บทที่ 4
การจัดการข้อความ

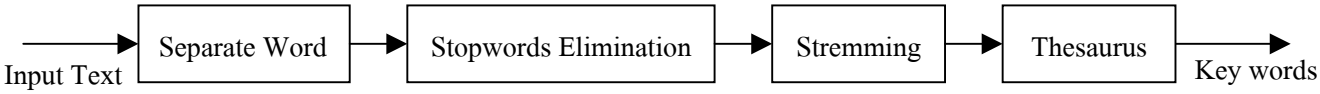
การจัดการข้อความมีความสำคัญอย่างมากในการสร้างระบบค้นหาข้อมูล เนื่องจากเป็นการกำจัดข้อมูลที่ไม่มีความสำคัญ ลดขนาดข้อมูลที่ใช้ แบ่งข้อมูลออกเป็นส่วนย่อย และวิเคราะห์ข้อมูลเหล่านั้น โดยจะมีการใช้การจัดการข้อความใน 2 ส่วนด้วยกัน คือ ส่วนของการทำindex อัตโนมัติ (Automatic Indexing) และส่วนการจัดการกับข้อความที่ผู้ใช้ป้อนเข้ามาเพื่อสืบค้น (Query)

4.1 ขั้นตอนในการจัดการข้อความ

ในส่วนของการจัดการกับข้อความที่ผู้ใช้ป้อนเข้ามาเพื่อสืบค้น มีขั้นตอนตามรูปที่ 4-1 ดังนี้

1. แบ่งคำ (Separate Word)
2. Stopwords Elimination
3. Stremming
4. Thesaurus

จากนั้นจึงนำคำที่ได้ไปใช้ในการค้นหา(Search) เพื่อเปรียบเทียบกับindexที่เก็บไว้

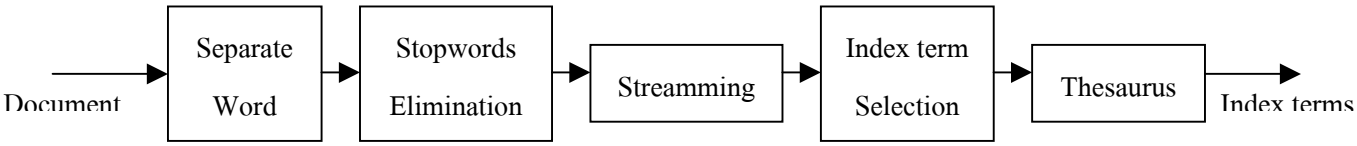


รูปที่ 4-1 แสดงขั้นตอนการจัดการข้อความที่ผู้ใช้ป้อนเข้ามาเพื่อสืบค้น

ในส่วนของการทำ index อัตโนมัติ มีขั้นตอน ดังนี้

1. แบ่งคำ (Separate Word)
2. Stopwords Elimination
3. Stremming
4. Index Term Selection
5. Thesaurus

จากนั้นจึงนำคำที่ได้ไปเก็บไว้ใช้เป็น Index Terms ของเอกสาร เพื่อใช้ในการสืบค้นข้อมูลตามรูป 4-2 ได้ดังนี้



รูปที่ 4-2 แสดงข้อความการจัดการข้อความภายในเอกสารเพื่อสร้าง index

4.1.1 การแบ่งคำ

การแบ่งคำเป็นกระบวนการในการเปลี่ยนกลุ่มของตัวอักษรให้กลายเป็นกลุ่มของคำ หรือก็คือการแบ่งข้อความที่ต่อเนื่องกันออกเป็นคำๆนั่นเอง

ในภาษาอังกฤษ ระหว่างคำจะถูกแบ่งด้วยช่องว่าง (Space) หรือเครื่องหมายต่างๆ เช่น . , - ‘’ : เป็นต้น โดยเครื่องหมายเหล่านี้จะใช้ในการแบ่งคำออกจากกัน ซึ่งในบางกรณีนั้น เครื่องหมายเหล่านี้ก็มีความสำคัญต่อความหมายของคำ เมื่อตัดทิ้งไปจะไม่สามารถเข้าใจความหมายได้ เช่น Unix-like OS แต่ในบางกรณีก็ไม่มีผลมากนัก เช่น State-of-the-art ซึ่งเมื่อแยกออกเป็นคำๆแล้ว ก็ยังคงความหมายเดิมอยู่นอกจากนี้คำบางคำที่เมื่ออยู่ด้วยกันจะมีความหมายพิเศษเพิ่มขึ้น เช่น Object Orient ก็จะทำให้เกิดความยุ่งยากในการแบ่งออกเป็นคำๆ

นอกจากนี้ตัวเลขก็ยังเป็นปัญหาที่สำคัญ เพราะตัวเลขอาจมีความสำคัญต่อความหมายของเอกสารหรือไม่มีก็ได้ นอกจากนี้ยังไม่มีความหมายในตัวเองหากขาดกลุ่มคำที่เกี่ยวข้อง ไม่เหมาะกับการเลือกเป็น Index term เพราะการค้นหาให้ได้ข้อมูลตรงกับความต้องการเป็นไปได้ยาก

เพื่อแก้ปัญหาต่างๆเหล่านี้ จึงมีการกำหนดรูปแบบในการแบ่งคำ เพื่อให้ได้ข้อมูลที่ตรงกับความต้องการมากที่สุด ดังนี้

1. แบ่งคำ โดยใช้เครื่องหมายต่างๆเป็นตัวแบ่ง ยกเว้น – และ . ซึ่งต้องไปพิจารณาเงื่อนไขข้อ 2
2. กรณี – จะคิดเป็นคำเดียวกันต่อเมื่อ
   - มีบางกลุ่มในกลุ่มที่เชื่อมกันด้วย – เป็น ตัวอักษรพิมพ์ใหญ่ทั้งหมด
     เช่น N-Layer, N-P-N
   - มีบางกลุ่มในกลุ่มที่เชื่อมกันด้วย – เป็นตัวเลขทั้งหมด
     เช่น B-79, M-16

   กรณี . จะคิดเป็นคำเดียวกันก็ต่อเมื่อ เป็นตัวอักษรพิมพ์ใหญ่ 1 ตัว สลับกับ . 1 จุด
   เช่น W.T.O. , W.H.O.
3. สำหรับกลุ่มของตัวเลขที่ถูกแบ่งจากข้อ 1 โดยไม่มีตัวอักษรอยู่ในกลุ่มเลยจะถูกกำจัด ส่วนที่มีตัวอักษรอยู่ในกลุ่มจะคงไว้ เช่น Direct3D, RFC2345
4. ตัวอักษรทั้งหมดจะถูกทำให้เป็นพิมพ์เล็ก (Lower Case) เพราะโดยปกติแล้ว Case ของตัวอักษรจะไม่มีความสำคัญในการสืบค้นข้อมูล

ในกรณีที่กลุ่มคำ 2 กลุ่ม มีความสัมพันธ์กัน และมีความหมายพิเศษเมื่ออยู่รวมกัน หรือ ไม่ตรงตามกรณีที่กำหนดแต่มีความสัมพันธ์กันนั้น

หากคำทั้งคู่นั้นมีจำนวนคำที่ใกล้เคียงกัน อยู่ในตำแหน่งใกล้เคียงกัน คือรวมกันอยู่เป็นคู่ๆ แล้ว หากคำใดคำหนึ่งได้รับเลือกให้เป็น Index Term อีกคำหนึ่งก็ควรได้รับเลือกด้วยเช่นเดียวกัน และมีโอกาสน้อยมากที่คำทั้งคู่นี้ที่อยู่ในเอกสารอื่น ซึ่งไม่มีความสัมพันธ์กันระหว่างคำทั้งสอง จะได้รับเลือกให้เป็น Index Term พร้อมกัน

ในกรณีการสืบค้นก็เช่นเดียวกัน หากคำทั้งคู่ถูกใช้ในการสืบค้นพร้อมกัน เอกสารที่มี Index term ทั้งคู่อยู่ย่อมจะอยู่ในตำแหน่งที่ดีกว่าเอกสารที่มีเพียงคำใดคำหนึ่งเป็น Index term ดังนั้นทำให้เราสามารถละเว้นการพิจารณาความสัมพันธ์ระหว่างคำเหล่านี้ไปได้

ตัวอย่างการแบ่งคำ เช่น

I have 3 documents. → i | have | documents

จะเห็นได้ว่า 3 และ . จะหายไป และ I เปลี่ยนเป็น i

หรือ How are you ? จะแบ่งได้เป็น how are you

? หายไป และ H เปลี่ยนเป็น h

X = 8*(9+3/2) จะแบ่งได้เป็น x

ตัวเลข (8932) และสัญลักษณ์ (*+/) หายไปทั้งหมด

4.1.2 Stopwords Elimination

Stopwords คือ คำที่ไม่มีความหมาย ไม่มีความสำคัญ ทำหน้าที่เชื่อมระหว่างคำต่างๆ พบได้ทั่วไปในเอกสารหรือข้อความ ไม่เกี่ยวข้องและไม่สามารถสื่อถึงสาระสำคัญของเอกสารได้ จึงไม่เหมาะที่จะใช้เป็นตัวแทนของเอกสารและไม่เหมาะที่จะนำมาทำการวิเคราะห์ สืบค้น ตัวอย่างเช่น

- Articles - a, an, the
- Pronouns - he, she, it, they, I, you
- Verbs - is, am, are, can, should
- Conjunctions - and, but, or
- Adjectives - all, still
- Adverbs - already, about, also, along
- Noun - example

ส่วนใหญ่นั้น Stopwords มักเป็นคำที่ถูกใช้บ่อยๆและพบทั่วไปในประโยคภาษาอังกฤษ ซึ่งโดยเฉลี่ยแล้วจะพบมากถึง 40 - 50 % ของคำทั้งหมดในเอกสารเลยทีเดียว ดังนั้นการตัด Stopwords ทิ้งไปจะช่วยลดระยะเวลาในการทำงาน และช่วยเพิ่มความถูกต้องด้วย

ลักษณะที่สำคัญของ Stopwords คือ มักจะมีความถี่ในการพบสูงมากๆ ซึ่งสูงกว่าความถี่ในการพบคำสำคัญทั่วไป ส่วนใหญ่แล้วจะเป็นพวก Articles, Conjunctions และ Prepositions แต่ที่เป็นคำประเภทอื่นๆก็มีเช่นเดียวกัน

ส่วนคำที่มีลักษณะตรงข้ามกับ Stopwords คือ เป็นคำสำคัญของเอกสาร ส่วนใหญ่มักเป็นคำนาม (Noun) และคำขยายนาม (Adjective) และมักมีความถี่ในการพบไม่มากหรือน้อยจนเกินไป อาจดูได้จากส่วนของเอกสารที่พบคำนั้นๆด้วย

ในการกำจัด Stopwords นั้น การใช้ Stopwords List เป็นวิธีที่ง่ายและค่อนข้างได้ผล เพราะส่วนใหญ่คำเหล่านี้มักเหมือนกันในทุกเอกสาร และเป็นกลุ่มของคำที่มีจำนวนไม่มากนัก โดยเราจะตรวจสอบคำทุกคำที่ได้จากกระบวนการตัดคำกับกลุ่มของ Stopwords ที่รวบรวมไว้ในฐานข้อมูล และกำจัดคำที่เป็น Stopwords ทิ้งไป ทั้งนี้อาจต้องระวังในการเลือกคำที่นำมาใส่ใน Stopwords Listให้เป็นคำที่ไม่มีความสำคัญต่อเอกสารจริงๆ เพราะหากผิดพลาดแล้ว อาจทำให้คำสำคัญของเอกสารนั้นๆหายไป ซึ่งส่งผลกระทบรุนแรงทั้งต่อ Index terms ที่ได้และการสืบค้นข้อมูล

ตัวอย่างในการกำจัด Stopwords เช่น

เนื่องจาก i และ have อยู่ใน Stopwords List จึงกำจัดคำเหล่านี้ทิ้ง

หรือ this, is, a, test, documents — this, is, a จะถูกกำจัดออกไป

เหลือเพียง test, documents

there, are, the, ant, here — there, are, the, here จะถูกกำจัดออกไป

เหลือเพียง ant

4.1.3 Stremming

คำว่า Strem หมายถึง ส่วนของคำที่เหลืออยู่หลังจากนำ Affixes (Prefixes และ Suffixes) ออกไป

ดังนั้น การทำ Stremming ก็คือ การทำให้คำศัพท์ต่างๆที่ถูกแปลงรูปไป เพื่อเปลี่ยนหน้าที่ของคำและนำไปใช้ในประโยค แต่ยังคงความหมายเดิม กลับมาอยู่ในรูปของรากศัพท์ (Root Word) เดิม ทั้งนี้เพื่อให้การสืบค้นและวิเคราะห์ข้อมูลมีประสิทธิภาพ และลดปริมาณการจัดเก็บข้อมูลที่ซ้ำซ้อนอยู่ในรูปของคำที่มีความหมายเดียวกัน แต่มีหลายรูปเหล่านี้

ตัวอย่างของคำเหล่านี้ เช่น

connect, connected, connecting, connection, connections, connector

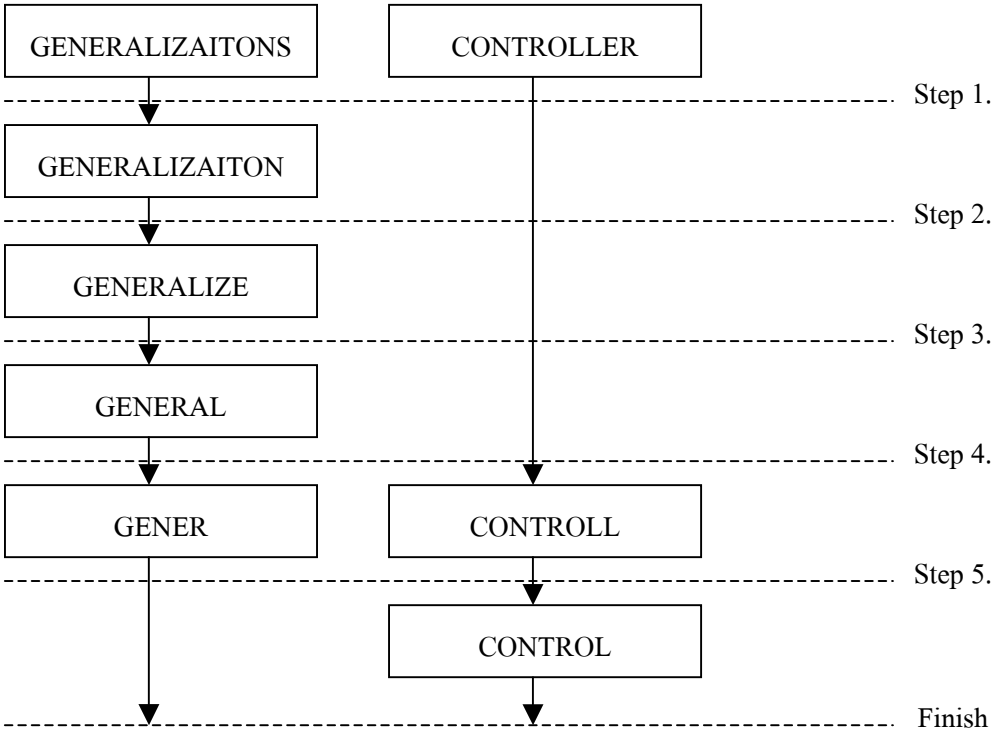
protect, protected, protecting, protection, protections, protector

เป็นต้น

ในการเปลี่ยนคำเหล่านี้กลับสู่รากศัพท์นั้น จำเป็นที่จะต้องใช้ความรู้ด้านไวยากรณ์ในภาษาอังกฤษ ทั้งในด้านการเปลี่ยนรูปของคำและโครงสร้างของประโยค ทั้งนี้กระบวนการพื้นฐานอยู่ที่การเปลี่ยนหรือลบ Prefix และ Suffix ต่างๆออกจากคำ แต่ส่วนที่สำคัญจริงๆแล้วอยู่ที่ Suffix เพราะส่วนใหญ่แล้วการเพิ่ม Prefix ให้กับคำ มักทำให้คำเปลี่ยนความหมายไป

ปัจจุบันอัลกอริทึม (Algorithm) ในการกำจัด Suffix ที่เป็นที่ยอมรับ คือ Porter's algorithm ซึ่งอาศัยการหา Suffix ที่ยาวที่สุดที่ตรงกับที่ได้กำหนดไว้ โดยใช้หลักไวยากรณ์ภาษาอังกฤษ

Porter's Algorithm แบ่งออกเป็น 5 ขั้นตอน ซึ่งคำแต่ละคำต้องผ่านขั้นตอนเหล่านี้ทั้งหมด ซึ่งคำจะเปลี่ยนแปลงไปเรื่อยๆ เมื่อผ่านแต่ละขั้นตอน เช่นดังรูปที่ 4-3 ซึ่งเป็นการยกตัวอย่างได้ดังนี้



รูปที่ 4-3 แสดงการเปลี่ยนแปลงของคำ ในขั้นตอนต่างๆของ Porter's Algorithm

จะเห็นได้ว่า คำเปลี่ยนรูปไปเรื่อยๆตามขั้นตอน จากผลการทดลองกับคำ 10000 คำ ได้ผลลัพธ์ ดังตาราง

Step	Number of reduced words
1	3597
2	766
3	327
4	2424
5	1373
Total	6350

ตาราง 4-1 แสดงจำนวนของคำที่เปลี่ยนแปลงในขั้นตอนต่างๆของ Porter's Algorithm (ต่อ 10000 คำ)

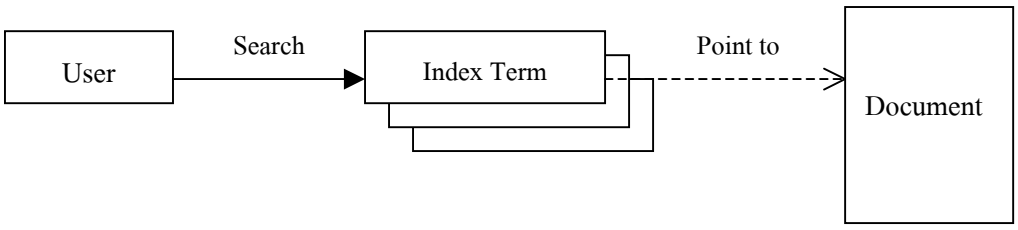
เนื่องจากคำที่ได้จากอัลกอริทึมนี้ อาจไม่ถูกต้องตรงตามรากศัพท์เดิม เพราะอัลกอริทึมมีหน้าที่เพียงทำให้คำที่มาจากรากศัพท์เดียวกันมีรูปเหมือนกัน โดยไม่รับประกันว่าคำที่ได้จะมีความหมายหรือไม่ ดังนั้นผลลัพธ์จึงอาจเป็นคำที่ไม่มีความหมาย เช่น happy -> happi, airline -> airlin, revive -> reviv เป็นต้น ดังนั้นจึงมีความจำเป็นจะต้องเก็บคำอ้างอิงไว้ เพื่อจะสามารถแสดงผลให้ผู้ใช้ได้ในภายหลัง โดยเลือกเก็บคำใดคำหนึ่งจากกลุ่มคำที่มาจากรากศัพท์เดียวกัน

นอกจากการใช้อัลกอริทึมในการตัด Suffix ทิ้งแล้ว ในบางกรณีที่คำมีการเปลี่ยนไปโดยสิ้นเชิง หรือเป็นกรณีที่เป็นข้อยกเว้น จำเป็นจะต้องมีฐานข้อมูลเกี่ยวกับคำเหล่านี้และรากศัพท์เดิมเอาไว้ เพื่อใช้ในการแปลงกลับมาเป็นรากศัพท์เดิมด้วย

4.1.4 การเลือก index (Index Terms Selection)

การเลือก Index terms เป็นขั้นตอนที่สำคัญมากในการจัดการเอกสาร เพราะ Index terms ถือเป็นตัวแทนของเอกสาร ในการสืบค้น เราจะไม่สืบค้นที่ตัวเอกสารโดยตรง แต่จะสืบค้นผ่าน Index terms ดังนั้นผลลัพธ์ของการสืบค้นจะถูกต้องหรือไม่ ส่วนใหญ่ขึ้นอยู่กับ Index terms ที่เลือกใช้

คำหรือกลุ่มคำที่ถูกเลือกมาเป็น Index terms จะต้องเป็นคำที่มีความสัมพันธ์กับเนื้อหาภายในเอกสาร เป็นส่วนที่เกี่ยวข้องกับใจความของเอกสาร และมีลักษณะเฉพาะเอกสาร ไม่ใช่คำที่พบได้ทั่วไปในเอกสารต่างๆ และเมื่อพิจารณา Index terms ทั้งหมดแล้วต้องเป็นตัวแทนของเอกสารทั้งฉบับได้โดยสมบูรณ์ดังแสดงไว้รูปที่ 4-4



รูปที่ 4-4 แสดงการค้นหาเอกสารผ่าน Index terms

จำนวนของ Index terms ก็มีส่วนสำคัญ หากเลือก Index terms น้อยเกินไปก็จะไม่สามารถเป็นตัวแทนของเอกสารทั้งฉบับได้ หากเลือกมากเกินไปก็จะทำให้การค้นหาขาดความแม่นยำ ทั้งนี้จำนวนของ Index term ในเอกสารแต่ละฉบับอาจมีไม่เท่ากันได้ ทั้งนี้เนื่องจากความยาวและเนื้อหาภายในเอกสารที่แตกต่างกัน

การเลือก Index terms นั้น แบ่งออกเป็น 2 วิธี คือ

1. <u>การเลือกโดยมนุษย์</u> วิธีนี้ใช้คนเป็นผู้อ่านเอกสาร แล้วเลือกคำสำคัญออกมาเป็น Index terms เป็นวิธีที่มีความถูกต้องแม่นยำมาก หากผู้เลือกมีความชำนาญ แต่เป็นวิธีที่ต้องใช้เวลานาน และต้องการ

บุคลากรที่มีความสามารถ และมีความรู้เกี่ยวกับเนื้อหาภายในเอกสารเป็นอย่างดี ซึ่งปกติจะเป็นหน้าที่ของเจ้าของเอกสารหรือบรรณารักษ์

เนื่องจากเป็นกระบวนการที่ใช้คน เอกสารเดียวกันจึงอาจมี Index terms แตกต่างกันได้ ขึ้นกับผู้เลือก นอกจากนี้ยังอาจมีโอกาสเกิด Index terms ไม่ครบถ้วนได้ง่าย

2. <u>การเลือกโดยอัตโนมัติ</u> วิธีนี้ใช้โปรแกรมเป็นผู้พิจารณาเลือก Index terms โดยส่วนใหญ่พิจารณาจากความถี่และตำแหน่งของคำที่พบในเอกสาร ซึ่งต้องมีขั้นตอนในการปรับค่า และการคำนวณต่างๆ เพื่อให้มีความถูกต้องเพิ่มขึ้น

วิธีการนี้มีโอกาสผิดพลาดได้มากกว่าการเลือกโดยมนุษย์ ซึ่งส่วนใหญ่มักเกิดจากการเลือก Index terms ผิด จึงมักใช้คนตรวจสอบผลลัพธ์ที่ได้อีกครั้ง

สำหรับรายละเอียดของวิธีการเลือก Index terms อัตโนมัติจะกล่าวถึงในบทต่อไป

4.1.5 Thesauri

Thesaurus คือ กลุ่มคำที่มีความหมายใกล้เคียงกันหรือเหมือนกัน กล่าวคือ สิ่งหนึ่งๆอาจถูกเรียกได้ในหลายรูปแบบขึ้นกับผู้เรียก เช่น เราอาจเรียกจักรยานว่า bicycle, cycle หรือ bike ก็มีความหมายถึงจักรยาน 2 ล้อเช่นเดียวกัน

Thesaurus อาจเป็นได้ทั้งคำเหมือน (Synonyms) กลุ่มคำ (Phrases) หรือคำที่เกี่ยวข้องในหัวข้อนั้นๆ จุดประสงค์ในการสร้าง Thesauri List ก็คือ

1. การพยายามหาคำที่เป็นตัวแทนของกลุ่ม เพื่ออ้างอิงทั้งในการสร้าง Index term และการค้นหาข้อมูล
2. ช่วยให้ผู้ใช้สามารถค้นหาข้อมูลได้ตรงกับความต้องการมากขึ้น
3. อาจช่วยผู้ใช้ในการแบ่งระดับความสัมพันธ์ (Hierarchical relationship) ของคำที่ค้นหา ว่าต้องการให้ค้นหาในความหมายที่กว้างมากแค่ไหน

ตัวอย่างในการใช้ Thesaurus เพื่อสร้างคำอ้างอิง เพื่อเป็นตัวแทนของกลุ่ม เช่น

coward	USE FOR	craven
craven	USE	coward

จะเห็นได้ว่า เราใช้คำว่า coward ในการอ้างอิงแทนคำว่า craven ซึ่งในการสร้าง Index term เราจะใช้ coward แทน craven และเมื่อผู้ใช้ค้นหาคำว่า craven เราจะเปลี่ยนเป็นค้นหาคำว่า coward แทน โดยที่ความสัมพันธ์ระหว่างคำในกรณีนี้ต้องเป็นความสัมพันธ์ที่สามารถใช้ทดแทนกันได้ โดยสมบูรณ์

นอกจากนี้ในกรณีที่คำสามารถใช้แทนกันได้โดยสมบูรณ์นี้ เราอาจนำ Thesaurus มาใช้เพียงด้านใดด้านหนึ่ง เช่น ใช้เฉพาะส่วน Query หรือ ใช้ เฉพาะส่วนของการทำ Index terms ก็ได้ ซึ่งจะให้ผลลัพธ์เหมือนกัน แต่กรณีที่ใช้ส่วน Query จะทำให้การสืบค้นช้าลง ส่วนกรณีที่ใช้ส่วนการทำ Index terms ก็จะทำให้ใช้เนื้อที่ในการเก็บข้อมูลมากขึ้น

นอกจาก Synonyms แล้ว เรายังสามารถใช้คำที่เป็น Near-synonyms มาใช้เป็น Thesaurus ได้ โดยการแบ่งระดับความสัมพันธ์ออกเป็นลำดับชั้น (Hierarchical relationship) ซึ่งส่วนใหญ่จะแบ่งออกเป็น ความหมายเชิงกว้าง (Broader Term : BT) และความหมายเชิงแคบ (Narrower Term : NT) ซึ่งผู้ใช้จะเป็นผู้กำหนดว่าต้องการให้สืบค้นแบบใด โดยในกรณีที่ไม่มีการแบ่งระดับความสัมพันธ์ เราจะเรียกรวมว่าเป็น Related Term (RT)

$$RT = NT \cup BT$$

การแบ่งความสัมพันธ์ออกเป็นระดับชั้นนั้น อาจทำให้ผู้ใช้เกิดความสับสนได้ง่าย ต้องใช้เนื้อที่จัดเก็บข้อมูลมาก ทำให้เกิดความยุ่งยากและเพิ่มเวลาที่ใช้ในการสืบค้นด้วย ดังนั้น จึงมักนิยมใช้ Thesaurus แบบความหมายเหมือนกันมากกว่า

ข้อควรระวังในการสร้าง Thesauri ก็คือ ความสัมพันธ์ระหว่างคำ ในบางกรณีจะมีลักษณะเป็นความสัมพันธ์แบบทางเดียว ($A \in B$ แต่ $B \notin A$) เช่น tunnel = tube แต่ tube $\neq$ tunnel เป็นต้น และนอกจากนี้ คำศัพท์หนึ่งคำอาจมีได้หลายความหมาย ซึ่งในแต่ละความหมายอาจมีกลุ่มของคำเหมือนที่แตกต่างกันไป