ (Vector Space Classification)	133
9.1	การจำแนกปริภูมิเวกเตอร์ของร็อกคิโอ (Rocchio vector space classification)	134
9.2	การจำแนกเอกสารด้วยวิธี kNN (K Nearest Neighbor Classification)	136
9.3	การจำแนกเอกสารด้วยวิธีการของ SVM (Support Vector Machine Classification)	143
9.3.1	การคำนวณระยะห่างระหว่างระนาบ +1 กับ -1	145
9.3.2	การจำแนกที่มีระยะห่างของขอบแบ่งเขตที่มากที่สุด	147
9.3.3	ตัวอย่างการคำนวณหาค่า SVM	148
9.4	คำถามท้ายบท	150
บทที่ 10	การจัดกลุ่มเอกสาร (Text Clustering)	151
10.1	เค-เฉลี่ย (K-means)	153
10.2	K-medoids	157
10.2.1	ระยะแมนแฮตตัน (Manhattan distance)	157
10.2.2	การจัดกลุ่มเอกสารแบบ K-medoids	158
10.3	การประเมินผลการจัดกลุ่มเอกสาร	160
10.3.1	วิธีการ Purity	161
10.3.2	วิธีการ Normalized Mutual Information (NMI)	162
10.3.3	วิธีการ Rand Index (RI)	163
10.3.4	วิธีการ F-Measure	165
10.4	การประยุกต์ใช้งาน	166
10.5	คำถามท้ายบท	169

บทที่ 11 การจัดกลุ่มเอกสารแบบมีลำดับชั้น (Hierarchical Text Clustering)	171
11.1 การจัดกลุ่มเอกสารแบบผสม (Agglomerative Text Clustering)	173
11.2 ความคล้ายคลึงของเอกสารระหว่างกลุ่ม (Inter-similarity)	175
11.2.1 การเชื่อมเดี่ยว (Single link)	176
11.2.2 การเชื่อมสมบูรณ์ (Complete link)	177
11.2.3 ค่าเฉลี่ยกลุ่ม (Group Average)	180
11.2.4 เซนทรอยด์ (Centroid)	180
11.2.5 เปรียบเทียบรูปแบบการผสมสำหรับการจัดกลุ่มเอกสาร	181
11.3 การจัดกลุ่มเอกสารแบบแบ่งแยก (Divisive Text Clustering)	183
11.4 คำถามท้ายบท	184
 เอกสารอ้างอิงและเอกสารอ่านประกอบ	185
อภิธานศัพท์ (อังกฤษ-ไทย)	200
อภิธานศัพท์ (ไทย-อังกฤษ)	208
ประวัติผู้แต่ง	216

คำนำ

ตำราภาษาไทยมีจำกัดมากโดยเฉพาะตำราสำหรับวิชาการทางด้านวิทยาการคอมพิวเตอร์ ทำให้นักศึกษามีโอกาสทำความเข้าใจทางด้านเนื้อหาได้อย่างจำกัด วิทยาการด้านคอมพิวเตอร์มีพัฒนาการและการเปลี่ยนแปลงที่เร็วมาก จนบางครั้งก็เคยคิดว่าถ้าจะตามวิทยาการทางด้านนี้ให้ทัน ก็จำเป็นต้องอ่านตำราที่เป็นภาษาอังกฤษที่เป็นสากลและมีผู้แต่งจำนวนมากเท่านั้น อย่างไรก็ตามเนื้อหาส่วนใหญ่ก็จะประกอบไปด้วยบริบทที่เป็นภาษาอังกฤษและมีตัวอย่างการประยุกต์ใช้ที่แตกต่างจากภาษาไทยเป็นอย่างมาก เมื่อวิทยาการการค้นคว้าข้อมูลมีส่วนเกี่ยวข้องกับการใช้ภาษาอยู่มาก ดังนั้นการแต่งตำราเป็นภาษาไทยน่าจะเป็นประโยชน์สำหรับคนอ่านได้มากกว่า ได้เห็นตัวอย่างที่เป็นรูปธรรมโดยไม่ต้องแปลเทียบ และเป็นการเพิ่มทางเลือกในการอ่านสำหรับตำราภาษาไทยที่มีอยู่น้อยมากเมื่อเทียบกับภาษาอังกฤษหรือภาษาอื่นที่เป็นภาษาหลัก อย่างไรก็ตามการใช้คำศัพท์ทางวิชาการที่เป็นภาษาไทยก็ยังเป็นอุปสรรคอย่างยิ่งในการหาคำศัพท์ที่เหมาะสมสำหรับการอ่านทำความเข้าใจได้อย่างสะดวก และหลายครั้งก็จำเป็นที่จะต้องกำกับด้วยภาษาอังกฤษที่เป็นต้นฉบับของความคิดและทฤษฎีนั้นๆ ดังนั้นจึงแนะนำให้ทำความเข้าใจในคำศัพท์ที่เป็นภาษาต้นฉบับด้วย โดยตั้งใจอย่างยิ่งที่จะให้คำอธิบายที่เป็นพื้นฐานแห่งความคิดและทฤษฎีที่กระจ่างที่สุด

งานเขียนหนังสือเล่มนี้ได้เริ่มขึ้นจากความตั้งใจที่จะให้ผู้อ่านได้มีความเข้าใจที่ถ่องแท้ จึงได้รวบรวมและแปลงเป็นการอธิบายที่ครอบคลุมและให้ข้อมูลที่มาที่ไปของหลักการมากที่สุดเท่าที่จะทำได้ โดยมีตัวอย่างและข้อมูลที่เกิดจากงานวิจัยประกอบ เนื้อหาจึงได้มาจากการค้นคว้าจากหลายแหล่งและจากบทความวิจัยที่ผ่านมาโดยเฉพาะในส่วนที่เกี่ยวข้องกับการประมวลผลภาษาไทย ผลการประยุกต์ใช้ในการอธิบายจึงน่าจะสะท้อนปัญหาของการเตรียมการสำหรับการค้นคืนได้อย่างตรงประเด็นที่สุด

หนังสือเล่มนี้ประกอบด้วยเนื้อหาที่เป็นพื้นฐานของ “การค้นคืนสารสนเทศ” เริ่มตั้งแต่ความจำเป็น ลักษณะเฉพาะทางภาษาที่มีผลต่อการสร้างดัชนีของคำค้น การขยายคำค้นด้วยวิธีการต่างๆ เพื่อให้ได้ผลลัพธ์ที่มีความครอบคลุมมากยิ่งขึ้น การเพิ่มประสิทธิภาพของการค้นคืน และที่เป็นหลักทฤษฎีที่เป็นหัวใจของการค้นคืน อันได้แก่ การกำหนดคะแนนหรือน้ำหนักสำหรับคำสำคัญ ไม่ว่าจะเป็นการใช้ค่าความถี่หรือค่าน้ำหนักของคำ ที่คำนวณได้จากค่า tf-idf และการใช้วิธีการของการจำแนกเอกสาร รวมถึงการประมวลผลเอกสารในรูปของปริภูมิเวกเตอร์ (Vector space) ทั้งนี้จะให้ข้อมูลที่เป็นพื้นฐานเพื่อความเข้าใจในทฤษฎีอย่างครบถ้วน

ขอแสดงความขอบคุณสำหรับนักศึกษาปริญญาโทและเอก ได้แก่ เรือเอกภัทร โชติกะพุกกะนะ, ภาณุวัฒน์ อัครวิจิตรเพ็ชร, อภินันท์ กวางแก้ว, วรณรัชฎ์ วิริยะวิทย์, เสฏฐวุฒิ เกียรติศักดิ์โสภณ, กิตติยา สุริยฉาย, และติมาภักดิ์ บวรกิจบำรุง ที่ได้ให้ความช่วยเหลือในการสร้างตัวอย่างและตรวจสอบรายละเอียดของบทความ

ดร.วิรัช ศรีเลิศล้ำวาณิช

การค้นคืนแบบบูลีน (Boolean Retrieval)

การค้นคืนสารสนเทศในความหมายที่แคบลงจะเป็นการค้นคืนเอกสาร ซึ่งเป็นกิจกรรมที่เราๆ จะต้องเผชิญอยู่เป็นประจำในการดำรงชีพปัจจุบันนี้ ไม่ว่าจะเป็นชีวิตส่วนตัวที่ต้องการค้นหาเอกสาร หรือข้อมูลที่ต้องการ หรือการค้นหาข้อมูลแหล่งท่องเที่ยว ที่พัก หรือแม้แต่การจะซื้อหาของใช้ในปัจจุบันก็ไม่สามารถจะหลีกเลี่ยงการค้นคืนสารสนเทศได้ หรือว่าจะเป็นชีวิตการทำงานที่มีเอกสารล้นมือ เนื่องจากความสะดวกในการเขียนเอกสารมีมากขึ้นทุกอย่างสามารถที่จะบันทึกได้ทุกแห่ง ด้วยความหลากหลายของโปรแกรมประยุกต์ เอกสารจึงอาจถูกจัดเก็บไว้ในหลากหลายรูปแบบและมักจะเป็นฐานข้อมูลเพื่อความสะดวกในการค้นคืนรูปแบบของเอกสารหลักๆ ก็สามารที่จะแบ่งออกตามลักษณะโครงสร้างการจัดเก็บได้ ดังสามารถดูแบบต่อไปนี้

• ข้อมูลแบบมีโครงสร้าง (Structured Data)

เป็นข้อมูลที่ถูกจัดเก็บโดยมีโครงสร้างที่แสดงความสัมพันธ์ระหว่างส่วนประกอบต่างๆ อย่างชัดเจน ส่วนใหญ่จะถูกจัดเก็บตามรูปแบบของระบบฐานข้อมูล (DBMS—Database Management System หรือ RDBMS—Relational Database Management System) ที่ใช้กันอย่างแพร่หลาย และจะมีมาตรฐานของภาษาที่ใช้ในการติดต่อ เรียกว่า SQL (Structured Query Language) ฐานข้อมูลจะถูกจัดเก็บเป็นตารางแสดงความสัมพันธ์ระหว่างแถวกับสดมภ์ เช่น

ตารางที่ 1.1 ตัวอย่างฐานข้อมูลลูกค้า (Customer)

ID	Name	City
1	กมล	เชียงใหม่
2	ชาญชัย	กำแพงเพชร
3	สรเสริญ	ปราจีนบุรี
4	อารี	ตรัง

ซึ่งแต่ละแถวสามารถแสดงเป็นรูปแบบของเรคคอร์ด (record) และส่วนประกอบที่เป็นฟิลด์ (field) ได้ดังนี้

Customer (ID: 1 Name: กมล City: เชียงใหม่)	Customer (ID: 2 Name: ชาญชัย City: กำแพงเพชร)	
--	---	-------

รูปที่ 1.1 ตัวอย่างการแสดงผลฐานข้อมูลในรูปแบบเรคคอร์ดและฟิลด์

โดยที่มี Customer เป็นชื่อของเรคคอร์ด ID, Name, City เป็นชื่อของฟิลด์ และ 1, กมล, เชียงใหม่ เป็นค่าของฟิลด์นั้นๆ

ข้อมูลแบบมีโครงสร้างนี้จะระบุแต่ละส่วนประกอบอย่างชัดเจน ทำให้สามารถที่จะค้นคืนได้อย่างมีประสิทธิภาพ และภาษา SQL ก็รองรับการใช้เงื่อนไขต่างๆ ในการค้นคืนได้

• ข้อมูลแบบกึ่งมีโครงสร้าง (Semi-Structured Data)

เป็นข้อมูลที่มีการกำกับองค์ประกอบและความสัมพันธ์ระหว่างองค์ประกอบในข้อความหรือเอกสารเป็นบางส่วน ซึ่งจะนำมากำกับเป็นรูปแบบของไฮเปอร์เท็กซ์ (Hypertext) ปัจจุบันข้อมูลประเภทนี้จะใช้ในการเผยแพร่และแลกเปลี่ยนกันเป็นส่วนใหญ่ในการสื่อสารทางอินเทอร์เน็ต เนื่องจากความแพร่หลายของการพัฒนามาตรฐานโครงสร้างข้อมูลสำหรับการสื่อสาร เพื่อความสะดวกและควมมีประสิทธิภาพในการแลกเปลี่ยนข้อมูล เช่น HTML (HyperText Markup Language), XML (Extensible Markup Language), JSON (JavaScript Object Notation) เป็นต้น จากที่ HTML เป็นภาษาที่ใช้กันอย่างแพร่หลายเมื่อเว็บเบราว์เซอร์ที่พัฒนาขึ้นได้กำหนดใช้ HTML เป็นมาตรฐานในการสร้างเว็บเพจ (Webpage) ที่มีลักษณะเป็นไฮเปอร์เท็กซ์ (Hypertext) สำหรับการเชื่อมโยงเว็บเพจเข้าด้วยกันบนโปรโตคอลที่เป็น HTTP (Hypertext Transfer Protocol) สำหรับการให้บริการบนอินเทอร์เน็ตที่เรียกว่า เวิลด์ไวด์เว็บ (World Wide Web—WWW) และ XML ก็เป็นภาษาสำหรับการกำกับข้อมูลที่มีความยืดหยุ่นมากขึ้นโดยที่ผู้ใช้สามารถที่จะนิยามองค์ประกอบและความสัมพันธ์ขึ้นมาใหม่ได้ ส่วน JSON ก็เป็นตัวอย่างมาตรฐานโครงสร้างข้อมูลที่นำมาใช้ทดแทน XML เพื่อที่จะให้มีประสิทธิภาพในการสื่อสารที่สูงขึ้น มีความสิ้นเปลืองในการส่งผ่านข้อมูลที่ลดลง

```
<p>
นายกมลต้องการซื้อคอนโดที่เชียงใหม่ราคา 2.5 ล้านบาท ในต้นปี 2559 นี้<br />
จึงขอทำเรื่องกู้เงินจาก<a href = "http:// http://www.bankthailand.info/BankInThailand.htm">
ธนาคารกสิกรไทย</a>จำนวน 1.5 ล้านบาท โดยจะผ่อนชำระในเวลา 10 ปี
</p>
```

รูปที่ 1.2 ตัวอย่างข้อมูลในรูปแบบ HTML

และข้อมูลแบบกึ่งมีโครงสร้างที่พบกันบ่อยก็เป็นข้อความในรูปแบบของอีเมล ข้อมูลหลักๆ ที่มีการกำกับจะเป็น วันที่ (Date) ชื่อเรื่อง (Subject) ชื่อผู้ส่ง (From) ชื่อผู้รับ (To) และประเภทเนื้อหา (Content-Type)

```
MIME-Version: 1.0
Received: by 10.36.105.85 with HTTP; Sat, 14 May 2016 22:13:29 -0700 (PDT)
Reply-To: kamol@worldmail.com
Date: Sun, 15 May 2016 12:13:29 +0700
Delivered-To: kamol@worldmail.com
Message-ID: <CAEPNvEAVTKytQ0BwHKF-N26iJ_dqAOvP_jQ0SqJJ4fgVMbRmHw@mail.gmail.com>
Subject: Home loan
From: Kamol <kamol@worldmail.com >
To: "banker@worldmail.com" <banker@worldmail.com>
Content-Type: text/plain; charset=UTF-8
```

เรียน นายธนาคาร

นายกมลต้องการซื้อคอนโดที่เชียงใหม่ราคา 2.5 ล้านบาท ในต้นปี 2559 นี้
จึงขอทำเรื่องกู้เงินจากธนาคารกสิกรไทยจำนวน 1.5 ล้านบาท โดยจะผ่อนชำระในเวลา 10 ปี

ด้วยความเคารพ

กมล

รูปที่ 1.3 ตัวอย่างข้อความในอีเมล

• ข้อมูลแบบไม่มีโครงสร้าง (Unstructured Data)

เป็นข้อมูลที่อยู่ในลักษณะที่เป็นข้อมูลดิบ (raw data) หรือเป็นข้อมูลที่ไม่ได้มีการระบุโครงสร้างที่แน่นอน หากเป็นข้อความก็เป็นข้อความดิบ (raw text) ที่มีแต่เฉพาะเนื้อหา ไม่ได้มีการระบุโครงสร้างของข้อความที่ชัดเจน การประมวลผลจึงจำเป็นที่จะต้องมีการรู้จำองค์ประกอบหรือโครงสร้างเอง ไม่ว่าจะมาจากการเรียนรู้จากตัวอย่างที่ผ่านมา (Supervised learning) หรือการเรียนรู้ที่ได้จากการคัดแยกเองจากการคำนวณในรู้แบบที่น่าจะเป็น (Unsupervised learning) ข้อความส่วนใหญ่จะเป็นข้อมูลแบบไม่มีโครงสร้างเนื่องจากสร้าง

ได้ง่าย โดยเฉพาะที่เป็นข้อความ เนื่องจากไม่จำเป็นต้องเรียนรู้หรือใช้เครื่องมือพิเศษใดๆ ในการเขียนข้อความ เพราะเป็นภาษาที่ทุกคนเขียนได้และเข้าใจ แต่สำหรับการประมวลผลทางคอมพิวเตอร์แล้วจำเป็นที่จะต้องใช้เทคนิคหลากหลายเพื่อที่จะเข้าใจโครงสร้างและองค์ประกอบ ซึ่งเป็นความจำเป็นพื้นฐานสำหรับการค้นคืนข้อมูล

นายกมลต้องการซื้อคอนโดที่เชียงใหม่ราคา 2.5 ล้านบาท ในต้นปี 2559 นี้
จึงขอทำเรื่องกู้เงินจากธนาคารกสิกรไทยจำนวน 1.5 ล้านบาท โดยจะผ่อนชำระในเวลา 10 ปี

รูปที่ 1.4 ตัวอย่างข้อมูลแบบไม่มีโครงสร้าง

ไม่ว่าข้อมูลจะถูกเก็บไว้ในลักษณะไหน กระบวนการการค้นคืนต้องสามารถที่จะให้ผู้ใช้งานระบุความต้องการของตนเองโดยการนำเสนอในลักษณะข้อมูลที่ระบบสามารถที่จะประมวลผลได้ เนื่องจากการค้นคืนมาจากความต้องการของผู้ใช้ ซึ่งจำเป็นต้องมีการตีความหมายให้ออกมาเป็นคำพูดหรือชุดคำศัพท์ที่สามารถใช้ในการประมวลผลการค้นคืนต่อไป ดังนั้นผลที่ได้ก็จะเป็นข้อความที่สามารถสร้างความพึงพอใจให้กับผู้ใช้ได้มากน้อยเพียงใด อย่างไรก็ตามผลความพึงพอใจก็ยังคงต้องขึ้นอยู่กับดุลยพินิจของแต่ละคนที่อาจแตกต่างกันได้

การค้นคืนแบบบูลีนจึงเป็นการค้นคืนเอกสารที่ประเมินจากการมีอยู่ของชุดคำที่แปลงมาจากความต้องการของผู้ใช้ ว่ามีปรากฏอยู่ในเอกสารมากน้อยเพียงไร หรืออีกนัยหนึ่งก็เป็นการวัดความใกล้เคียงระหว่างชุดคำที่ใช้ในการค้นคืนกับเอกสารที่มีอยู่ในฐานข้อมูล โดยคำนึงถึงเฉพาะการปรากฏอยู่ของชุดคำ แต่ไม่ได้คำนึงถึงความถี่ของการเกิดของคำในเอกสาร

1.1 ปัญหาของการค้นคืน

เนื่องจากผลสัมฤทธิ์ของการค้นคืนนั้นคือค่าความพึงพอใจของผู้ใช้ ดังนั้นจึงเป็นการยากที่จะวัดเปรียบเทียบคุณภาพของรูปแบบที่ใช้ในการค้นคืน เมื่อแปลงความต้องการมาเป็นชุดคำที่ใช้ในการค้นคืน ปัญหาของการค้นคืนจึงเป็นเรื่องของคุณภาพในการค้นหาคำตอบวัดค่าความใกล้เคียงด้วยรูปแบบที่เหมาะสม และประสิทธิภาพในการค้นคำตอบให้ได้ในระยะเวลาที่จำกัดในกรณีที่มีเอกสารที่จะต้องทำการเปรียบเทียบอยู่เป็นจำนวนมาก และสิ่งที่ตามมา

เนื่องจากปริมาณเอกสารที่มีอยู่มากแล้ว การลำดับความใกล้เคียงของเอกสารต่อคำค้นจึงมีความหมายขึ้นมาก

ภายใต้สถานการณ์ทั่วไปเอกสารจะมีอยู่เป็นจำนวนมากและมีการเปลี่ยนแปลงอยู่เสมอ ดังนั้นปัญหาของการค้นคืนจึงพอจะให้ข้อสังเกตได้ดังนี้

- ปริมาณเอกสารที่มีอยู่เป็นจำนวนมากและมีการเปลี่ยนแปลงอยู่เสมอ การค้นคืนจึงจำเป็นต้องคำนึงถึงประสิทธิภาพในการประมวลผล การจัดเก็บ และการปรับแต่ง
- เงื่อนไขที่จะสามารถใช้ในการระบุคำค้น การระบุการมีอยู่ของคำอย่างเดียวยังไม่เพียงพอ บางครั้งจำเป็นที่จะต้องระบุระยะห่างระหว่างคำด้วยเพื่อให้คำเหล่านั้นอยู่ในบริบทเดียวกันด้วย เช่น ประโยค หรือวลีเดียวกัน เป็นต้น โดยเฉพาะคำในภาษาไทยที่อาจมีคำแทรก การลดรูปหรือคำสะกดต่างอยู่ค่อนข้างบ่อย เช่น “การออกกำลังกาย” ซึ่งมักจะใช้ไม่แตกต่างไปจากคำว่า “ออกกำลังกาย” หรือ การลดรูปของคำว่า “รถสีแดง” เป็น “รถแดง” “พิพิธภัณฑสถานแห่งชาติเจ้าสามพระยา” เป็น “พิพิธภัณฑเจ้าสามพระยา”
- การเรียงลำดับความใกล้เคียงของเอกสารที่มีต่อคำค้น ในบางครั้งเมื่อมีผลลัพธ์เป็นจำนวนมาก การพิจารณาแต่ละเอกสารจะใช้เวลานาน และโดยพฤติกรรมของผู้ใช้แล้วย่อมที่จะทบทวนผลลัพธ์ได้เฉพาะเอกสารที่อยู่ต้นๆ จำนวนหนึ่งเท่านั้น เช่น Google อาจคืนผลกลับมาเป็นสามสิบกว่าล้านเว็บไซต์สำหรับคำค้นคำว่า “รถ” แต่ผู้ใช้จะสามารถดูไปได้สักกี่หน้า

ดังนั้นในขั้นแรกสำหรับการค้นคืนเอกสารนั้นจำเป็นต้องเตรียมเอกสารสำหรับการค้นคืนด้วยการถอดเอกสารออกมาเป็นตารางความสัมพันธ์ระหว่างคำกับเอกสาร (Term-document matrix) ค่าที่ปรากฏในตารางนั้นจะเป็น 0 หรือ 1 เท่านั้น โดยที่ 0 หมายความว่าคำๆ นั้นไม่ได้ปรากฏอยู่ในเอกสาร ส่วน 1 หมายความว่าคำๆ นั้นปรากฏอยู่ในเอกสาร

ตารางที่ 1.2 ตารางความสัมพันธ์ระหว่างคำกับเอกสาร (Term-document matrix) ค่าของสมาชิกในเมทริกซ์ $(t, d) = 0$ เมื่อไม่ปรากฏคำ (t) ในเอกสาร (d)

	เศรษฐกิจ	บันเทิง	ต่างประเทศ	วิถีชีวิต	การเมือง	สังคม	กีฬา
อัตราดอกเบี้ย	1	0	0	0	0	1	0
ช่อง7	0	1	0	0	1	1	0
แผ่นดินไหว	0	0	1	1	1	1	0
แคลอรี	0	0	0	1	0	0	0
ร่างรัฐธรรมนูญ	0	0	1	1	1	1	0
ยาบ้า	0	0	0	0	0	1	0
เลสเตอร์ซิตี	0	0	0	0	1	1	1

ตารางความสัมพันธ์ระหว่างคำกับเอกสาร (Term-document matrix) ได้จากการสำรวจการปรากฏของคำในเอกสารของแต่ละหมวด แล้วบันทึกค่าของสมาชิกในเมทริกซ์ (t, d) โดยค่าของสมาชิกในเมทริกซ์ $(t, d) = 1$ เมื่อพบคำ (t) ในเอกสาร (d) และค่าของสมาชิกในเมทริกซ์ $(t, d) = 0$ เมื่อไม่พบคำ (t) ในเอกสาร (d) เช่น (แคลอรี, เศรษฐกิจ) = 0 เมื่อไม่พบคำว่า “แคลอรี” ในเอกสารเศรษฐกิจ แต่ (ร่างรัฐธรรมนูญ, การเมือง) = 1 แสดงว่าพบคำว่า “ร่างรัฐธรรมนูญ” ในเอกสารการเมือง ดังนั้นเมื่อดูจากตารางความสัมพันธ์ระหว่างคำกับเอกสารเท่านั้นก็จะทำให้เราสามารถที่จะค้นคืนเอกสารที่เกี่ยวข้องกับคำที่เราสนใจได้

การคำนวณในการค้นคืนแบบบูลีน (Boolean retrieval) เวกเตอร์ของคำจะถูกสร้างขึ้นจากแถวของตาราง และเมื่อต้องการรู้ผลของการค้นคืนว่ามีเอกสารในหมวดไหนที่เขียนถึงแผ่นดินไหว ร่างรัฐธรรมนูญ และ เลสเตอร์ซิตี บ้าง ก็สามารถที่จะคำนวณได้จาก แผ่นดินไหว AND ร่างรัฐธรรมนูญ AND เลสเตอร์ซิตี

$$0011110 \text{ AND } 0011110 \text{ AND } 0000111 = 0000110$$

คำตอบคือ การเมือง และ สังคม เนื่องจากค่าของบิตที่ 5 และบิตที่ 6 เป็น 1 ซึ่งเป็นผลลัพธ์ของการคำนวณของเวกเตอร์ระดับบิตด้วยตัวดำเนินการ AND