บทที่ 2
การวิเคราะห์ข้อความ

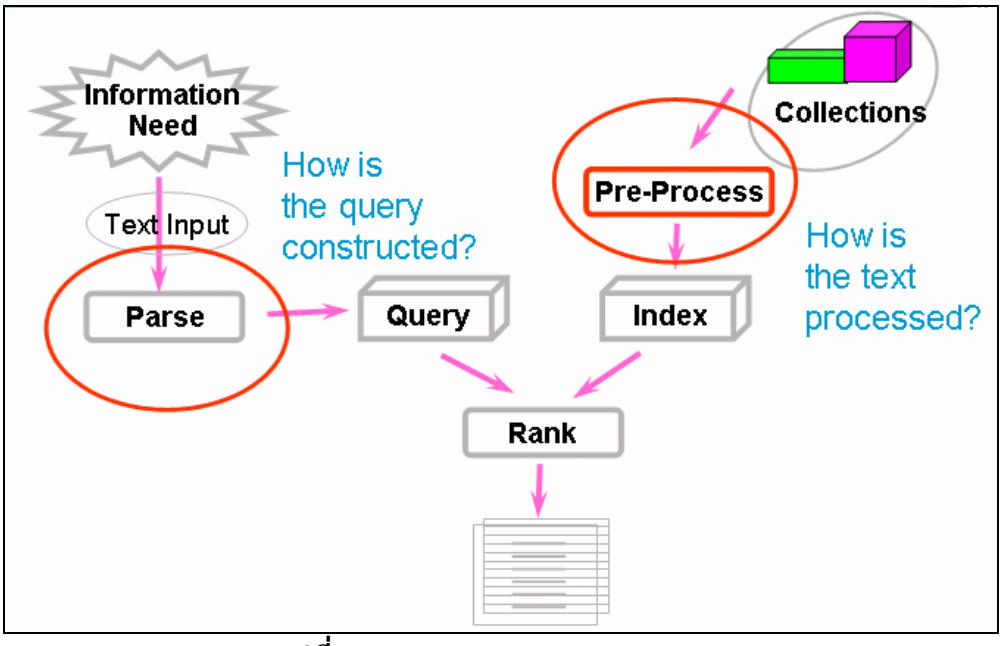
วัตถุประสงค์

1. เพื่อให้นักศึกษาทราบทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์ข้อความ
2. เพื่อให้นักศึกษาทราบถึงวิธีการสร้างตัวแทนเอกสาร
3. เพื่อให้นักศึกษาทราบถึงประเภทของดรรชนีที่ใช้ในระบบค้นคืนสารสนเทศ
4. เพื่อให้นักศึกษาทราบถึงปัญหาในการวิเคราะห์ข้อความ
5. เพื่อให้นักศึกษาทราบถึงวิธีการแก้ปัญหาในการวิเคราะห์ข้อความ

สารบัญ

2.1 บทนำ

การเก็บข้อความทั้งหมดในเอกสารนั้นใช้เนื้อที่ในหน่วยความจำและค่าใช้จ่ายที่สูงมาก ดังนั้นในการแก้ไขปัญหานี้จึงมีการจัดเก็บข้อความเอกสารต่างๆเหล่านั้นในรูปของตัวแทนเอกสาร ดังรูปที่ 2.1 :แสดง Content Analysis Area ซึ่งวิธีการนั้นจะนำข้อความซึ่งอาจเป็นชื่อหนังสือ บทคัดย่อ สารบัญ บรรณานุกรม หรืออาจเป็นข้อความทั้งหมดของเอกสาร นำมาวิเคราะห์เพื่อหาตัวแทนของเอกสารเหล่านั้นให้เหมาะสม



รูปที่ 2.1 : แสดง Content Analysis Area

วิธีการวิเคราะห์ข้อความให้อยู่ในรูปแบบของตัวแทนเอกสาร แบ่งออกได้เป็น 2 วิธีการด้วยกันคือ

1. วิเคราะห์ทางด้านภาษาศาสตร์ วิธีการนี้นั้นมีความยุ่งยากและ เสียค่าใช้จ่ายสูงมาก
2. วิเคราะห์ทางด้านสถิติศาสตร์ เป็นวิธีการที่นิยมซึ่งวิธีการนี้ได้ถูกตรวจสอบและทดลองเพื่อให้สามารถนำมาประยุกต์ใช้กับระบบค้นคืนสารสนเทศ(IR) ได้ดี โดยหลักการพื้นฐานนั้นใช้ทฤษฎีของลุนซ์ (Luhn) ในการสร้างตัวแทนเอกสาร ซึ่งมีการขจัดคำที่มีความถี่ของการเกิดขึ้นสูงออกไป ดังรูปที่ 2.2 นอกจากนี้ยังทำการขจัดคำที่ฟุ่มเฟือยเป็นคำหรือเทอมที่ไม่มีความหมาย รวมทั้งขจัดคำที่มีความหมายคล้ายคลึงกันหรือกลุ่มคำที่มีความหมายเดียวกันออกไปทำให้เหลือเฉพาะคำสำคัญที่สามารถสร้างเป็นตัวแทนเอกสารได้ นำคำสำคัญต่างๆเหล่านี้มาสร้างดรรชนี สำหรับกระบวนการค้นคืนสารสนเทศต่อไป

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

รูปที่ 2.2 : แสดงถึงคำที่มีความถี่ของการเกิดขึ้นสูงในเอกสาร

2.2 ทฤษฎีของลุนซ์ (Luhn)

ลุนซ์ ได้ให้ความเห็นในการวิเคราะห์ข้อความไว้ตั้งแต่ปี คศ. 1958 ว่า

“จำนวนครั้งของ**คำ**ๆหนึ่งเกิดขึ้นภายใน**เอกสารหนึ่ง** จะใช้เป็นวิธีวัดความสำคัญของคำได้วิธีหนึ่ง ความเกี่ยวพันธ์ของตำแหน่งภายในประโยคของคำจะให้ค่าของความสำคัญที่เป็นประโยชน์ในการวัดเพื่อกำหนดความสำคัญของประโยค ปัจจัยที่สำคัญของประโยคจะอยู่บนพื้นฐานของการรวมกันของการวัดทั้งสองนี้ “

'It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements.'

หรืออาจกล่าวได้ว่า ความถี่ของข้อมูล (Frequency Data) สามารถถูกใช้ เพื่อนำมาใช้วัดความสำคัญของคำและประโยคที่ใช้แทนเอกสารหนึ่งได้
คำและประโยคที่ใช้แทนเอกสารหนึ่งได้

สมมุติว่า

f เป็นความถี่ (frequency) ของการเกิดขึ้นของคำ ใน**ตำแหน่ง** ต่างๆของข้อความ

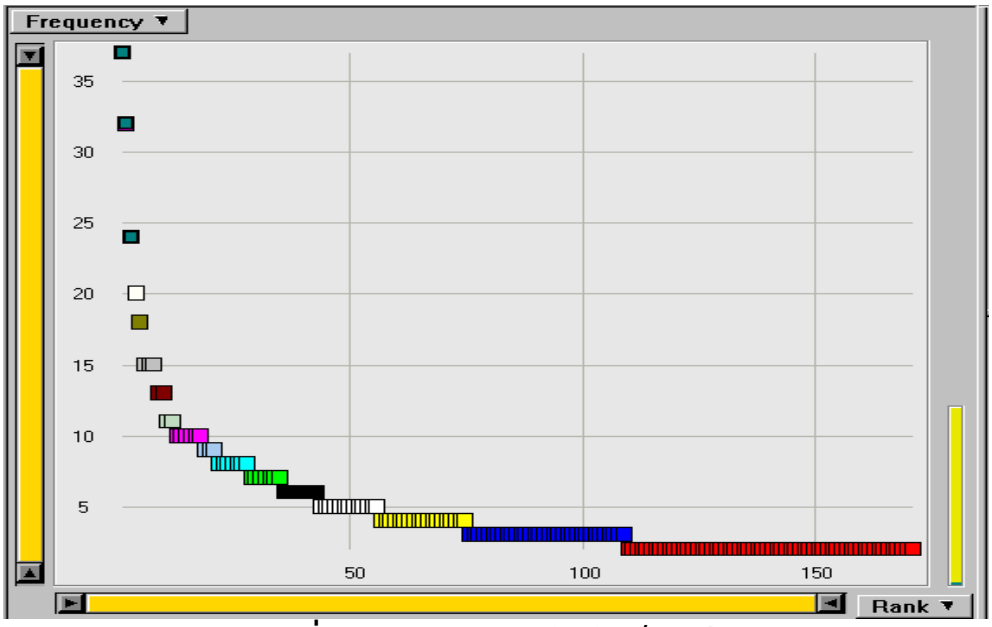
r เป็น rank order หรือ**ลำดับของระดับ** ของการ**เกิดขึ้นของคำ ๆ**นั้นในเอกสาร

ความสัมพันธ์ของ f และ r ทำให้เกิด Hyperbolic Curve กล่าวคือ ค่าของ f สูงจะส่งผลให้ r มีค่าต่ำ

ตัวอย่าง 2.1 แสดงความถี่ของการเกิดขึ้นของคำและrank order ของคำๆนั้น ดังนี้

Rank(r)	***Freq(f)***	***Term***
1	37	system
2	32	knowledge
3	24	base
4	20	problem
5	18	abstract
6	15	model
7	15	language
8	15	implement
9	13	reason
10	13	information
11	11	expert
12	11	analysis
13	10	rule
14	10	program
15	10	operation
16	10	evaluation
17	10	compute
18	10	case
19	9	generation
20	9	form

นำความสัมพันธ์ระหว่าง f และ r มาแสดงเป็นกราฟได้ ดังรูปแสดงที่ 2.3



รูปที่ 2.3 : แสดง ความสัมพันธ์ระหว่าง f และ r

กฎของ Zipf กล่าวว่า

“การคูณค่าของความถี่ของคำ(f) กับ Rank Order(r) มีค่าค่อนข้างคงที่”

f ∝ 1/r

ซึ่ง Rank(r) คือตัวเลขตำแหน่งของคำในรายการที่เรียงลำดับที่มีการลดความถี่(f)

$$f \propto \frac{1}{r} \qquad f \cdot r = k \ \ (\text{for constant } k)$$

ถ้าความน่าจะเป็นของคำของตำแหน่ง r คือ p และ N คือจำนวนทั้งหมดของคำที่เกิดขึ้น

$$p_r = \frac{f}{N} = \frac{A}{r} \quad \text{for corpus indp. const. } A \approx 0.1$$

สำหรับคำที่ปรากฏเป็นจำนวน n เที่ยวนั้นมี rank ดังนี้

Rn = AN/n

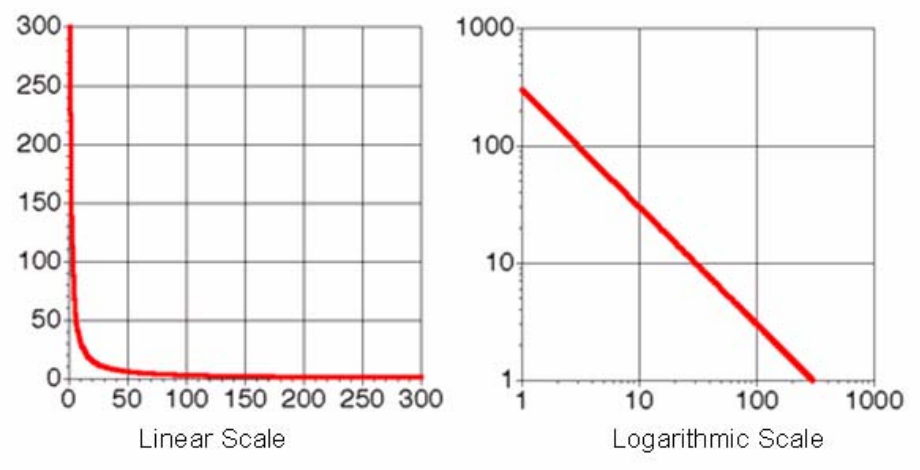
จำนวนคำ(I) ที่ปรากฏ n เที่ยว

$$I_n = r_n - r_{n+1} = \frac{AN}{n} - \frac{AN}{n+1} = \frac{AN}{n(n+1)}$$

สมมุติตำแหน่งสุงสุดของเหตุการณ์ที่มี rank D = AN/1 เศษส่วนของคำที่มีความถี่ n คือ

$$\frac{I_n}{D} = \frac{1}{n(n+1)}$$

ซึ่งเศษส่วนของคำจะปรากฏครึ่งหนึ่งเท่านั้น ดังรูปที่ 2.4 :แสดง Zipf Distribution



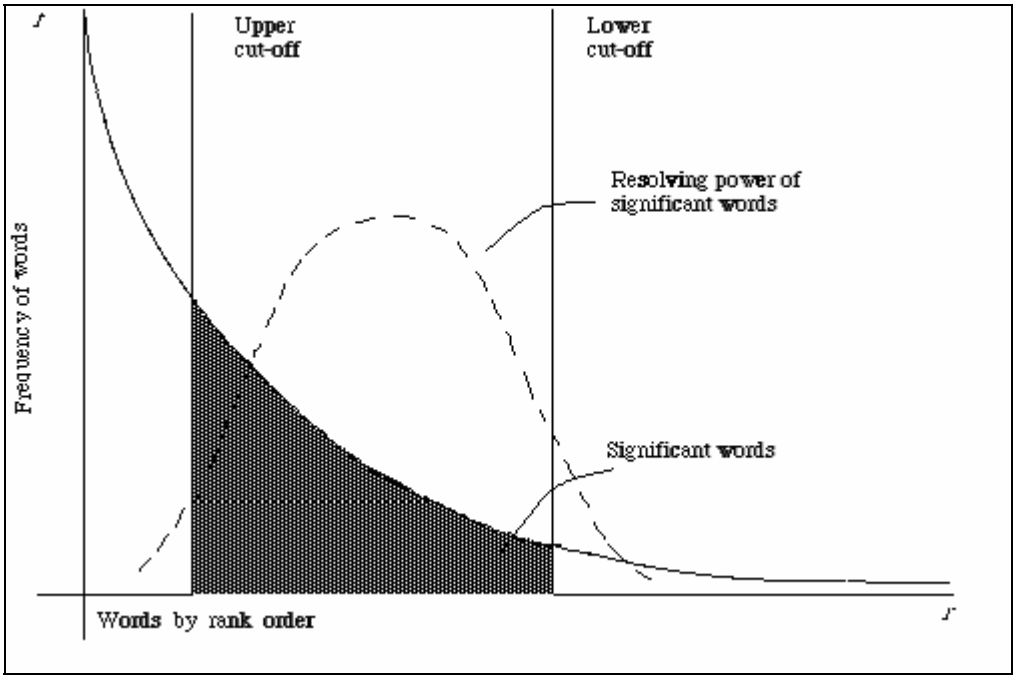
รูปที่ 2.4 : แสดง Zipf Distribution

Zipt's Law มีผลกระทบกับระบบค้นคืนสารสนเทศ **ข้อดี**คือ กำจัดคำหรือข้อความที่ไม่สำคัญ ทำให้ลดค่าใช้จ่ายและเนื้อที่ในหน่วยความจำ **ข้อเสีย**คือ คำส่วนมากจะถูกรวบรวมให้เพียงพอสำหรับการวิเคราะห์ทางสถิติ เพื่อใช้สำหรับการสอบถาม(query) ซึ่งเป็นการยากสำหรับที่จะค้นหาคำต่างๆเหล่านั้น

ลุนซ์ ได้ใช้กฎของ Zipt's สร้างจุด Cut-off ขึ้นมา 2 จุดคือ Upper Cut-off และ Lower Cut-off เพื่อขจัดคำที่ไม่สำคัญในเอกสารออกไป เพื่อให้เหลือข้อความในเอกสารให้น้อยลง การขจัดคำในส่วนนี้ได้แก่ คำที่อยู่เลย Upper Cut-off ถือว่าเป็นคำที่ไม่สำคัญ เป็นคำที่เกิดขึ้นบ่อยๆ คำเหล่านี้ไม่ได้แสดงเนื้อหาของเอกสาร เช่น an a but and in เป็นต้น นอกจากนี้ยังขจัดคำที่อยู่ใต้ Lower Cut-off คำในกลุ่มนี้เป็นคำที่ค่อนข้างจะเกิดขึ้นน้อยครั้งในเอกสาร คำเหล่านี้ไม่ให้ความสำคัญแก่เนื้อเรื่องของเอกสาร ดังแสดงตามรูปที่ 2.5 ต่อจากนั้นลุนซ์ได้เสนอวิธีวัดหาความสำคัญ ดังนี้

Resolving Power ของคำที่สำคัญ หรือ ความสามารถของคำที่จะแยกแยะเนื้อหา ไว้ว่า

" จะขึ้นถึงจุดสุดยอดในตำแหน่งประมาณกึ่งกลางทางของ Rank Order ระหว่าง Cut-off 2 จุด และจะลดลงเกือบใกล้ถึงศูนย์ เมื่ออยู่ใกล้ๆตอนปลายหรือที่จุด Cut-off "



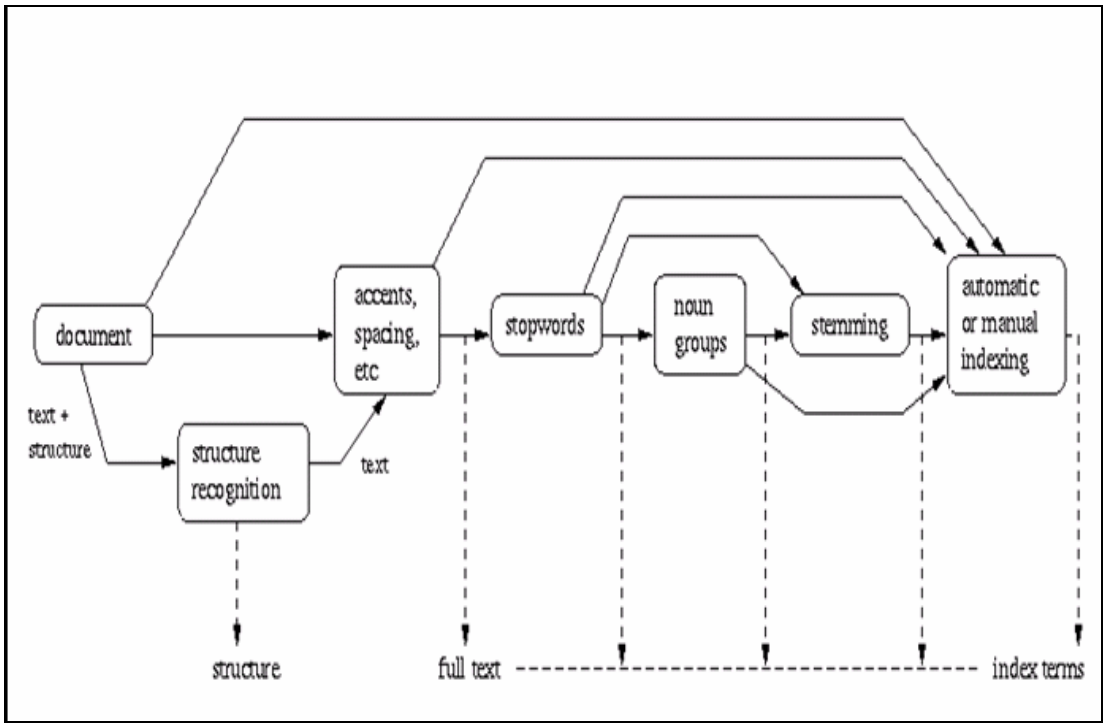
รูปที่ 2.5 :แสดงความสัมพันธ์ระหว่างความถี่ของการเกิดขึ้นของคำ(f) และRank Order(r)

ปัญหาที่เกิดขึ้นคือการกำหนดจุด Cut-off 2 จุดนี้ให้เหมาะสม ทำอย่างไรเพราะไม่ทราบว่าจุดใดที่เหมาะสมที่สุด การแก้ปัญหานี้อาจกระทำได้จากการสังเกตและการทดลองโดยสมมุติจุด Cut-off ทั้งสองจุดเพื่อดูผลลัพธ์ที่เกิดขึ้นและทำการปรับเปลี่ยนค่าจุด Cut-off โดยกระทำซ้ำๆ เพื่อหาค่าCut-off ทั้งสองจุดที่เหมาะสมที่สุด

นอกจากนี้ ลุนซ์ได้เสนอวิธีการสร้างบทคัดย่อโดยอัตโนมัติ(Automatic Abstract) โดยวัดความสำคัญของประโยคในเชิงปริมาณเป็นตัวเลข โดยพิจารณาถึงจำนวนคำที่สำคัญและไม่สำคัญในแต่ละส่วนของประโยค ซึ่งจากการวัดความสำคัญนี้ประโยคที่มีลำดับความสำคัญในลำดับสูงสุดจะถูกเก็บไว้ในบทคัดย่อโดยอัตโนมัติดังกล่าว

2.3 การสร้างตัวแทนเอกสาร-Conflation

การสร้างตัวแทนเอกสารนั้นเป็นการเลือกคำหรือประโยคที่มีความสำคัญของเอกสารนั้น เพื่อแทนข้อมูลทั้งหมดของเอกสาร ซึ่งเป็นการสร้าง คีย์เวิร์ด(keywords) หรือ ดรรชนี(index) ที่ใช้สำหรับการสืบค้นเอกสารนั่นเอง วิธีการในการวิเคราะห์ข้อความเพื่อหาตัวแทนเอกสาร จาก Metadata ที่อาจเป็น ชื่อผู้แต่ง, ชื่อเรื่อง, วันที่ , ISBN , ความยาว , สำนักพิมพ์ , บทคัดย่อ, สารบัญ , บรรณานุกรม หรือ ข้อความทั้งหมดของเอกสาร ที่เป็นหนังสือ ,วารสาร , หนังสือพิมพ์ ,งานวิจัย เราสามารถใช้วิธีที่เรียกว่า conflation algorithm ในการสร้างตัวแทนเอกสารต่างๆเหล่านี้ ดังขั้นตอนที่จะกล่าวต่อไปนี้



รูปที่ 2.6 : แสดงขั้นตอนการสร้างตัวแทนเอกสาร

ขั้นตอนในการสร้างตัวแทนเอกสารตามรูปแสดงที่ 2.5 แบ่งออกเป็น 3 ขั้นตอนใหญ่ๆได้แก่

1. **การขจัดคำที่มีความถี่ของการเกิดขึ้นสูงออกไปหรือคำหยุด (stopwords)**

ขั้นตอนนี้เป็นการปฏิบัติการของ Upper Cut-off โดยขจัดคำที่มีความถี่สูงออกไป เนื่องจากคำเหล่านี้ไม่ให้ความหมายหรือแสดงเนื้อหาของเอกสาร เราเรียกคำประเภทนี้ว่า Stop Words หรือ Common Words หรือ Fluff Words คำต่างๆเหล่านี้ได้แก่

- คำสรรพนาม สันธาน บุพบท เช่น he, and, in, or
- อักขระโรมัน A ถึง Z
- ตัวเลข 0 ถึง 9
- สัญลักษณ์ต่าง ๆ เช่น { , [, * , #

รวมประมาณ 250 คำ ช่วยลดขนาดของเนื้อที่เก็บประมาณ 30-50 %

ตารางที่ 2.1 : แสดงตัวอย่างของคำที่ไม่สำคัญซึ่งควรจะขจัดออกไป

A	CANNOT	INTO	OUR	THUS
ABOUT	CO	IS	OURS	TO
ABOVE	COULD	IT	OURSELVES	TOGETHER
ACROSS	DOWN	ITS	OUT	TOO
AFTER	DURING	ITSELF	OVER	TOWARD
AFTERWARDS	EACH	LAST	OWN	TOWARDS
AGAIN	EG	LATTER	PER	UNDER
AGAINST	EITHER	LATTERLY	PERHAPS	UNTIL
ALL	ELSE	LEAST	RATHER	UP
ALMOST	ELSEWHERE	LESS	SAME	UPON
ALONE	ENOUGH	LTD	SEEM	US
ALONG	ETC	MANY	SEEMED	VERY
ALREADY	EVEN	MAY	SEEMING	VIA
ALSO	EVER	ME	SEEMS	WAS
ALTHOUGH	EVERY	MEANWHILE	SEVERAL	WE

ALWAYS	EVERYONE	MIGHT	SHE	WELL
AMONG	EVERYTHING	MORE	SHOULD	WERE
AMONGST	EVERYWHERE	MOREOVER	SINCE	WHAT
AN	EXCEPT	MOST	SO	WHATEVER
AND	FEW	MOSTLY	SOME	WHEN
ANOTHER	FIRST	MUCH	SOMEHOW	WHENCE
ANY	FOR	MUST	SOMEONE	WHENEVER
ANYHOW	FORMER	MY	SOMETHING	WHERE
ANYONE	FORMERLY	MYSELF	SOMETIME	WHEREAFTER
ANYTHING	FROM	NAMELY	SOMETIMES	WHEREAS
ANYWHERE	FURTHER	NEITHER	SOMEWHERE	WHEREBY
ARE	HAD	NEVER	STILL	WHEREIN
AROUND	HAS	NEVERTHELESS	SUCH	WHEREUPON
AS	HAVE	NEXT	THAN	WHEREVER
AT	HE	NO	THAT	WHETHER
BE	HENCE	NOBODY	THE	WHITHER
BECAME	HER	NONE	THEIR	WHICH
BECAUSE	HERE	NOONE	THEM	WHILE
BECOME	HEREAFTER	NOR	THEMSELVES	WHO
BECOMES	HEREBY	NOT	THEN	WHOEVER
BECOMING	HEREIN	NOTHING	THENCE	WHOLE
BEEN	HEREUPON	NOW	THERE	WHOM
BEFORE	HERS	NOWHERE	THEREAFTER	WHOSE
BEFOREHAND	HERSELF	OF	THEREBY	WHY
BEHIND	HIM	OFF	THEREFORE	WILL
BEING	HIMSELF	OFTEN	THEREIN	WITH
BELOW	HIS	ON	THEREUPON	WITHIN
BESIDE	HOW	ONCE	THESE	WITHOUT

BESIDES	HOWEVER	ONE	THEY	WOULD
BETWEEN	I	ONLY	THIS	YET
BEYOND	IE	ONTO	THOSE	YOU
BOTH	IF	OR	THOUGH	YOUR
BUT	IN	OTHER	THROUGH	YOURS
BY	INC	OTHERS	THROUGHOUT	YOURSELF
CAN	INDEED	OTHERWISE	THRU	YOURSELVES

2. ดึงเอาคำท้ายออก (Suffix)

คำหลายคำที่เป็นคำที่มีความหมายคล้ายคลึงกันและจัดเป็นคำประเภทเดียวกัน เช่น Compatible กับ Compatibility ถ้าตัด –ible กับ -ibility รากศัพท์(stem) ที่เหลือคือคำเดียวกัน

กฎเกณฑ์ในการ Suffix

(1).ความยาวของรากศัพท์(stem) ที่เหลือต้องมากกว่าจำนวนที่กำหนด โดยทั่วไปคือ 2

(2).ตอนท้ายของรากศัพท์(Stem-ending) ต้องถูกต้องตามเงื่อนไขตามที่กำหนดเช่น ไม่ลงท้ายด้วย q เป็นต้น

ตัวอย่างที่ 2.2 : แสดงคำที่นำเอานำคำท้ายออกทำให้ผิดความหมาย

- organization, organ → organ
- police, policy → polic
- arm, army → arm

ตารางที่ 2.2 : แสดงตัวอย่างของคำที่มีความหมายคล้ายคลึงกัน

ABILITIES	ALISES	ANCIAL	ARISABILITY	ASISINGFUL
ABILITY	ALISING	ANCIALS	ARISABLE	ASISINGLY
ABLE	ALISINGFUL	ANCIES	ARISATION	ASISINGS
ABLED	ALISINGLY	ANCING	ARISATIONS	ASIZABLE
ABLEDLY	ALISINGS	ANCINGFUL	ARISE	ASIZE
ABLENESS	ALISM	ANCINGLY	ARISED	ASIZED
ABLENESSES	ALISMS	ANCINGS	ARISEDLY	ASIZEDLY
ABLER	ALIST	ANCY	ARISER	ASIZER
ABLES	ALISTIC	ANEOUS	ARISES	ASIZES
ABLING	ALISTICALLY	ANEOUSLY	ARISING	ASIZING
ABLINGFUL	ALISTICISM	ANEOUSNESS	ARISINGFUL	ASIZINGFUL
ABLINGLY	ALISTICISMS	ANT	ARISINGLY	ASIZINGLY
ABLY	ALISTICS	ANTANEOUS	ARISINGS	ASIZINGS
ACEOUS	ALISTS	ANTANEOUSLY	ARISM	ASM
ACEOUSLY	ALITIES	ANTED	ARISMS	ASMS
ACEOUSNESS	ALITY	ANTEDLY	ARIST	AST
ACEOUSNESSES	ALIZATION	ANTIALITIES	ARISTIC	ASTIC
ACIES	ALIZATIONAL	ANTIALITY	ARISTICISM	ASTICAL
ACIDOUS	ALIZATIONALLY	ANTIALNESS	ARISTICISMS	ASTICALLY
ACIDOUSLY	ALIZATIONS	ANTIALNESSES	ARISTICS	ASTICISM
ACIOUSNESS	ALIZE	ANTIC	ARISTS	ASTICISMS
ACIOUSNESSES	ALIZED	ANTICISM	ARITIES	ASTICS
ACITIES	ALIZEDLY	ANTICISMS	ARITY	ASTMENT
ACITY	ALIZER	ANTICS	ARIZABILITIES	ASTMENTS
ACY	ALIZES	ANTING	ARIZABILITY	ASTRIES
AE	ALIZING	ANTINGFUL	ARIZABLE	ASTRY
AGE	ALIZINGFUL	ANTINGLY	ARIZATION	ASTS
AGED	ALIZINGLY	ANTINGS	ARIZATIONS	ASY
AGEDLY	ALIZINGS	ANTLY	ARIZE	ATA
AGER	ALLED	ANTMENT	ARISED	ATABILITIES
AGES	ALLEDLY	ANTMENTS	ARISEDLY	ATABILITY
AGING	ALLIC	ANTRESS	ARIZER	ATABLE
AGINGFUL	ALLICALLY	ANTRESSES	ARIZES	ATABLES
AGINGLY	ALLICISM	ANTRY	ARIZING	ATABLY
AIC	ALLICISMS	ANTS	ARIZINGFUL	ATAL
AICAL	ALLICS	AR	ARIZINGLY	ATE
AICALLY	ALLING	ARIAL	ARIZINGS	ATED
AICALS	ALLINGFUL	ARIALS	ARLY	ATEDLY
AICISM	ALLINGLY	ARIAN	AROID	ATELY
AICISMS	ALLMENT	ARIANS	AROIDS	ATENESS
AICS	ALLY	ARIC	ARS	ATENESSES
AL	ALMENT	ARICISM	ARY	ATER
ALISATION	ALNESS	ARICISMS	ASIS	ATES
ALISATIONAL	ALNESSES	ARICS	ASISE	ATIC
ALISATIONALLY	ALS	ARIES	ASISEABLE	ATICAL
ALISATIONS	ANCE	ARILINESS	ASISED	ATICALLY
ALISE	ANCED	ARILY	ASISEDLY	ATICISM
ALISED	ANCEDLY	ARINESS	ASISER	ATICISMS
ALISEDLY	ANCER	ARINESSES	ASISES	ATICS
ALISER	ANCES	ARISABILITIES	ASISING	ATING

3. Equivalent Stem-Ending และ Noun group(กลุ่มคำ)

ปัญหา มีอยู่หลายคำที่เอาคำท้ายออกไปแล้ว มีความหมายถูกต้อง แต่บางคำไม่เป็นอย่างนั้นเช่น ABSORB- กับ ABSORPT- ซึ่งคำสองคำนี้เป็นคำที่มีความหมายเดียวกัน แต่ลงท้ายต่างกัน ซึ่ง –B กับ –PT ดังนั้นเราจึงมีการเก็บรากศัพท์ในลักษณะนี้ในรายการของ Equivalent Stem-Endings เพื่อใช้ในการตรวจสอบ พร้อมกับระบุความหมายของคำนั้นไว้ในวงเล็บต่อท้ายคำนั้นด้วย การขจัดคำต่างๆเหล่านี้ไม่ได้มีความถูกต้อง 100 % แต่มีความผิดพลาดของการสร้างตัวแทนเอกสารในลักษณะนี้อยู่ประมาณ 5 %

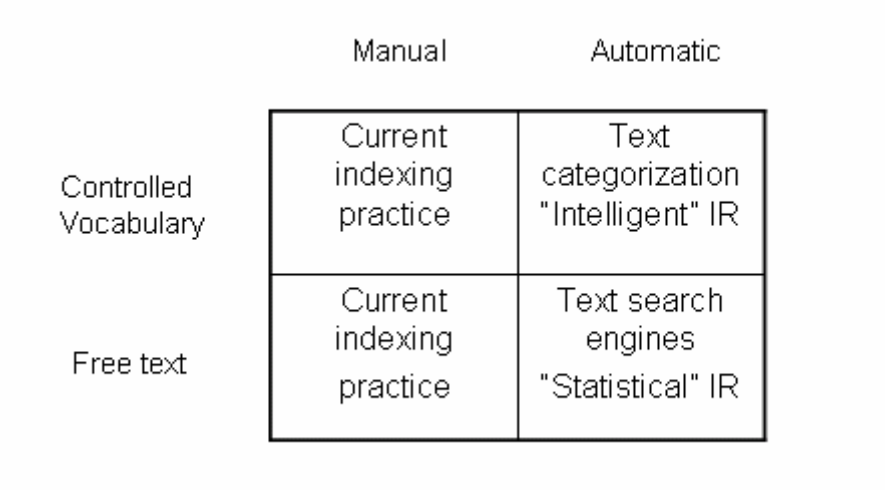
นอกจากนี้ยังมีการจัดกลุ่มของคำที่มีความหมายเดียวกัน สำหรับคำบางคำที่ตัวสะกดต่างกันแต่มีความหมายเดียวกัน เช่น SUN กับ SOLAR ส่งผลให้เกิดความผิดพลาดในการสร้างตัวแทนเอกสารที่มีประสิทธิภาพ การแก้ปัญหานี้โดยนำคำที่มีความหมายเดียวกันนี้นำมารวมกลุ่มเป็นคลาส(class) และตั้งเอาคำใดคำหนึ่งเป็นชื่อของกลุ่มแทน

ผลของการทำงานของอัลกอริทึมนี้ ได้ผลลัพธ์ เป็นคำสำคัญ(Significant Words) หรือคำที่สามารถแสดงเนื้อหาของเอกสาร (Content-Bearing Words) เป็นตัวแทนของเอกสารทั้งหมดซึ่งสามารถนำไปเก็บไว้ในระบบ ทำให้ประหยัดเนื้อที่ของหน่วยความจำได้มาก

2.4 ดรรชนี(indexing)

ภาษาดรรชนีเป็นภาษาที่ใช้บรรยายเอกสารและข้อคำถาม สมาชิกของภาษาดรรชนีคือ เทอมดรรชนี ซึ่งอาจดึงมาจากข้อความของเอกสารหรืออาจจะได้รับโดยอิสระ ภาษาดรรชนีมีการบรรยายเป็น pre-coordinated หรือ post-coordinated ก็ได้โดย pre-coordinated นั้นเป็นการกำหนดดรรชนีไว้ก่อน ซึ่งมีการกำหนดข้อตกลงทางตรรกะ เทอมดรรชนีที่ใช้เป็นส่วนของlabel ที่กำหนดคลาสของเอกสาร สำหรับ post-coordinated นั้นจะสร้างขึ้นในขณะที่ทำการค้นหา โดยมีการสืบค้นจากคลาสที่ระบุไว้แล้วจากการรวมกลุ่มของ label กับเทอมดรรชนีที่กำหนด(index term)

คำศัพท์ของภาษาดรรชนีอาจจะถูกควบคุมหรือไม่ถูกควบคุมก็ได้ ในกรณีที่มีการควบคุมนั้นอาจเป็นเทอมดรรชนีหรือข้อตกลงดรรชนี (term index) ที่เป็นความสัมพันธ์ที่เป็นลำดับชั้นที่กระทำโดยมนุษย์ (Manual or human indexing) สำหรับภาษาดรรชนีที่สร้างโดยอัตโนมัตินั้น(Automatic indexing:)โปรแกรมในการสร้างดัชนีจะตัดสินใจจากคำ วลีหรือจุดเด่นอื่นๆเพื่อใช้สำหรับทดสอบเอกสารเพื่อสร้างดรรชนีขึ้นมา ดังรูปที่ 2.7 : แสดง Manual vs. Automatic Indexing



รูปที่ 2.7 : **แสดง** Manual vs. Automatic Indexing

โครงการ Smart (1966) พัฒนาการสร้างดรรชนีอัตโนมัติ(automatic indexing) ซึ่งสามารถสรุปได้ว่า ค่าเฉลี่ยของกระบวนการสร้างดรรชนีโดยการกำหนดเอกสารที่ให้หรือการสอบถามเป็นเซตของเทอมหรือข้อตกลงที่มีน้ำหนัก(weight)และไม่มีน้ำหนักที่บรรจุเอกสารหรือข้อความสอบถามนั้น ผลสรุปของการวิจัยนั้นเห็นได้ว่าประสิทธิผลของระบบค้นคืนสารสนเทศที่ดีนั้นควรใช้เทอมที่มีน้ำหนัก ในการดึงจากเอกสารที่กำหนดซึ่งความยาวน้อยที่สุดของเอกสารที่ใช้ควรเป็น Metadata ที่เป็น รายการย่อ(abstract)

ดรรชนีหรือตัวแทนเอกสารนี้แบ่งออกเป็นหลายประเภทตามลักษณะการใช้งาน เช่น

- ดรรชนีที่ถูกควบคุมคำศัพท์(Controlled Vocabulary Index)
- ดรรชนีที่เป็นแบบฉบับ(Conventional Index)
- ดรรชนีที่นำมาเชื่อมโยงกัน(Coordinate Index)
- ดรรชนีที่เอามาจากชื่อเรื่อง(Keyword from Title Index)
- ดรรชนีข้อความ(Citation Index)

ดรรชนีควบคุมคำศัพท์ (Controlled Vocabulary Index)

เป็นคำศัพท์มาตรฐานที่มีผู้จัดทำไว้ในคู่มือหรือในบัญชีคำศัพท์ที่เรียกว่า Thesaurus ซึ่งคำศัพท์เหล่านี้สามารถช่วยในการขจัดความซ้ำซ้อนและจัดหมวดหมู่ของคำของเอกสาร

ดรรชนีที่เป็นแบบฉบับ(Conventional Index)

เป็นดรรชนีประเภทหนึ่งที่สามารถแบ่งเป็น 3 ชนิดคือ

1. Alphabetical Subject Index

เป็นดรรชนีที่จัดเรียงตามลำดับตัวอักษร ไม่มีโครงสร้างตายตัว มีความยืดหยุ่น เช่นดรรชนีชื่อ ดรรชนีผู้แต่ง ดรรชนีชื่อเรื่อง ข้อเสียของดรรชนีชนิดนี้คือ ไม่มีการควบคุมคำศัพท์ ทำให้เกิดปัญหาในเรื่องของความหมาย เช่น Building Structure กับ Structure of Buildings

2. Hierarchically Classified Index

เป็นดรรชนีที่จัดแยกหมวดหมู่และเรียงตามลำดับชั้น เช่นดรรชนีในสาขาวิชาการต่างๆ เช่น การจัดหมวดหมู่หนังสือระบบทศนิยมดิวอี้

600	Economics and management
630	Economics evaluation
633	Manufacturing cost estimation
633.1	Direct cost
633.13	Utilities
633.132	Water

ข้อเสียของดรรชนีชนิดนี้คือ ผู้จัดทำต้องคาดการณ์จัดเตรียมเนื้อที่ไว้ให้เพียงพอ ไม่เหมาะกับการสืบค้นเรื่องราวที่มีความเฉพาะเจาะจง เนื่องจากไม่สะดวกและเสียเวลาในการสืบค้น

3. Alphabetic-Classified Index

เป็นดรรชนีผสมระหว่างแบบที่ 1 และ 2 ซึ่งมีการแบ่งหมวดหมู่ของคำและเรียงตามลำดับตัวอักษร ทำให้สะดวกมากขึ้น

ดรรชนีที่นำมาเชื่อมโยงกัน(Coordinate Index)

เป็นดรรชนีที่นำเอาคำตั้งแต่ 2 คำขึ้นไปมาเชื่อมโยงกันเป็นคำใหม่ แบ่งเป็น 2 แบบคือ

1. Pre-Coordinate Index

เป็นการเตรียมรายการดรรชนีที่เป็นคำเชื่อมโยงไว้ให้ผู้ใช้ สามารถค้นหาข้อมูลได้ทันที เช่น ประโยคว่า University Library Administration ผู้ใช้สามารถสืบค้นได้อย่างสะดวก

2. Post-Coordinate Index

ผู้ใช้ต้องนำเอาคำแต่ละคำมาใช้เป็นดรรชนีในการค้นหา ซึ่งใช้ประสบการณ์และความสามารถของบุคคลที่จะนำคำสำคัญต่าง ๆมาเชื่อมโยงเอง

ดรรชนีที่เอามาจากชื่อเรื่อง(Keyword from Title Index)

เป็นการทำงานอัตโนมัติโดยใช้คอมพิวเตอร์ช่วย แบ่งเป็น

1. KWIC Index(Key-word-In-Context Index)

เป็นดรรชนีที่นำเอาคำสำคัญจากชื่อเรื่องของเอกสารมาหมุนเวียนเป็นดรรชนี โดยจัดเรียงตามลำดับตัวอักษร แบ่งเป็น 3 ส่วนคือ

1.1 รหัสหรือหมายเลขเอกสาร (Document Identification Code)

1.2 คำสำคัญต่าง ๆ (Index Words) ที่ปรากฏอยู่ในชื่อเรื่อง และนำมาหมุนเวียนเพื่อทำดรรชนี

1.3 คำต่าง ๆในชื่อเรื่องที่เหลือจากคีย์เวิร์ด เราเรียกว่า context ช่วยให้ความหมายของคำสำคัญเด่นชัดขึ้น และช่วยให้ผู้ใช้ทราบชื่อเรื่องทั้งหมดของเอกสารด้วย

ตัวอย่างเอกสารชื่อ “User Preference in Published Indexes”
สามารถกระทำดรรชนีโดยวิธี KWIC Index ได้ดังนี้

RENCE IN PUBLISHED INDEXES/USER PREFE 4.411

USER PREFERENCE IN PUBLISHED INDEXES/ 4.411

ทั้งสามส่วนอยู่บรรทัดเดียวกัน

- คำสำคัญจะอยู่สดมภ์กลางและเรียงตามลำดับอักษร
- context จะเรียงต่อจากคำสำคัญ ถ้าหมดเนื้อที่จะมาต่อที่สดมภ์ซ้ายสุด
- ตอนจบของชื่อเรื่องลงท้ายด้วย /
- หมายเลขเอกสารจะอยู่ที่สดมภ์ขวาสุด

2. KWOC Index(Keyword-Out-of Context Index) หรือ KWAC

เป็นการทำดรรชนีที่นำเอาคำสำคัญจากชื่อเรื่องของเอกสาร โดยนำคำสำคัญมาไว้ที่ริมซ้ายสุด ตามด้วยชื่อเรื่องทั้งหมดของเอกสาร

จากตัวอย่างข้างต้นสามารถกระทำโดยวิธี KWOC ได้ดังนี้

INDEXES

USER PREFERENCE IN PUBLISHED INDEXES 4.411

PUBLISHED

USER PREFERENCE IN PUBLISHED INDEXES 4.411

ข้อดีของการทำดรรชนีจากชื่อเรื่อง รวดเร็ว และลดค่าใช้จ่ายได้มาก นอกจากนี้ยังสามารถประมวลผลเอกสารจำนวนมากได้ ซึ่งวิธีการทำดรรชนีจากชื่อเรื่องนี้จะเหมาะกับเอกสารสาขาเทคโนโลยีและวิทยาศาสตร์ มากกว่าสาขามนุษย์ศาสตร์และสังคมศาสตร์

ข้อเสีย ไม่มีการควบคุมคำศัพท์ ทำให้เกิดปัญหาในการค้นหา กล่าวคือค้นหาแล้วไม่ได้เอกสารในเรื่องที่ต้องการอาจเป็นเรื่องอื่นๆที่ไม่เกี่ยวข้อง

ดรรชนีข้อความ(Citation Index)

กระทำโดยคอมพิวเตอร์ โดยนำรายการอ้างอิงท้ายเอกสารซึ่งเป็นแหล่งที่มาของข้อมูลในเอกสารมาทำการวิเคราะห์หาดรรชนี เหมาะกับสิ่งพิมพ์ประเภทต่างๆ เช่น ดรรชนีของบทความ ดรรชนีของวารสาร ดรรชนีของหนังสือพิมพ์ ในรูปของบัตรรายการหนังสือในห้องสมุด เป็นต้น

2.5 ดรรชนีเพิ่มน้ำหนักคำศัพท์เฉพาะ(Index term weighting)

ดรรชนีที่ดีนั้นจากการวิจัยนั้นควรเป็นดรรชนีที่มีการใช้น้ำหนักเป็นตัวตัดสินในการเลือกคำที่สำคัญหรือดรรชนี

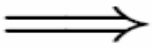
ระบบที่ใช้ในการสร้างดรรชนี แบ่งได้เป็น 2 ประเภทคือ

1. ระบบการกำหนดดรรชนีด้วยมือคน (Assigned-Term System) ใช้ความรู้ความสามารถ วิจารณญาณในการพิจารณา มีการควบคุมคำศัพท์ ประเภทของดรรชนีที่ได้จากระบบนี้ได้แก่ รายการหัวเรื่อง(subject-heading lists) และบัญชีศัพท์สัมพันธ์ (thesaurus)
2. ระบบการดึงดรรชนีจากเอกสาร(Derived-Term System) เช่น ดรรชนีผู้แต่ง ดรรชนีชื่อเรื่อง ดรรชนีข้อความ บางครั้งเรียกว่า Extraction Method

ตัวอย่างที่ 2.3 แสดงการสร้างดรรชนี(index)

สมมุติว่าเอกสาร 1 ถึง 6 ประกอบด้วยข้อความดังนี้

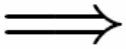
Document	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old



Number	Text	Documents
1	cold	1, 4
2	days	3, 6
3	hot	1, 4
4	in	2, 5
5	it	4, 5
6	like	4, 5
7	nine	3, 6
8	old	3, 6
9	pease	1, 2
10	porridge	1, 2
11	pot	2, 5
12	some	4, 5
13	the	2, 5

การทำงานนั้นจะทำการแยกคำต่างๆ โดยนำมาจัดเรียงลำดับตามตัวอักษรจากน้อยไปหามาก และระบุว่าคำนั้นอยู่ในเอกสารใด ต่อจากนั้นมีการระบุตำแหน่งที่อยู่ของคำนั้นในเอกสารดังนี้

Document	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old



Number	Text	(Document; Word)
1	cold	(1; 6), (4; 8)
2	days	(3; 2), (6; 2)
3	hot	(1; 3), (4; 4)
4	in	(2; 3), (5; 4)
5	it	(4; 3, 7), (5; 3)
6	like	(4; 2, 6), (5; 2)
7	nine	(3; 1), (6; 1)
8	old	(3; 3), (6; 3)
9	pease	(1; 1, 4), (2; 1)
10	porridge	(1; 2, 5), (2; 2)
11	pot	(2; 5), (5; 6)
12	some	(4; 1, 5), (5; 1)
13	the	(2; 4), (5; 5)

ต่อจากนั้นนับความถี่ของคำในเอกสารเพื่อให้น้ำหนักคำแต่ละคำ สำหรับระบบน้ำหนักคำนี้ผู้คิดค้นในการเอาความถี่ของคำมาใช้คือ H.P Luhn ซึ่งรู้จักกันในชื่อของ Zipf หลักการคือ ให้นำความถี่ไปคูณด้วยลำดับของคำจะได้เป็นค่าคงที่หนึ่งดังนี้

ลำดับ	***คำ***	***ความถี่***	***ลำดับ*ความถี่***
1	the	69,971	69,971
2	of	36,411	72,822
3	and	28,852	86,556

ความถี่ของคำ (Term Frequency) freqi,j เป็นจำนวนความถี่ของคำที่ ki ในเอกสาร dj ความถี่ของคำในรูปปกติ (Normalization form) คือ

$$f_{i,j} = \frac{freq_{i,j}}{\max l.freq_{l,j}}$$

ในการกำหนดน้ำหนักของเทอมนั้น เราวัดจาก tf*idf โดย

- Term frequency (tf)
- Inverse document frequency (idf)

ซึ่งต้องคำนวณน้ำหนักของแต่ละเอกสาร

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

โดย

$T_k = \text{term } k \text{ in document } D_i$

$tf_{ik} = \text{frequency of term } T_k \text{ in document } D_i$

$idf_k = \text{inverse document frequency of term } T_k \text{ in } C$

$N = \text{total number of documents in the collection } C$

$n_k = \text{the number of documents in } C \text{ that contain } T_k$

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

ซึ่งค่าของ idf จะมีค่าสูงถ้าเป็นเทอมที่หายาก และค่าของ idf จะมีค่าต่ำถ้าเป็นเทอมที่มีความถี่สูง

สมมุติว่ามีเอกสารทั้งหมดจำนวนเท่ากับ 10000 ฉบับ

$$\log\left(\frac{10000}{10000}\right) = 0$$

$$\log\left(\frac{10000}{5000}\right) = 0.301$$

$$\log\left(\frac{10000}{20}\right) = 2.698$$

$$\log\left(\frac{10000}{1}\right) = 4$$

ตัวอย่างที่ 2.4 ถ้ามีคำ 3 คำ คือ “ค้นคืน” , .”คอมพิวเตอร์. และ “เอกสาร.

สมมุติว่า คำว่า “**ค้นคืน**” ปรากฏในเอกสารจำนวนเท่ากับ 100

คำว่า “**คอมพิวเตอร์**” ปรากฏในเอกสารจำนวนเท่ากับ 500

และคำว่า “**เอกสาร**” ปรากฏในเอกสารจำนวนเท่ากับ 900

จะเห็นว่า คำว่า “เอกสาร” แทบจะไม่สามารถเป็นตัวแทนของเอกสารได้เพราะเกิดในเอกสารทุกฉบับ

ค่าของ idf ของคำว่า “ค้นคืน” มีค่าเท่ากับ 4.222 ,

ค่าของ idf ของคำว่า “คอมพิวเตอร์” มีค่าเท่ากับ 2.00

ค่าของ idf ของคำว่า “เอกสาร” มีค่าเท่ากับ 1.132

ถ้ามีเอกสารทั้งสิ้น 1000 ฉบับ การนำ idf ไปคูณเข้าจะเป็นการเพิ่มความสำคัญให้กับคำดังกล่าว

การวิจัยของ Sparck Jones ทำการวิจัยทางสถิติโดยนำเอกสาร จำนวน N เอกสาร และ index term ที่เกิดขึ้นใน n ของเอกสารเหล่านั้นมาหาน้ำหนักของคำ ผลของการทดลองทำให้ทราบว่า น้ำหนักที่มีค่า idf ที่มีค่าเท่ากับ log(N/n) + 1 จะมีประสิทธิภาพในการดึงที่มากกว่าการใช้แบบไม่มีน้ำหนัก Salton and Yang ได้มีการรวมวิธีการของการให้น้ำหนักทั้งความถี่ทั้งภายในเอกสารและภายนอกเอกสาร ผลของการวิจัยทำให้ทราบว่าคำศัพท์ที่เกิดขึ้นบ่อยๆนั้นไม่มีประโยชน์ในการค้นคืนสารสนเทศ แต่คำศัพท์ที่เกิดขึ้นในระดับกลางๆกลับมีประโยชน์มากในการค้นคืนมากกว่า

2.6 Probabilistic Indexing

เป็นการใช้โมเดลเชิงปริมาณ(Quantitative Model) บนพื้นฐานของสมมุติฐานทางสถิติมาใช้สำหรับการสร้างดรรชนีแบบอัตโนมัติเพื่อแสดงถึงการกระจายของคำในข้อความ Bookstrin,Swanson and Harter ได้ทำการศึกษาความแตกต่างระหว่าง word-type กับ word-token

คำที่สำคัญในตัวแบบได้แก่

1. Word-Type ชนิดของคำหนึ่ง
2. Word-Token เป็นการเกิดขึ้นของคำโดยเฉพาะที่จุดหนึ่งในข้อความของเอกสาร หรือ บทคัดย่อ (Abstract)

ความถี่ของการเกิดขึ้นของคำ w ในเอกสารหนึ่ง หมายถึง จำนวน Word-Tokens ที่เกิดขึ้นในเอกสารนั้นที่สอดคล้องกับของ Word-Type หนึ่งที่ไม่ซ้ำกัน

Stone and Rubinoff ,Demeraul and Dennis ได้มีการวิจัยทางสถิติเพื่อดูพฤติกรรมของคำที่ระบุ(specialty) ว่าแตกต่างจากคำฟังก์ชัน (function)คำอย่างไร ซึ่งเขาพบว่า ฟังก์ชันของคำนั้นจะกระจายไปยังเอกสารทั้งหมด(poisson distribution) ดังนั้นถ้าต้องการค้นหาการกระจายของคำ w ในข้อความนั้น ดังนั้นความน่าจะเป็น f(n) เป็นความน่าจะเป็นที่จะมีการเกิดขึ้นของคำ w โดย n เป็นเหตุการณ์ของฟังก์ชันคำ w

$$f(n) = e^{-x} x^{n} / n$$

โดย x คือเป็นจำนวนเฉลี่ยของการเกิดขึ้นของคำ w ในข้อความ ซึ่งเปลี่ยนแปลงตามความยาวของข้อความ เช่น คำว่า FOR จะเกิดขึ้นกระจายไปทั่วเอกสาร แต่สำหรับคำว่า WAR เป็นคำเจาะจงซึ่งให้เนื้อหาได้ดีกว่า และพบหรือเกิดขึ้นในเอกสารค่อนข้างน้อยกว่า

Harter ได้มีการวิจัยและทดลองสมมุติฐานเพื่อหากฎเกณฑ์สำหรับกำหนดเทอมดรรชนี (index term) ที่เหมาะสมเพื่อเป็นตัวแทนของเอกสาร สมมุติฐานคือ

(1) ความน่าจะเป็น ของเอกสารที่เกี่ยวข้องที่พบจากการร้องขอสารสนเทศนั้น เป็นฟังก์ชันของขอบเขตความสัมพันธ์ที่เกี่ยวข้องกับหัวข้อในเอกสาร
(2) จำนวนของการเกิดขึ้นของคำ(token) ในเอกสาร เป็นฟังก์ชันของขอบเขตที่อ้างอิงไปยังคำในเอกสาร

เขาได้มีการกำหนดหัวข้อ(topic) ด้วยเนื้อหา(subject) เพื่อทำการร้องขอ(request)จากการอ้างอิงด้วยคำ(word) การกระจายของ w สามารถผสมการกระจายของคำได้ที่เรียกว่า 2-P0isson model ดังนี้

$$f(n) = \frac{p_i e^{-x1} x1^n}{n} + \frac{(1 - p_i e^{-x2} x2^n)}{n}$$

โดย pi หมายถึงความน่าจะเป็นของการสุ่มเอกสารที่เป็นชุดย่อย

x1 และ x2 เป็นค่าเฉลี่ยของเหตุการณ์ทั้งสอง

ซึ่งเป็นการใช้สถิตินั้นบรรยายพฤติกรรมอง content-bearing ซึ่งเป็นการสร้างลักษณะเฉพาะในคลาสของเอกสารเพื่อให้สามารถค้นหาหรือเข้าถึงได้อย่างแม่นยำในขอบเขตที่คาดหวังได้ ในกรณีที่เราสามารถคำนวณความน่าจะเป็นของความเกี่ยวพันธ์ของเอกสารหนึ่งไปยังคลาสได้นั้น การคำนวณจะมีการเปลี่ยนแปลงดังนี้

$$\frac{p_i e^{-x1} x1^k}{p_i e^{-x1} x1^k + (1 - p_i e^{-x2} x2^k)}$$

ผลลัพธ์ที่ได้คืออัตราส่วนที่ใช้กำหนดเทอมดรรชนี(index term) w ของเหตุการณ์ k ครั้งในเอกสาร อัตราส่วนนี้เป็นข้อเท็จจริงของความน่าจะเป็นที่คำ w ของเอกสารที่ระบุที่เป็นของคลาสหนึ่งนั้น ค่าเฉลี่ยของขอบเขต x1 ที่ให้นั้นบรรจุเหตุการณ์ k อัตราส่วนนี้จะเปรียบเทียบถึงค่าใช้จ่ายของฟังก์ชันในการที่ผู้ใช้มีการเข้าถึงข้อมูลที่ผิดพลาด อาจเป็นการค้นคืนสารสนเทศที่ตรงวัตถุประสงค์

2.7 ปัญหาและการแก้ไขในการวิเคราะห์ข้อความ

ในการวิเคราะห์ข้อความ เพื่อสร้างเป็นดรรชนีนั้นมีวิธีการสร้างดรรชนี(Index Construction Methods) มีอยู่ด้วยกันหลายวิธีเช่น Memory-based inversion , Sort-based inversion ,All above, combined with compression , FAST-INV ,Based on text partitioning และ ฯลฯ

จากการศึกษานั้นพบว่า ถ้าขนาดของข้อความ 5 GB มีเอกสารมากถึง 5 ล้าน และมีหน่วยความจำหลัก 40 MB มีผลการปฏิบัติงานดังนี้

Method	Memory (Mb)	Disk (Mb)	Time (hours)
Linked lists (memory)	4000	0	6
Linked lists (disk)	30	4000	1100
Sort-based	40	8000	20
⋮			
FAST-INV (no extra disk)	40	0	79
FAST-INV (extra disk)	40	4000	12
Text-based partition	40	35	15

ในแต่ละวิธีจะมีจุดเด่นและจุดด้อยต่างๆกัน ดังนั้นการเลือกวิธีการใดนั้นขึ้นอยู่กับความเหมาะสมรวมทั้งประสิทธิภาพและประสิทธิผลที่ต้องการ

ปัจจัยที่สำคัญที่ช่วยควบคุมประสิทธิภาพของดรรชนีคือ

1. indexing exhaustivity ถูกกำหนดเป็นจำนวนของหัวข้อดรรชนีที่แตกต่างกัน ซึ่งบรรจุดรรชนีต่างๆที่ให้ความหมายครอบคลุมไปทุกๆเรื่องราวสาขาที่มีอยู่ในกลุ่มของเอกสารทั้งหมด โดยค่า Exhaustivity สูงจะให้ค่า recall ที่สูงซึ่งเป็นความสามารถของระบบในการดึงเอกสารที่เกี่ยวข้องออกมา และให้ค่า precision ที่ต่ำ
2. Index Language specificity เป็นความสามารถของภาษาดรรชนีที่บรรยายหัวข้ออย่างแม่นยำ บรรจุดรรชนีที่มีความหมายเจาะจงในวงแคบและถูกต้อง โดยค่าของ Specificity ที่สูงจะให้ค่า Precision สูงด้วย และได้ค่า recall ที่ต่ำ กล่าวคือเอกสารที่ถูกดึงออกมาส่วนใหญ่เป็นเอกสารที่เกี่ยวข้อง โดยดรรชนีในลักษณะนี้เป็นดรรชนีที่มีตัวบ่งชี้เนื้อหา (Content Indicators) เพิ่มเติม เช่น น้ำหนักของคำหรือความสัมพันธ์ระหว่างคำนี้กับดรรชนีอื่นๆ

ในการวิเคราะห์ข้อความนั้น เกิดปัญหาต่างๆในหลายๆเรื่องเกี่ยวกับความหมายของคำในภาษาอังกฤษ สามารถสรุปได้เป็นข้อๆดังนี้

1. คำๆเดียวกันที่บางครั้งมีความสำคัญต่อเนื้อหาใจความ แต่บางครั้งไม่สำคัญ เช่น คำว่า CAN เป็นคำกริยาแปลว่า สามารถ แต่ CAN เมื่อใช้เป็นคำนาม กลับแปลว่ากระป๋อง
2. คำที่เหมือนกันมีความหมายต่างกัน เมื่อใช้ในเรื่องที่ต่างกัน เช่น Baseball กับ Military Base

3. คำที่เขียนต่างกันแต่มีความหมายเหมือนกัน เช่นคำว่า “restaurant “ และ “café”
4. คำบางคำไม่ได้ให้ความสำคัญต่อใจความเลย เช่น คำสรรพนาม คำบุพบท คำสันธาน ต้องขจัดออกจากข้อความ
5. ความหมายของคำในประโยคต่างๆอาจไม่ชัดเจน ต้องพิจารณาจากประโยคก่อนหน้า หรือมีความคลุมเครือ(ambiguous) เช่น
    - คำว่า “bat” ไม่ทราบว่าเป็น baseball หรือ mamal
    - คำว่า “apple” ไม่ทราบว่าเป็น company หรือ fruit
    - คำว่า “bit” ไม่ทราบว่าเป็น unit of data หรือ act of eating
6. คำบางคำแปรเปลี่ยนไปตามยุคสมัย เช่น Monitor หรือ Operating System

วิธีการแก้ไขปัญหาต่างๆเหล่านี้มีการใช้ **<u>พจนานุกรมในระบบ IR</u>** ซึ่งพจนานุกรมนี้มีด้วยกันหลายชนิดได้แก่

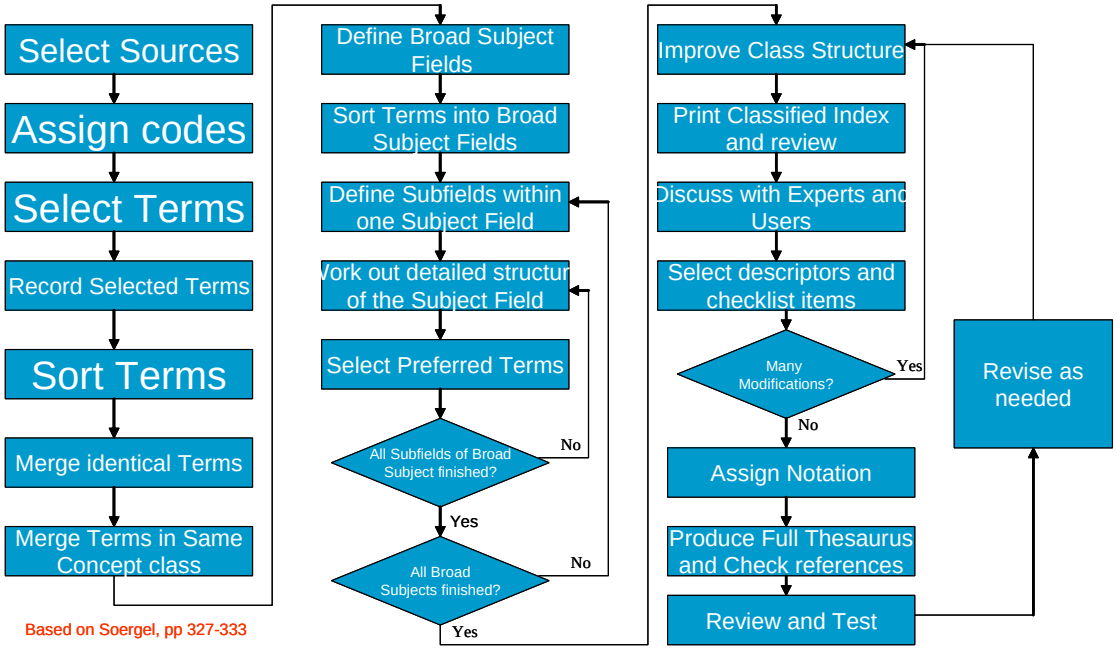
1. Negative Dictionary เป็นพจนานุกรมที่เก็บคำต่างๆที่ต้องขจัดออกจากข้อความ ในการวิเคราะห์ข้อความ เช่น stop word list , การตัด Suffix
2. Synonym Dictionary หรือ บัญชีศัพท์สัมพันธ์(Thesaurus) เป็นพจนานุกรมที่เก็บคำพ้อง(Synonym) ที่แยกเป็นลำดับขั้น และคำต่างๆที่มีความสัมพันธ์กับคำนั้นในหลายๆระดับ เช่น มีความหมายพ้องกัน (Related Term: RT) ,มีความหมายเจาะจงมากกว่า (Narrower Term : NT) , มีความหมายกว้างกว่า(Boarder Term :BT) โดยแสดงถึงคำศัพท์ที่ถูกควบคุม

“ A Thesaurus is a collection of selected vocabulary (preferred terms or descriptors) with links among synonymous, equivalent, broader, narrower and other related terms “

หลักเกณฑ์ในการสร้างบัญชีศัพท์สัมพันธ์

- เก็บถ้อยคำในรูปแบบต่างๆที่ไม่ค่อยพบบ่อยนัก
- ไม่เก็บคำที่มีความถี่สูงหรือพบบ่อยมาก เช่น and, in, but , the
- คำที่ไม่แสดงถึงความสำคัญอย่างชัดเจน เช่น hand
- คำที่มีหลายๆความหมาย เก็บความหมายของคำไว้ด้วย

โดยมีกระบวนการสร้างบัญชีคำศัพท์ ดังรูปที่ 2.8



รูปที่ 2.8 : แสดง Flow of Work in Thesaurus Construction

ตัวอย่างที่ 2.4 แสดงบัญชีคำศัพท์คำว่า Secondary Schools (โรงเรียนมัธยม) จะมีการเก็บ

BT	Education	(การศึกษา)
RT	Primary School	(โรงเรียนประถม)
NT	Classes	(ชั้นเรียน , เวลาเรียน)

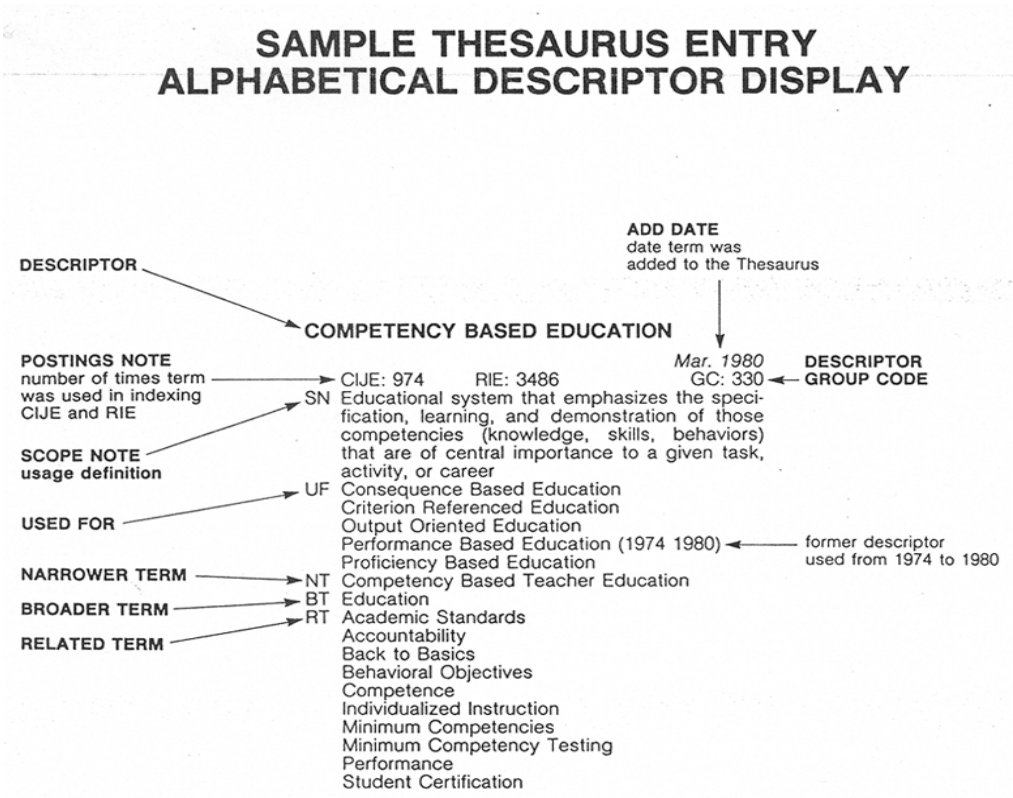
การสร้างบัญชีคำศัพท์นั้นมีหลายๆองค์กรได้มีการสร้างไว้เพื่อให้ใช้เป็นมาตรฐานสากล ดังนี้

National and International Standards for Thesauri

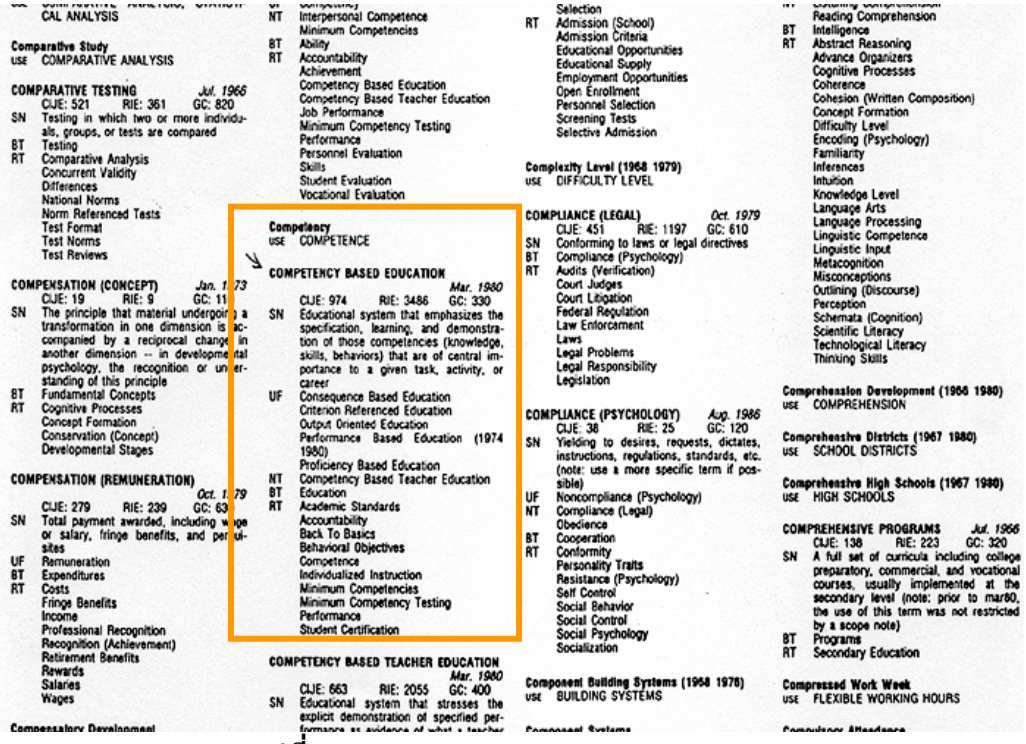
- ANSI/NISO z39.19-1994 — American National Standard Guidelines for the Construction, Format and Management of Monolingual Thesauri
- ANSI/NISO Draft Standard Z39.4-199x — American National Standard Guidelines for Indexes in Information Retrieval

- ISO 2788 — Documentation — Guidelines for the establishment and development of monolingual thesauri
- ISO 5964 — Documentation — Guidelines for the establishment and development of multilingual thesauri

ตัวอย่างที่ 2.5 แสดง ERIC –Thesaurus ในการสร้างดรรชนีในลักษณะต่าง ๆ



รูปที่ 2.9 :แสดง ERIC Thesaurus – Entry



CAL ANALYSIS

Comparative Study
use COMPARATIVE ANALYSIS

COMPARATIVE TESTING *Jul. 1966*
CIJE: 521 RIE: 361 GC: 820
SN Testing in which two or more individuals, groups, or tests are compared
BT Testing
RT Comparative Analysis
Concurrent Validity
Differences
National Norms
Norm Referenced Tests
Test Format
Test Norms
Test Reviews

COMPENSATION (CONCEPT) *Jan. 1973*
CIJE: 19 RIE: 9 GC: 11
SN The principle that material undergoing a transformation in one dimension is accompanied by a reciprocal change in another dimension -- in developmental psychology, the recognition or understanding of this principle
BT Fundamental Concepts
RT Cognitive Processes
Concept Formation
Conservation (Concept)
Developmental Stages

COMPENSATION (REMUNERATION) *Oct. 1979*
CIJE: 279 RIE: 239 GC: 63
SN Total payment awarded, including wage or salary, fringe benefits, and perquisites
UF Remuneration
BT Expenditures
RT Costs
Fringe Benefits
Income
Professional Recognition
Recognition (Achievement)
Retirement Benefits
Rewards
Salaries
Wages

NT Interpersonal Competence
Minimum Competencies
BT Ability
RT Accountability
Achievement
Competency Based Education
Competency Based Teacher Education
Job Performance
Minimum Competency Testing
Performance
Personnel Evaluation
Skills
Student Evaluation
Vocational Evaluation

Competency
use COMPETENCE

COMPETENCY BASED EDUCATION *Mar. 1980*
CIJE: 974 RIE: 3486 GC: 330
SN Educational system that emphasizes the specification, learning, and demonstration of those competencies (knowledge, skills, behaviors) that are of central importance to a given task, activity, or career
UF Consequence Based Education
Criterion Referenced Education
Output Oriented Education
Performance Based Education (1974 1980)
Proficiency Based Education
NT Competency Based Teacher Education
BT Education
RT Academic Standards
Accountability
Back To Basics
Behavioral Objectives
Competence
Individualized Instruction
Minimum Competencies
Minimum Competency Testing
Performance
Student Certification

COMPETENCY BASED TEACHER EDUCATION *Mar. 1980*
CIJE: 663 RIE: 2055 GC: 400
SN Educational system that stresses the explicit demonstration of specified per-

Selection
RT Admission (School)
Admission Criteria
Educational Opportunities
Educational Supply
Employment Opportunities
Open Enrollment
Personnel Selection
Screening Tests
Selective Admission

Complexity Level (1968 1979)
use DIFFICULTY LEVEL

COMPLIANCE (LEGAL) *Oct. 1979*
CIJE: 451 RIE: 1197 GC: 610
SN Conforming to laws or legal directives
BT Compliance (Psychology)
RT Audits (Verification)
Court Judges
Court Litigation
Federal Regulation
Law Enforcement
Laws
Legal Problems
Legal Responsibility
Legislation

COMPLIANCE (PSYCHOLOGY) *Aug. 1986*
CIJE: 36 RIE: 25 GC: 120
SN Yielding to desires, requests, dictates, instructions, regulations, standards, etc. (note: use a more specific term if possible)
UF Noncompliance (Psychology)
NT Compliance (Legal)
Obedience
BT Cooperation
RT Conformity
Personality Traits
Resistance (Psychology)
Self Control
Social Behavior
Social Control
Social Psychology
Socialization

Component Building Systems (1968 1978)
use BUILDING SYSTEMS

Reading Comprehension
BT Intelligence
RT Abstract Reasoning
Advance Organizers
Cognitive Processes
Coherence
Cohesion (Written Composition)
Concept Formation
Difficulty Level
Encoding (Psychology)
Familiarity
Inferences
Intuition
Knowledge Level
Language Arts
Language Processing
Linguistic Competence
Linguistic Input
Metacognition
Misconceptions
Outlining (Discourse)
Perception
Schemata (Cognition)
Scientific Literacy
Technological Literacy
Thinking Skills

Comprehension Development (1966 1980)
use COMPREHENSION

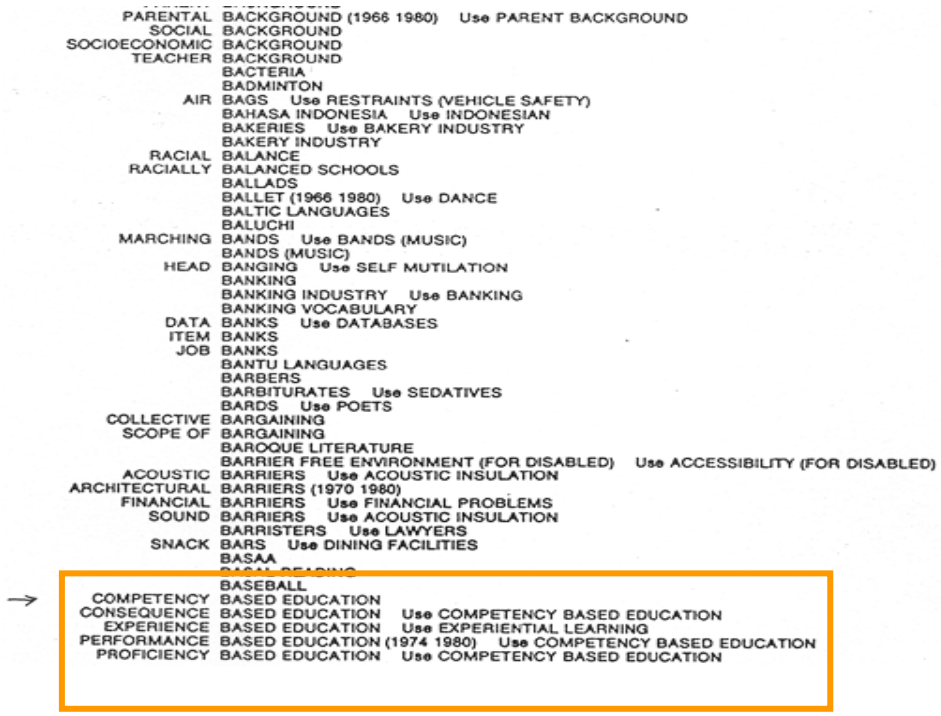
Comprehensive Districts (1967 1980)
use SCHOOL DISTRICTS

Comprehensive High Schools (1967 1980)
use HIGH SCHOOLS

COMPREHENSIVE PROGRAMS *Jul. 1966*
CIJE: 138 RIE: 223 GC: 320
SN A full set of curricula including college preparatory, commercial, and vocational courses, usually implemented at the secondary level (note: prior to mar80, the use of this term was not restricted by a scope note)
BT Programs
RT Secondary Education

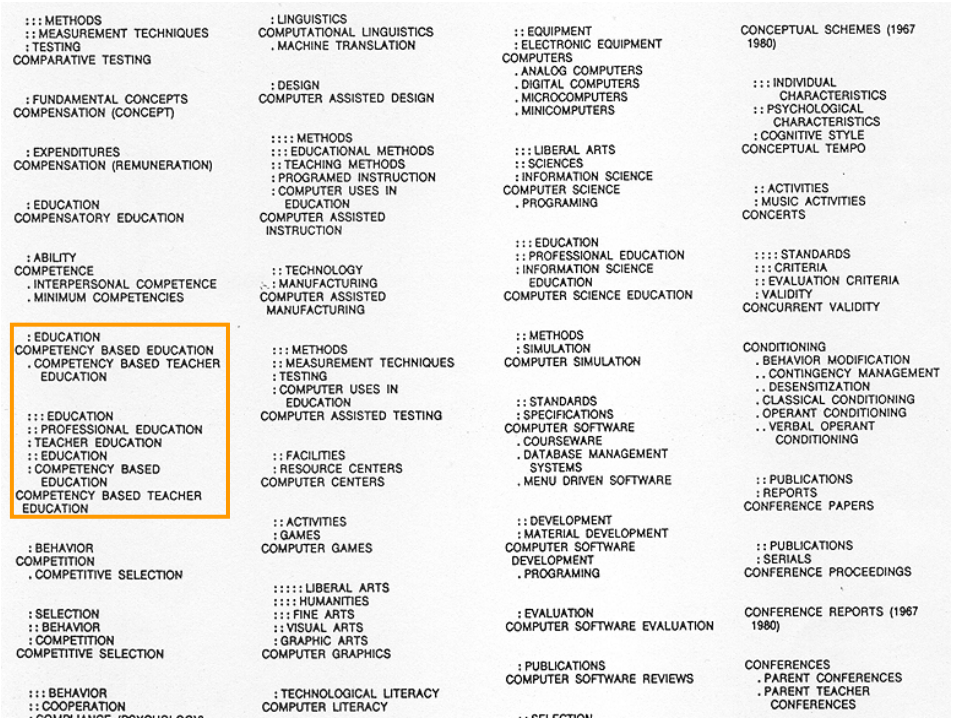
Compressed Work Week
use FLEXIBLE WORKING HOURS

รูปที่ 2.10 :แสดง ERIC Thesaurus – Alphabetic



PARENTAL BACKGROUND (1966 1980) Use PARENT BACKGROUND
SOCIAL BACKGROUND
SOCIOECONOMIC BACKGROUND
TEACHER BACKGROUND
BACTERIA
BADMINTON
AIR BAGS Use RESTRAINTS (VEHICLE SAFETY)
BAHASA INDONESIA Use INDONESIAN
BAKERIES Use BAKERY INDUSTRY
BAKERY INDUSTRY
RACIAL BALANCE
RACIALLY BALANCED SCHOOLS
BALLADS
BALLET (1966 1980) Use DANCE
BALTIC LANGUAGES
BALUCHI
MARCHING BANDS Use BANDS (MUSIC)
BANDS (MUSIC)
HEAD BANGING Use SELF MUTILATION
BANKING
BANKING INDUSTRY Use BANKING
BANKING VOCABULARY
DATA BANKS Use DATABASES
ITEM BANKS
JOB BANKS
BANTU LANGUAGES
BARBERS
BARBITURATES Use SEDATIVES
BARDS Use POETS
COLLECTIVE BARGAINING
SCOPE OF BARGAINING
BAROQUE LITERATURE
BARRIER FREE ENVIRONMENT (FOR DISABLED) Use ACCESSIBILITY (FOR DISABLED)
ACOUSTIC BARRIERS Use ACOUSTIC INSULATION
ARCHITECTURAL BARRIERS (1970 1980)
FINANCIAL BARRIERS Use FINANCIAL PROBLEMS
SOUND BARRIERS Use ACOUSTIC INSULATION
BARRISTERS Use LAWYERS
SNACK BARS Use DINING FACILITIES
BASAA
BASAL READING
BASEBALL
COMPETENCY BASED EDUCATION
CONSEQUENCE BASED EDUCATION Use COMPETENCY BASED EDUCATION
EXPERIENCE BASED EDUCATION Use EXPERIENTIAL LEARNING
PERFORMANCE BASED EDUCATION (1974 1980) Use COMPETENCY BASED EDUCATION
PROFICIENCY BASED EDUCATION Use COMPETENCY BASED EDUCATION

รูปที่ 2.11 :แสดง ERIC Thesaurus – KWIC Index



: : : METHODS
: : MEASUREMENT TECHNIQUES
: TESTING
COMPARATIVE TESTING

: FUNDAMENTAL CONCEPTS
COMPENSATION (CONCEPT)

: EXPENDITURES
COMPENSATION (REMUNERATION)

: EDUCATION
COMPENSATORY EDUCATION

: ABILITY
COMPETENCE
. INTERPERSONAL COMPETENCE
. MINIMUM COMPETENCIES

: EDUCATION
COMPETENCY BASED EDUCATION
. COMPETENCY BASED TEACHER EDUCATION

: : : EDUCATION
: : PROFESSIONAL EDUCATION
: TEACHER EDUCATION
: : EDUCATION
: COMPETENCY BASED EDUCATION
COMPETENCY BASED TEACHER EDUCATION

: BEHAVIOR
COMPETITION
. COMPETITIVE SELECTION

: SELECTION
: : BEHAVIOR
: COMPETITION
COMPETITIVE SELECTION

: : : BEHAVIOR
: : COOPERATION

: LINGUISTICS
COMPUTATIONAL LINGUISTICS
. MACHINE TRANSLATION

: DESIGN
COMPUTER ASSISTED DESIGN

: : : : METHODS
: : : EDUCATIONAL METHODS
: : TEACHING METHODS
: PROGRAMED INSTRUCTION
: COMPUTER USES IN EDUCATION
COMPUTER ASSISTED INSTRUCTION

: : TECHNOLOGY
: MANUFACTURING
COMPUTER ASSISTED MANUFACTURING

: : : METHODS
: : MEASUREMENT TECHNIQUES
: TESTING
: COMPUTER USES IN EDUCATION
COMPUTER ASSISTED TESTING

: : FACILITIES
: RESOURCE CENTERS
COMPUTER CENTERS

: : ACTIVITIES
: GAMES
COMPUTER GAMES

: : : : : LIBERAL ARTS
: : : : HUMANITIES
: : : FINE ARTS
: : VISUAL ARTS
: GRAPHIC ARTS
COMPUTER GRAPHICS

: TECHNOLOGICAL LITERACY
COMPUTER LITERACY

: : EQUIPMENT
: ELECTRONIC EQUIPMENT
COMPUTERS
. ANALOG COMPUTERS
. DIGITAL COMPUTERS
. MICROCOMPUTERS
. MINICOMPUTERS

: : : LIBERAL ARTS
: : SCIENCES
: INFORMATION SCIENCE
COMPUTER SCIENCE
. PROGRAMING

: : : EDUCATION
: : PROFESSIONAL EDUCATION
: INFORMATION SCIENCE EDUCATION
COMPUTER SCIENCE EDUCATION

: : METHODS
: SIMULATION
COMPUTER SIMULATION

: : STANDARDS
: SPECIFICATIONS
COMPUTER SOFTWARE
. COURSEWARE
. DATABASE MANAGEMENT SYSTEMS
. MENU DRIVEN SOFTWARE

: : DEVELOPMENT
: MATERIAL DEVELOPMENT
COMPUTER SOFTWARE DEVELOPMENT
. PROGRAMING

: EVALUATION
COMPUTER SOFTWARE EVALUATION

: PUBLICATIONS
COMPUTER SOFTWARE REVIEWS

CONCEPTUAL SCHEMES (1967 1980)

: : : INDIVIDUAL CHARACTERISTICS
: : PSYCHOLOGICAL CHARACTERISTICS
: COGNITIVE STYLE
CONCEPTUAL TEMPO

: : ACTIVITIES
: MUSIC ACTIVITIES
CONCERTS

: : : : STANDARDS
: : : CRITERIA
: : EVALUATION CRITERIA
: VALIDITY
CONCURRENT VALIDITY

CONDITIONING
. BEHAVIOR MODIFICATION
. . CONTINGENCY MANAGEMENT
. . DESENSITIZATION
. CLASSICAL CONDITIONING
. OPERANT CONDITIONING
. . VERBAL OPERANT CONDITIONING

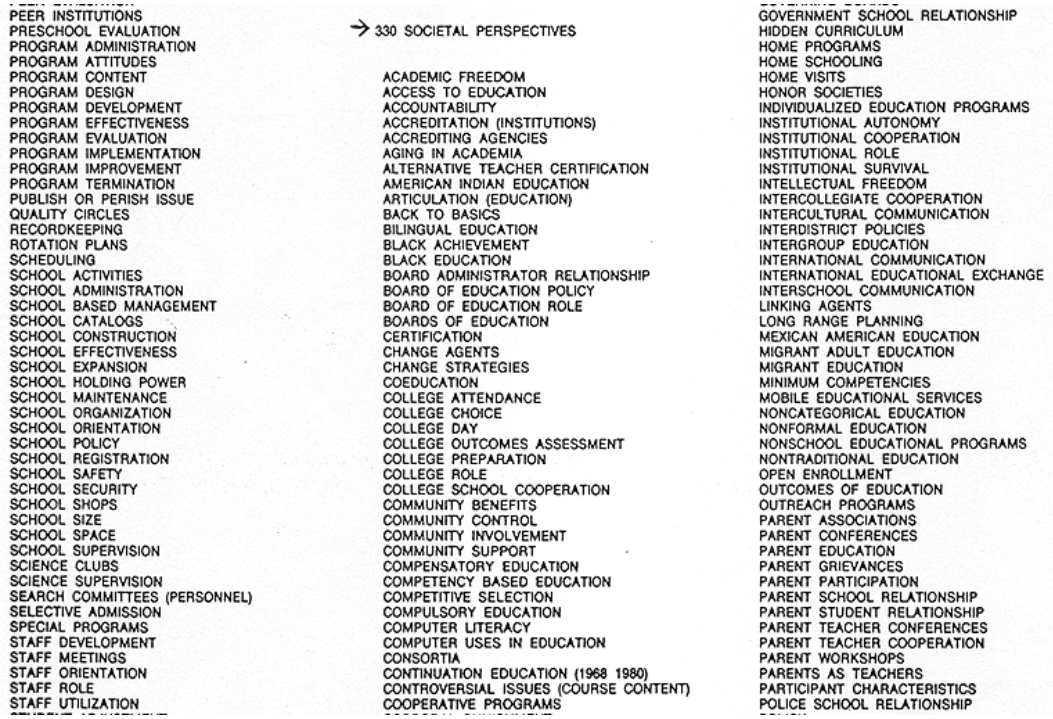
: : PUBLICATIONS
: REPORTS
CONFERENCE PAPERS

: : PUBLICATIONS
: SERIALS
CONFERENCE PROCEEDINGS

CONFERENCE REPORTS (1967 1980)

CONFERENCES
. PARENT CONFERENCES
. PARENT TEACHER CONFERENCES

รูปที่ 2.12 :แสดง ERIC Thesaurus – Hierarchies



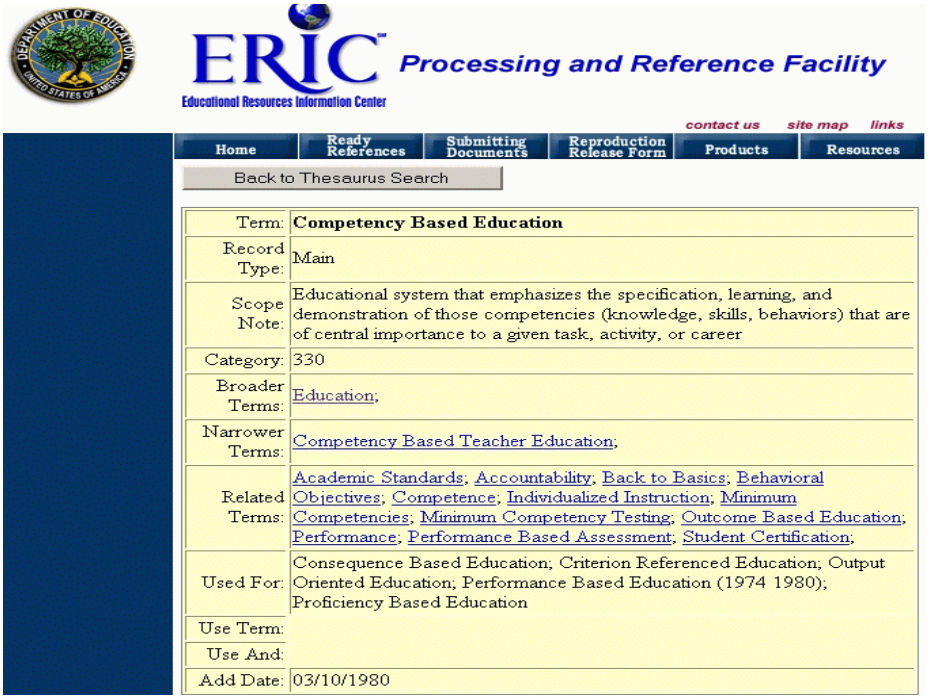
PEER INSTITUTIONS
PRESCHOOL EVALUATION
PROGRAM ADMINISTRATION
PROGRAM ATTITUDES
PROGRAM CONTENT
PROGRAM DESIGN
PROGRAM DEVELOPMENT
PROGRAM EFFECTIVENESS
PROGRAM EVALUATION
PROGRAM IMPLEMENTATION
PROGRAM IMPROVEMENT
PROGRAM TERMINATION
PUBLISH OR PERISH ISSUE
QUALITY CIRCLES
RECORDKEEPING
ROTATION PLANS
SCHEDULING
SCHOOL ACTIVITIES
SCHOOL ADMINISTRATION
SCHOOL BASED MANAGEMENT
SCHOOL CATALOGS
SCHOOL CONSTRUCTION
SCHOOL EFFECTIVENESS
SCHOOL EXPANSION
SCHOOL HOLDING POWER
SCHOOL MAINTENANCE
SCHOOL ORGANIZATION
SCHOOL ORIENTATION
SCHOOL POLICY
SCHOOL REGISTRATION
SCHOOL SAFETY
SCHOOL SECURITY
SCHOOL SHOPS
SCHOOL SIZE
SCHOOL SPACE
SCHOOL SUPERVISION
SCIENCE CLUBS
SCIENCE SUPERVISION
SEARCH COMMITTEES (PERSONNEL)
SELECTIVE ADMISSION
SPECIAL PROGRAMS
STAFF DEVELOPMENT
STAFF MEETINGS
STAFF ORIENTATION
STAFF ROLE
STAFF UTILIZATION

→ 330 SOCIETAL PERSPECTIVES

ACADEMIC FREEDOM
ACCESS TO EDUCATION
ACCOUNTABILITY
ACCREDITATION (INSTITUTIONS)
ACCREDITING AGENCIES
AGING IN ACADEMIA
ALTERNATIVE TEACHER CERTIFICATION
AMERICAN INDIAN EDUCATION
ARTICULATION (EDUCATION)
BACK TO BASICS
BILINGUAL EDUCATION
BLACK ACHIEVEMENT
BLACK EDUCATION
BOARD ADMINISTRATOR RELATIONSHIP
BOARD OF EDUCATION POLICY
BOARD OF EDUCATION ROLE
BOARDS OF EDUCATION
CERTIFICATION
CHANGE AGENTS
CHANGE STRATEGIES
COEDUCATION
COLLEGE ATTENDANCE
COLLEGE CHOICE
COLLEGE DAY
COLLEGE OUTCOMES ASSESSMENT
COLLEGE PREPARATION
COLLEGE ROLE
COLLEGE SCHOOL COOPERATION
COMMUNITY BENEFITS
COMMUNITY CONTROL
COMMUNITY INVOLVEMENT
COMMUNITY SUPPORT
COMPENSATORY EDUCATION
COMPETENCY BASED EDUCATION
COMPETITIVE SELECTION
COMPULSORY EDUCATION
COMPUTER LITERACY
COMPUTER USES IN EDUCATION
CONSORTIA
CONTINUATION EDUCATION (1968 1980)
CONTROVERSIAL ISSUES (COURSE CONTENT)
COOPERATIVE PROGRAMS

GOVERNMENT SCHOOL RELATIONSHIP
HIDDEN CURRICULUM
HOME PROGRAMS
HOME SCHOOLING
HOME VISITS
HONOR SOCIETIES
INDIVIDUALIZED EDUCATION PROGRAMS
INSTITUTIONAL AUTONOMY
INSTITUTIONAL COOPERATION
INSTITUTIONAL ROLE
INSTITUTIONAL SURVIVAL
INTELLECTUAL FREEDOM
INTERCOLLEGIATE COOPERATION
INTERCULTURAL COMMUNICATION
INTERDISTRICT POLICIES
INTERGROUP EDUCATION
INTERNATIONAL COMMUNICATION
INTERNATIONAL EDUCATIONAL EXCHANGE
INTERSCHOOL COMMUNICATION
LINKING AGENTS
LONG RANGE PLANNING
MEXICAN AMERICAN EDUCATION
MIGRANT ADULT EDUCATION
MIGRANT EDUCATION
MINIMUM COMPETENCIES
MOBILE EDUCATIONAL SERVICES
NONCATEGORICAL EDUCATION
NONFORMAL EDUCATION
NONSCHOOL EDUCATIONAL PROGRAMS
NONTRADITIONAL EDUCATION
OPEN ENROLLMENT
OUTCOMES OF EDUCATION
OUTREACH PROGRAMS
PARENT ASSOCIATIONS
PARENT CONFERENCES
PARENT EDUCATION
PARENT GRIEVANCES
PARENT PARTICIPATION
PARENT SCHOOL RELATIONSHIP
PARENT STUDENT RELATIONSHIP
PARENT TEACHER CONFERENCES
PARENT TEACHER COOPERATION
PARENT WORKSHOPS
PARENTS AS TEACHERS
PARTICIPANT CHARACTERISTICS
POLICE SCHOOL RELATIONSHIP

รูปที่ 2.13 :แสดง ERIC Thesaurus – Groups



รูปที่ 2.14 :แสดง ERIC Thesaurus – Online

ที่มา :http://www.ericfacility.net/extra/pub/thessearch.cfm

การสร้างบัญชีศัพท์สัมพันธ์ ได้ 2 วิธีคือ

(1) Manual Thesaurus ใช้หลักการทางความหมาย โดยตรวจดูคำพ้อง ความหมายโดยทั่วไป หรือเจาะจง ส่วนใหญ่ทำได้ 2 แบบคือ

1.1 การเชื่อมโยงคำที่เกี่ยวข้องกับหัวเรื่องเดียวกัน ให้อยู่ในกลุ่มเดียวกัน และใช้แทนกันได้ นำตัวแทนกลุ่มคำต่าง ๆเก็บในลิสต์ ผู้สร้างดรรชนีสามารถเลือกคำต่าง ๆจากลิสต์ไปสร้างเป็นตัวแทนเอกสาร หรือสร้างข้อคำถามในการดึงเอกสารที่ต้องการออกมาได้

1.2 เชื่อมโยงคำที่เกี่ยวข้องกับคำที่สัมพันธ์กัน เช่น จัดความสัมพันธ์เป็นลำดับขั้น

(2) อัตโนมัติ ใช้หลักการทางรูปแบบ(Syntactically) และสถิติ โดยใช้สมมุติฐานต่อไปนี้คือ

“ถ้าดรรชนี a และ b สามารถใช้แทนกันได้ หมายความว่า a และ b มีความหมายเหมือนกัน หรืออ้างอิงถึงเรื่องเดียวกัน หรือหัวข้อเรื่องเดียวกัน”

เรานำกลุ่มดรรชนีไปใช้ได้ 2 วิธีคือ

(1) โดยแทนที่แต่ละคีย์เวิร์ด ในตัวแทนเอกสาร หรือข้อคำถาม โดยชื่อของกลุ่มที่คีย์เวิร์ดหรือดรรชนีนั้นอยู่ ซึ่งช่วยพัฒนา recall ให้ดีขึ้น ซึ่งเพิ่มจำนวนของเอกสารที่ตรงกับข้อคำถาม(query)ของผู้ใช้

(2) โดยแทนที่แต่ละคีย์เวิร์ดด้วยทุกๆคีย์เวิร์ดที่เกิดขึ้นภายในกลุ่มเดียวกันกับคีย์เวิร์ดนั้น เป็นการเพิ่ม precision ช่วยให้เอกสารมีความเจาะจงตรงกับความต้องการของผู้ใช้

3. Phrase Dictionary

เป็นพจนานุกรมที่เก็บวลีของคำที่เกิดขึ้นบ่อยๆและแสดงเนื้อหาของเอกสารได้ ซึ่งให้ความหมายเจาะจงมากขึ้นกว่าการใช้ดรรชนีเดี่ยวๆ เช่น คำว่า “Programming Language “ วลี จะเป็นการรวบรวมคำ ตัวอย่างที่ใช้พจนานุกรมนี้คือ ระบบ Smart เป็นระบบค้นคืนสารสนเทศโดยอัตโนมัติ ได้กำหนดวิธีการในการค้นหาวลีที่สำคัญไว้ 2 วิธีคือ

(1) วิธีการค้นหาวลีที่สำคัญโดยอาศัยสถิติ (Statistical Phrase Detection) คำนึงเฉพาะคำที่เกิดขึ้นพร้อมกันบ่อยๆในเอกสาร ไม่สนใจรูปแบบของประโยค

(2) วิธีการค้นหาวลีที่สำคัญโดยอาศัยรูปแบบ (Syntactical Phrase Detection) คำนึงถึงลักษณะเฉพาะส่วนประกอบของวลี โดยเพิ่มข้อจัดในการประกอบกันของคำ โดยนิยามรูปแบบ เงื่อนไข ข้อจำกัด

65824 United States
61327 Article Type
33864 Los Angeles
18062 Hong Kong
17788 North Korea
17308 New York
15513 San Diego
15009 Orange County
12869 prime minister
12799 first time
12067 Soviet Union
10811 Russian Federation
9912 United Nations
8127 Southern California
7640 South Korea
7620 end recording
7524 European Union
7436 South Africa
7362 San Francisco
7086 news conference
6792 City Council
6348 Middle East
6157 peace process
5955 human rights
5837 White House

5778 long time
5776 Armed Forces
5636 Santa Ana
5619 Foreign Ministry
5527 Bosnia-Herzegovina
5458 words indistinct
5452 international community
5443 vice president
5247 Security Council
5098 North Korean
5023 Long Beach
4981 Central Committee
4872 economic development
4808 President Bush
4652 press conference
4602 first half
4565 second half
4495 nuclear weapons
4448 UN Security Council
4426 South Korean
4219 first quarter
4166 Los Angeles County
4107 State Duma
4085 State Council
3969 market economy
3941 World War II

รูปที่ 2.15 :แสดง Top Phrases from TIPSTER

4. Hierarchical Dictionary

เป็นพจนานุกรมที่มีรูปแบบคล้าย ๆกับ Tree และมีการแบ่งประเภทคำตามลำดับขั้นความสำคัญ

สรุปว่าการสร้างตัวแทนเอกสาร จะด้วยวิธีใดก็ตามต้องผ่านกระบวนการขจัดความซ้ำซ้อน(Normalization) ของข้อความในหลาย ๆระดับ โดยประกอบด้วยขั้นตอนดังนี้

(1) ขจัดเอาคำฟุ่มเฟือยและไม่มีความหมายพิเศษในการแสดงลักษณะเด่นของข้อความออกไป คำที่เหลือก็จะเป็นคำสำคัญ (Significant Words)

(2) นำเอาคำสำคัญต่าง ๆที่มีความหมายเดียวกันมาจัดรวมเป็นกลุ่ม และแต่ละกลุ่มจะมีคำสำคัญหนึ่งที่ใช้เป็นชื่อกลุ่ม และใช้ชื่อกลุ่มนี้เป็นดรรชนีหรือคีย์เวิร์ด

(3) การให้น้ำหนักดรรชนี ซึ่งอาจกระทำโดย

- ใช้ความถี่ของคำในเอกสารหรือใช้ความยาวของเอกสาร
- ใช้ความถี่ของคำในกลุ่มเอกสารทั้งหมด
- ใช้จำนวนเอกสารที่มีคำ ๆนั้นอยู่ในกลุ่มเอกสารทั้งหมด

ผลลัพธ์ที่เกิดขึ้นเป็นดรรชนี หรือตัวแทนเอกสารที่สามารถนำไปบันทึก และสามารถสืบค้นเอกสารต่าง ๆเหล่านี้ได้

ตัวอย่างที่ 2.6 แสดงการสร้างดัชนีหรือคำสำคัญ

- **Original text:**
  John Davenport, 52 years old, was appointed chief executive officer of this international telecommunications concern's U.S. subsidiary, Cable & Wireless North America Inc. Mr. Davenport, who succeeds John Zrno, is currently general manager of the group's operations in Bermuda.
- **One indexing result:**
  john davenport appoint chief executive officer international telecommunication concern subsidiary cable wireless north america davenport succeed john zrno current general manager group operation bermuda
- **Another possibility:**
  John_Davenport #person 52 years_old #age appoint chief_executive_officer international telecommunication concern #USA subsidiary Cable_&_Wireless_North_America #company Davenport #person succeed John_Zrno #person general_manager group operation Bermuda #foreigncountry

แบบฝึกหัด

1. ลุนซ์มีวิธีการวัดหาความสำคัญอย่างไร จงอธิบายพอเข้าใจ
2. จงอธิบายถึง Upper Cut-off และ Lower Cut-off พอเข้าใจ
3. ทำไมถึงต้องมีการขจัดคำที่มีความถี่เกิดขึ้นสูงและต่ำออกไป มีเหตุผลอย่างไรจงอธิบายพอเข้าใจ
4. ดรรชนีคืออะไร มีความสำคัญอย่างไรในระบบค้นคืนสารสนเทศ
5. จงอธิบายความแตกต่างระหว่า KWIC Index และ KWOC Index
6. ปัจจัยที่สำคัญที่ช่วยควบคุมประสิทธิภาพของดรรชนีมีอะไรบ้าง จงอธิบายพอเข้าใจ
7. ปัญหาในการวิเคราะห์ข้อความมีอะไรบ้าง
8. Thesaurus คืออะไร มีประโยชน์อย่างไรในระบบค้นคืนสารสนเทศ
9. จงอธิบายวิธีการในการให้น้ำหนักดรรชนี มีประโยชน์อย่างไรในระบบค้นคืนสารสนเทศ

บรรณานุกรม

สุมาลี เมืองไพศาล ,"การจัดการข้อมูล และการเรียกใช้ข้อมูล" ,สำนักพิมพ์มหาวิทยาลัยรามคำแหง ,CS337 ,2535

Art and Architecture Thesaurus ,
http://orange.sims.berkeley.edu/cgi-bin/flamenco/aa/Flamenco

ERIC Thesaurus , http://www.ericfacility.net/extra/pub/thessearch.cfm

Prof. Ray Larson & Prof. Marc Davis , "Information Organization and Retrieval : Lecture 22: Thesaurii and Metadata" , UC Berkeley SIMS , http://www.sims.berkeley.edu/academics/courses/is202/f04