บทที่ 8
การจัดทำดรรชนีอัตโนมัติ
(Automatic Indexing)

จากการศึกษามาจนถึงบทนี้ จะสรุปได้ว่า การจัดทำดรรชนีมี 2 ลักษณะ คือ ดรรชนีที่เกิดจากการกำหนดสำคัญขึ้น เพื่อใช้แทนเนื้อหาสาระของเอกสาร (Assignment indexing) และดรรชนีที่เกิดจากการดึงเอาคำสำคัญที่ปรากฏอยู่ในเนื้อหาของเอกสาร (Indexing by extraction) แต่ดรรชนีที่มีส่วนใหญ่เกิดจากมนุษย์เป็นผู้คัดเลือกคำสำคัญขึ้นเอง โดยใช้คู่มือควบคุมคำศัพท์เป็นเครื่องมือพิจารณาคัดเลือกคำที่เป็นมาตรฐาน แต่ดรรชนีที่เกิดจากการดึงเอาคำสำคัญในเอกสารมาเป็นศัพท์ดรรชนี ผู้จัดทำดรรชนีจะพยายามเลือกคำที่เป็นตัวบ่งชี้ที่ดีว่า เอกสารนั้นเป็นเรื่องเกี่ยวกับอะไร โดยพิจารณาว่า คำหรือวลีใดใช้มาก ใช้บ่อย และปรากฏอยู่ในส่วนสำคัญของเอกสาร อาทิ ชื่อเรื่อง บทสรุป หัวข้อเรื่อง หรือคำอธิบายภาพประกอบ และในส่วนที่เป็นเนื้อหาโดยตรง

8.1 แนวคิดและคำนิยาม

การจัดทำดรรชนีอัตโนมัติ (Automatic indexing) หมายถึง

1. การคัดเลือกคำสำคัญจากเอกสาร โดยวิธีการของคอมพิวเตอร์ เพื่อจัดทำเป็นรายการศัพท์ดรรชนี

2. การใช้คอมพิวเตอร์เพื่อดึงเอาคำสำคัญจากเอกสารออกมาและกำหนดให้เป็นศัพท์ดรรชนี โดยไม่ใช้มนุษย์เป็นผู้ช่วย หลังจากที่มนุษย์ได้ป้อนโปรแกรมคำสั่งต่าง ๆ ให้คอมพิวเตอร์รู้

การใช้คอมพิวเตอร์ช่วยจัดทำดรรชนีปรากฏครั้งแรก เมื่อประมาณเดือนพฤษภาคม ปี ค.ศ. 1958 โดย ฮานส์ ปีเตอร์ ลูน (Hans Peter Luhn) ชาวอเมริกันแห่งบริษัท IBM จัดทำดรรชนีโดยนำคำสำคัญจากชื่อเรื่องของเอกสารทางวิทยาศาสตร์มาหมุนเวียนเป็นศัพท์ดรรชนี เรียกว่า KWIC Index (Keywords – in – context) และได้ผลิตตัวอย่างดรรชนีออกมา โดยใช้โปรแกรม Routine I หนึ่งเดือนต่อมาในปีเดียวกัน ดรรชนีแบบ KWIC ได้ถูกจัดทำขึ้นสำหรับวรรณกรรมทางด้านการสืบค้นสารสนเทศและการแปลโดยเครื่องคอมพิวเตอร์ ซึ่งตลอดปี ค.ศ. 1958 – 1959 โปรแกรมเพื่อการจัดทำดรรชนีแบบ KWIC ได้รับการพัฒนาอย่างไม่หยุดยั้ง จนเป็นที่ยอมรับกันว่า เป็นทศวรรษ

160 IS 307

การจัดทำดรรชนีแบบ KWIC ได้รับการพัฒนาอย่างไม่หยุดยั้ง จนเป็นที่ยอมรับกันว่า เป็นทศวรรษ

แห่งความอัศจรรย์ และสิ่งสำคัญที่เกิดขึ้นในวงการสาขาวิชาเคมี คือ การประดิษฐ์หลอดแก้วทดลอง ผู้ผลิตดรรชนีได้ลดค่าใช้จ่ายการจัดทำดรรชนีลงทั้งในด้านการจ่ายให้แก่ผู้จัดทำดรรชนี บรรณาธิการดรรชนี ผู้พิมพ์ ผู้ตรวจสอบและผู้อ่าน การนำคอมพิวเตอร์เข้ามาช่วยจัดทำดรรชนีช่วยลดภาระค่าใช้จ่ายในการจัดพิมพ์ดรรชนี

ฮานส์ ปีเตอร์ ลูน เป็นบุคคลแรกที่มีบทบาทในการนำคอมพิวเตอร์มาใช้ในการจัดทำดรรชนี โดยการนำคำสำคัญจากชื่อเรื่องของบทความที่ปรากฏอยู่ในวารสาร Chemical Abstracts (วารสารสาระสังเขปสาขาวิชาเคมี) มาจัดทำเป็น Chemical Titles โดยจัดเรียงดรรชนีตามลำดับอักษร เพื่อช่วยในการค้นหาบทความ ต่อมาดรรชนีแบบ KWIC ได้รับความนิยมอย่างแพร่หลาย โปรแกรมง่าย ๆ ที่เขียนขึ้นเป็นการนับจำนวนความถี่ของคำที่ปรากฏในเนื้อหา โดยการกำหนดให้คอมพิวเตอร์รู้ว่า คำใดที่มีลักษณะเป็น "stop list" และจะต้องตัดออกไปไม่นำมาพิจารณาจัดเรียงตามลำดับอักษร ได้แก่ คำนำหน้านาม (article) คำบุพบท (prepositions) คำสันธาน (conjunctions) และคำสรรพนาม (pronouns) เป็นต้น คำใดที่มีความถี่ปรากฏในเอกสารบ่อย ๆ จะถูกเลือกมาจัดเป็นคำประเภท "top of the list" และจะถูกคัดเลือกเข้ามาเป็นศัพท์ดรรชนี (index terms) สำหรับเอกสารนั้น ๆ นอกจากคอมพิวเตอร์เลือกคำและวลีแล้ว คำสั่งโปรแกรมคอมพิวเตอร์สามารถเลือกคำที่มีรากศัพท์เดียวกัน โดยตัดคำลงท้ายคำ –ed หรือ –ing ออก

ลักษณะของดรรชนีแบบ KWIC ประกอบด้วย ส่วนสำคัญ 3 ส่วน คือ 1) Keyword ; 2) Context และ 3) Reference ดังรายละเอียดต่อไปนี้

1. คำสำคัญ (Keyword)

หมายถึง คำที่สำคัญทุกคำที่ปรากฏในชื่อเรื่องหรือสาระสังเขป คำสำคัญประเภทนี้เป็นคำอิสระ (free – text term) เป็นคำแบบไม่ควบคุมคำศัพท์ มีลักษณะเป็นคำธรรมชาติ (natural language) ตามที่ปรากฏเป็นภาษาเขียนที่ผู้แต่งใช้ในเอกสาร

การกำหนดว่า คำใดเป็นคำสำคัญนั้น มนุษย์จะเป็นผู้ป้อนโปรแกรมคำสั่งต่าง ๆ เข้าไปในเครื่องคอมพิวเตอร์ให้คอมพิวเตอร์รู้ คือ

– คำที่เป็น stop lists ได้แก่ คำนำหน้านาม (articles) คำบุพบท (preposition) คำสันธาน (conjunction) และคำสรรพนาม (pronoun) ซึ่งคอมพิวเตอร์จะไม่นำเอาคำเหล่านี้มาหมุนเวียนนำเป็นคำสำคัญ

– คำที่เป็น go lists ได้แก่ คำทุกคำที่ไม่ใช่คำ stop list นำมาหมุนเวียน จัดทำเป็นคำสำคัญได้

2. คำซึ่งเป็นบริบท (Context)

หมายถึง คำซึ่งอยู่ส่วนต้นหรือส่วนที่ต่อจากคำสำคัญ (Keyword) เป็นคำซึ่งอยู่แวดล้อมรอบคำสำคัญนั่นเอง ซึ่งไม่ได้ถูกนำมากำหนดให้เป็นคำสำคัญในขณะนั้น คำสำคัญดังกล่าวจะช่วยให้ความหมายของคำเด่นชัดขึ้น

3. ส่วนอ้างอิง (Reference)

หมายถึง ส่วนที่บ่งชี้บอกตำแหน่งที่ปรากฏสารสนเทศ โดยทำหน้าที่ชี้โยงจากคำสำคัญ ซึ่งถูกกำหนดเป็น index term ไปยังแหล่งที่ปรากฏสารสนเทศ ได้แก่ เลขลำดับที่ของบทความ หมายเลข / รหัสของเอกสาร ซึ่งจะให้ข้อมูลบรรณานุกรมของเอกสาร

โดยทั่วไปดรรชนีแบบ KWIC จะจัดพิมพ์ไว้เป็น 3 คอลัมน์ในบรรทัดเดียวกัน ข้อมูลในแต่ละบรรทัดจัดเรียงตามลำดับอักษรของคำสำคัญ (Keyword) ซึ่งอยู่ตรงกลางระหว่างบริบท (Context) ซึ่งอยู่ทางคอลัมน์ซ้ายและคอลัมน์ขวา ส่วนคอลัมน์ขวาสุดเป็นส่วนอ้างอิง (Reference) ที่จะเชื่อมโยงไปสู่ข้อมูลบรรณานุกรมของเอกสารที่ต้องการค้น ตัวอย่าง เช่น

CONTEXT	**KEYWORD**	CONTEXT	REFERENCE
SUGAR AND SUGAR	**ALCOHOLS**	METHOD FOR ANALYSIS OF	89
ALCOHOLS / METHODS FOR	**ANALYSIS**	OF SUGAR AND SUGAR	89
SUGAR AND SUGAR ALCOHOLS	**METHODS**	FOR ANALYSIS OF	89
ANALYSIS OF SUGARS AND	**SUGARS**	ALCOHOLS / METHOD FOR	89
METHOD FOR ANALYSIS OF	**SUGARS**	AND SUGAR ALCOHOLS	89

ดรรชนีแบบ KWIC เป็นที่ยอมรับกันและใช้กันอย่างแพร่หลาย ง่ายต่อการจัดทำและเสียค่าใช้จ่ายในการจัดทำถูก สะดวกในการใช้และเป็นตัวช่วยในการค้นหาสารสนเทศได้ดี แต่จะต้องอาศัยทักษะและความชำนาญจากนักเอกสารสนเทศหรือผู้เชี่ยวชาญเฉพาะสาขาวิชาช่วย จึงจะทำให้การค้นหาสารสนเทศสะดวกยิ่งขึ้น

การจัดทำดรรชนีแบบ KWIC นี้ค่อนข้างจะมีปัญหาด้านการจัดพิมพ์รายการ จึงมีการดัดแปลงรูปแบบการจัดพิมพ์ โดยให้คำสำคัญ (keyword) จัดวางอยู่ในตำแหน่งเหมือนหัวเรื่อง

(heading) ซึ่งอยู่ทางด้านซ้ายมือสุดของรายการ ส่วนคำซึ่งเป็นบริบทและส่วนอ้างอิงจะอยู่ภายใต้คำสำคัญ ดรรชนีรูปแบบที่ดัดแปลงใหม่นี้ เรียกว่า ดรรชนีแบบ KWOC (Keyword – out of – Context) จัดให้คำสำคัญดึงออกมาอยู่ภายนอกบริบท และดรรชนีแบบ KWOC ที่แท้จริงจะใช้เครื่องหมายดอกจัน (asterisk) แทนตำแหน่งคำสำคัญ ซึ่งดึงออกไปเป็นหัวเรื่อง

ตัวอย่าง เช่น

ANALYSIS	
METHOD FOR * OF SUGARS AND SUGAR ALCOHOLS	89
METHOD	
* FOR ANALYSIS OF SUGARS AND SUGAR ALCOHOLS	89
SUGARS	
METHOD FOR ANALYSIS OF * AND SUGAR ALCOHOLS	89

จุดประสงค์ในด้านการใช้คอมพิวเตอร์เพื่อจัดทำดรรชนีถือว่า คำศัพท์ที่ปรากฏใช้ในเอกสารจัดเป็นการจัดลำดับความต่อเนื่องของสัญลักษณ์ ตัวอักษร จำนวน และเครื่องหมาย ดังนั้นคอมพิวเตอร์สามารถจัดคำในรูปแบบต่าง ๆ ได้ สามารถนับจำนวนคำในเอกสารและคำนวณหาความหมายจำนวนของคำโดยเฉลี่ยในแต่ละประโยค และนับความถี่ของการใช้คำในเอกสาร

วิธีพื้นฐานของการจัดทำดรรชนีอัตโนมัติมี 2 วิธี คือ

1. **การวิเคราะห์เชิงสถิติของการใช้คำ (Statistical analysis of text)**

วิธีการเช่นนี้ยึดอยู่บนสมมติฐานว่า ความถี่ / จำนวนครั้งที่คำนั้นปรากฏใช้ในเอกสารมากเท่าไร หมายความว่า คำนั้นจะมีความสำคัญ เป็นตัวบ่งชี้ถึงเนื้อหาสาระของเอกสารได้มากยิ่งขึ้น คอมพิวเตอร์จะจัดเรียงคำเป็นกลุ่มและจัดเรียงตามลำดับอักษร โดยไม่นับคำ stop lists คำใดที่มาจากรากศัพท์เดียวกัน จะถือเป็นคำเดียวกัน ลูน (Luhn) จึงเป็นบุคคลแรกที่เริ่มใช้ความถี่ของคำเป็นตัวกำหนดความสำคัญของคำ ลูนกล่าวว่า วิธีเช่นนี้เป็นวิธีที่ง่ายที่สุด ไม่ซับซ้อน และไม่หวือหวาวิเศษโดดเด่นอะไร คอมพิวเตอร์ทำหน้าที่คัดเลือกคำที่มีความถี่การใช้สูง มีการใช้บ่อย ๆ เช่น กำหนดเกณฑ์ไว้ว่า หากมีการใช้ซ้ำกัน 10 – 12 ครั้ง หัวเรื่องที่มีประโยชน์อาจถูกตัดออกไป ดังนั้นจึงตั้งค่าทั่วไป ความถี่ลดลงเป็น 3 – 4 ครั้งก็ได้ แต่อาจทำให้เนื้อหาที่ไม่ใช่ประโยชน์เข้ามาด้วย

การประเมินแบบง่าย ๆ อีกแนวทางหนึ่ง คือ การพิจารณาดูความถี่ของการใช้คำจากส่วนต่าง ๆ ของเอกสาร ตามการพิจารณาความสำคัญของแต่ละส่วน โดยให้น้ำหนัก (weighting) ความสำคัญของ keyword ดังต่อไปนี้

1. คำที่ปรากฏในชื่อเรื่อง จะมีน้ำหนักของคำมากกว่าคำที่ปรากฏในเนื้อหา
2. คำที่มีความถี่ของความสัมพันธ์กัน ทั้งนี้ขึ้นอยู่กับความสัมพันธ์ระหว่างจำนวนของคำที่ปรากฏในเอกสารเปรียบเทียบกับคำเดียวกันที่ปรากฏความถี่ของการใช้คำในเอกสารอื่น ๆ
3. การใช้นามวลี โดยคัดเลือกจากชื่อเรื่องและสาระสังเขป
4. การใช้ศัพท์สัมพันธ์
5. การใช้ปัจจัยที่มีความเกี่ยวข้องกัน โดยดูระดับความสัมพันธ์ของคำในเอกสารเดียวกัน

การใช้เทคนิคใดก็ตามขึ้นอยู่กับปัจจัยต่าง ๆ ได้แก่ ประเภทของเอกสาร ลักษณะการวิจัย เป็นต้น

2. การวิเคราะห์เชิงโครงสร้างของคำและความหมายของคำ (Syntactic and Semantic analysis)

วิธีการนี้ได้รับการพัฒนาขึ้น เมื่อประมาณกลางทศวรรษ 1960 เกิดขึ้นภายหลังวิธีแบบการวิเคราะห์เชิงสถิติของการใช้คำ

– การวิเคราะห์เชิงโครงสร้างของคำ (Syntactic analysis) เป็นวิธีการทางภาษาศาสตร์ หมายถึง การพิจารณาบทบาทของคำในประโยค ดูระดับของคำเชิงไวยากรณ์ และความสัมพันธ์ระหว่างคำในประโยค และโครงสร้างของคำที่มีความหมายสัมพันธ์กับคำอื่น ๆ

– การวิเคราะห์เชิงความหมายของคำ (Semantic analysis) หมายถึง การวิเคราะห์ความสัมพันธ์ระหว่างคำ โดยแนวคิดที่แสดงออกถึงเนื้อหาและคำที่แสดงไว้ในเนื้อหาของเอกสาร คำแต่ละคำจะมีความหมายสมบูรณ์ในตัวของคำนั้น

การจัดทำดรรชนีอัตโนมัติโดยใช้วิธีการวิเคราะห์เชิงสถิติของการใช้คำที่เป็นที่นิยมใช้มากกว่าวิธีการวิเคราะห์เชิงภาษาศาสตร์ เนื่องจากเป็นวิธีที่ง่ายกว่า โดยใช้วิธีการดึงคำมาจากเอกสารโดยตรง ไม่ต้องอาศัยสติปัญญาของมนุษย์ในการวิเคราะห์เนื้อหาสาระของเอกสาร ไม่ต้องใช้เครื่องมือควบคุมคำศัพท์ควบคุมคำสำคัญให้เป็นคำมาตรฐาน แต่จะใช้คำที่ปรากฏในเอกสารโดยตรง และจะต้องได้รับการตรวจสอบแล้วว่า เป็นคำที่มีความถี่ในการใช้สูง (top of the lists) ซึ่งถือว่า

คำเหล่านี้สามารถสื่อความหมายของเรื่องได้ ยกเว้นคำที่เป็น stop lists เทคนิคนี้ได้พัฒนาขึ้นใช้และปรากฏว่า เป็นเทคนิคที่มีประสิทธิภาพมาก

การจัดทำดรรชนีอัตโนมัติจำแนกได้ 2 ลักษณะ คือ

1. การจัดทำดรรชนีโดยการกำหนดคำแบบอัตโนมัติ (Automatic Assignment Indexing)
2. การจัดทำดรรชนีโดยการดึงคำออกมาอัตโนมัติ (Automatic Extraction Indexing)

การจัดทำดรรชนีโดยการกำหนดคำแบบอัตโนมัติ (Automatic Assignment Indexing)

การจัดทำดรรชนีที่จัดทำโดยมนุษย์ โดยทั่วไปใช้วิธีการกำหนดสำคัญจัดเป็นระบบมือ (manual system) ซึ่งเป็นวิธีที่ค่อนข้างยากและซับซ้อนอยู่แล้ว เมื่อนำเอาคอมพิวเตอร์มาช่วยดูเหมือนจะเป็นงานที่ยากยิ่งขึ้น วิธีการที่นำมาใช้ คือ การกำหนด "โครงร่าง (profile)" ของคำหรือวลีที่มีแนวโน้มว่า เป็นคำหรือวลีที่มีปรากฏในเอกสารบ่อยมาก มีความถี่การใช้สูงที่ผู้จัดทำดรรชนีควรจะกำหนดให้เป็น index term เช่น ในโครงร่างของวลีว่า "acid rain = ฝนกรด" ถ้าดูในบัญชีหัวเรื่องจะปรากฏดังนี้

Acid rain	
BT	Atmosphere
	Air pollution
RT	Acid precipitation (Meteorology)
NT	Sulfur dioxide
	Sulfur acid

ถ้าคำทุกคำข้างต้นเป็นคำศัพท์ควบคุม มีโครงร่างของคำสัมพันธ์กัน สามารถใช้โปรแกรมคอมพิวเตอร์จับกลุ่มคำ / วลีที่มีในเอกสารเข้าไว้ด้วยกัน วิธีนี้ไม่ใช่เรื่องง่ายต้องอาศัยคำสั่งโปรแกรมคอมพิวเตอร์ที่ดี

IS 307 165

ถ้าวลี "acid rain" ปรากฏในบทความมากกว่า 10 ครั้ง เช่นนี้ถือว่า วลี "acid rain" สามารถใช้เป็นศัพท์ดรรชนีได้ แต่ถ้ามีวลีอื่น ๆ ปรากฏอยู่ด้วย แต่ไม่บ่อยนัก ได้แก่ atmosphere, air pollution, sulfur dioxide, sulfur acid ฯลฯ วลีเหล่านี้อาจนำมากำหนดเป็นศัพท์ดรรชนีได้เช่นกัน

ผู้จัดทำดรรชนีจะประสบปัญหาการใช้คอมพิวเตอร์ช่วยกำหนดคำสำคัญ เพราะคอมพิวเตอร์ไม่สามารถคาดคะเนได้ว่า ควรจะกำหนดคำใดจึงจะเหมาะสมกับเนื้อหาสาระของเอกสาร แม้ว่าโปรแกรมคอมพิวเตอร์จะคาดคะเนคำสำคัญได้ก็ตาม จากปัญหาดังกล่าวการจัดทำดรรชนีโดยการกำหนดคำแบบอัตโนมัติไม่ใคร่จะประสบความสำเร็จมากนัก หากเปรียบเทียบกับการจัดทำดรรชนีด้วยระบบมือโดยมนุษย์เป็นผู้จัดทำ แต่โดยแท้จริงแล้ว การจัดทำดรรชนีทั้งโดยระบบมือและระบบอัตโนมัติมีสมรรถนะในระดับใกล้เคียงกัน บางครั้งอาจกำหนดคำสำคัญน้อยเกินไป (underassignment) และอาจจะกำหนดคำมากเกินไปหรือละเอียดเกินไป (overassignment) มีจำนวนคำสำคัญที่ไม่ต้องการ / หรือไม่มีประโยชน์มากกว่าคำสำคัญที่ต้องการและจำเป็นสำหรับการสืบค้นสารสนเทศ

แม้จะมีความพยายามปรับปรุงวิธีการจัดทำดรรชนี โดยการกำหนดคำแบบอัตโนมัติมาเป็นระยะเกือบ 4 ทศวรรษแล้วก็ตาม การจัดทำดรรชนีโดยวิธีนี้ยังไม่ถึงขั้นที่น่าพึงพอใจเท่าที่ควร หากปราศจากการควบคุมการทำงานโดยคน แต่ปัจจุบันความสนใจในการจัดทำดรรชนีโดยการกำหนดคำแบบอัตโนมัติมีเพียงไม่มาก เว้นเสียแต่เป็นการจัดทำดรรชนีสำหรับเอกสารสิ่งพิมพ์และดรรชนีหนังสือ

การจัดทำดรรชนีโดยการดึงคำออกมาอัตโนมัติ (Automatic Extraction Indexing)

คอมพิวเตอร์ทำหน้าที่ดึงคำหรือวลีออกมาจากเอกสาร เป็นวิธีการจัดทำดรรชนีที่สะดวกและรวดเร็วกว่าการจัดทำดรรชนีโดยมนุษย์ โดยดึงคำตามที่มนุษย์ได้ป้อนคำสั่งไว้ในคอมพิวเตอร์ โดยดึงคำที่เป็น go lists ทุกคำ (ยกเว้น stop lists) เป็นศัพท์ดรรชนี

ในทางปฏิบัติ โดยแท้จริงระบบการจัดทำดรรชนีอัตโนมัติไม่มีระบบใดเป็นแบบอัตโนมัติอย่างแท้จริง เพราะทุกระบบจะต้องมีมนุษย์เป็นผู้กำหนดโปรแกรมคำสั่ง จึงทำให้รู้สึกว่าคอมพิวเตอร์ทำหน้าที่ช่วยให้มนุษย์มีความคล่องตัว สะดวกในการจัดทำดรรชนีมากยิ่งขึ้น โดยทำหน้าที่ ดังนี้

1. คอมพิวเตอร์ช่วยผู้จัดทำดรรชนีในการแสดงสารสนเทศบนเครือข่ายออนไลน์ ข้อผิดพลาดใด ๆ ของดรรชนีออนไลน์อาจทราบได้ทันทีและผู้จัดทำดรรชนีสามารถแก้ไขข้อผิดพลาดได้โดยเร็ว ได้แก่ การใช้คำศัพท์ที่ไม่ได้มาตรฐาน การใช้คำประสมของหัวเรื่องหลักหรือหัวเรื่องย่อยตามวิธีการของตรรกบูลีน (Boolean logic) ไม่ถูกต้อง

2. เราใช้โปรแกรมคอมพิวเตอร์ เพื่อดึงคำสำคัญจากเนื้อหาในส่วนที่เป็นชื่อเรื่องและ / หรือสาระสังเขปมาเป็นศัพท์ดรรชนี โดยคำ / วลีที่คอมพิวเตอร์ดึงออกมานี้อาจได้รับการตรวจสอบโดยผู้จัดทำดรรชนีและอาจเพิ่มคำสำคัญขึ้นใหม่ซึ่งเป็นคำศัพท์ที่คอมพิวเตอร์ไม่สามารถกำหนดขึ้นมาได้

ปัจจุบันการจัดทำดรรชนีอัตโนมัติมีความจำเป็นสำหรับเอกสาร / ข้อมูลฉบับเต็ม เพื่อเป็นเครื่องมือช่วยจัดเก็บ ค้นหาและการเข้าถึงสารสนเทศที่ต้องการได้อย่างรวดเร็ว

8.2 การใช้ศัพท์สัมพันธ์เพื่อการจัดทำดรรชนี

ปัจจุบันหน่วยงานที่ให้บริการสารสนเทศเฉพาะสาขาวิชาได้ใช้ "ศัพท์สัมพันธ์" (Thesaurus) เป็นเครื่องมือสำคัญช่วยการดำเนินงานการจัดเก็บและการค้นหาสารสนเทศได้เป็นจำนวนมากและในเวลาอันรวดเร็ว

8.2.1 แนวคิดและคำนิยาม

คำว่า "thesaurus" อ่านออกเสียงทับศัพท์ว่า "ธิซอรัส" มีรากศัพท์มาจากภาษากรีกว่า "thesauros" หมายถึง ขุมทรัพย์ หรือคลังแห่งความรู้ เช่นเดียวกับพจนานุกรม สารานุกรม ซึ่งหมายถึง แหล่งรวมคำ วลี หรือข้อความต่าง ๆ ที่นำมาจากวรรณกรรมต่าง ๆ เป็นหนังสือที่รวบรวมคำที่มีความหมายใกล้เคียงกันมารวมเป็นชุดคำเดียวกัน จึงเรียกว่า "ศัพท์สัมพันธ์"

ศัพท์สัมพันธ์ (Thesaurus) เป็นการรวมคำ หรือกลุ่มของคำ ซึ่งประกอบด้วย คำเชื่อมระหว่างคำ ซึ่งใช้ในเอกสารกับคำ ซึ่งเป็นคำสำคัญ (descriptors) เพื่อใช้สำหรับช่วยการสืบค้นสารสนเทศ โดยการแสดงความสัมพันธ์ของคำศัพท์ในเชิงความหมายของคำ ศัพท์สัมพันธ์จำนวนมากแสดงคำที่ใช้ในสาขาวิชาต่าง ๆ เพื่อประโยชน์ในการจัดทำดรรชนี ดังนั้นผู้จัดทำดรรชนีมักใช้ศัพท์สัมพันธ์เพื่อเป็นเครื่องมือควบคุมคำศัพท์ จากคำศัพท์ซึ่งเป็นภาษาธรรมชาติที่ใช้ในเอกสารให้เป็น

คำศัพท์ที่ใช้ในระบบภาษาดรรชนี ถ้าพิจารณาในรูปของโครงสร้างของศัพท์สัมพันธ์จะเห็นว่า ศัพท์สัมพันธ์เป็นคำศัพท์ควบคุมที่มีการหมุนเวียน รวบรวมคำศัพท์ที่มีความหมายเช่นเดียวกัน หรือมีความสัมพันธ์กันในเชิงกว้าง แต่ครอบคลุมองค์ความรู้เฉพาะด้านด้วย

ศัพท์สัมพันธ์จึงเป็นรายการคำศัพท์ที่รวบรวมมาจากเอกสาร หรือเป็นคำที่มีการบัญญัติขึ้นใช้ มีลักษณะเป็นคำศัพท์ควบคุม กำหนดให้คำหนึ่งมีหน้าที่ควบคุมหรือใช้แทนคำหลายคำที่มีความหมายเหมือนกัน โดยมีจุดมุ่งหมายให้มีการใช้คำที่เป็นมาตรฐาน และช่วยค้นคืนสารสนเทศที่ต้องการได้อย่างรวดเร็ว ศัพท์สัมพันธ์มีลักษณะเฉพาะ คือ ไม่มีการอธิบายความหมายของคำแบบพจนานุกรม แต่จะแสดงความสัมพันธ์ของคำในรูปของความสัมพันธ์ของศัพท์ที่มีความหมายเชิงกว้างกับศัพท์ที่มีความหมายเชิงแคบ โดยจำแนกคำศัพท์ตามลำดับชั้นของความสัมพันธ์

การสร้างศัพท์สัมพันธ์มีวัตถุประสงค์ เพื่อให้บริการสารสนเทศเฉพาะสาขาวิชาให้มีความสะดวกและรวดเร็ว โดยการใช้คอมพิวเตอร์เป็นเครื่องช่วยดำเนินการรวบรวมและจัดความสัมพันธ์ของคำศัพท์ องค์การที่ให้สารสนเทศส่วนใหญ่จำเป็นต้องใช้ศัพท์สัมพันธ์ในการกำหนดคำสำคัญ หรือหัวเรื่องสำหรับบันทึกในรายการข้อมูลทางบรรณานุกรมเก็บไว้ในฐานข้อมูลและการจัดเก็บแฟ้มเอกสารได้อย่างมีประสิทธิภาพ เพราะรายการคำศัพท์เพียงอย่างเดียว สามารถสื่อให้ทราบรายละเอียดของเรื่องทุกเรื่องที่เกี่ยวข้องกันได้อย่างชัดเจน

ในการจัดทำดรรชนีและการค้นคืนสารสนเทศโดยวิธีอัตโนมัติ ผู้จัดทำดรรชนีและนักเอกสารสนเทศจะใช้ประโยชน์จากศัพท์สัมพันธ์ เพื่อให้การปฏิบัติงานมีประสิทธิภาพได้ 2 ประการ

1. สามารถกำหนดคำสำคัญ / ศัพท์ดรรชนีที่ใช้แทนเนื้อหาสาระของเอกสารได้อย่างคงที่ คงเส้นคงวา มีมาตรฐานของการใช้คำ ถ้าผู้จัดทำดรรชนีแต่ละคนเลือกใช้คำหรือกำหนดคำโดยอิสระ ย่อมส่งผลให้การสืบค้นสารสนเทศไม่มีประสิทธิผลเท่าที่ควร

2. สามารถสืบค้นสารสนเทศเรื่องใดเรื่องหนึ่งที่เกี่ยวข้องกันได้อย่างครบถ้วน โดยที่ผู้ใช้ไม่จำเป็นต้องรู้จักคำศัพท์ได้เองทั้งหมด

วิธีการดังกล่าวทั้งสองประการนี้ เป็นวิธีการจัดทำดรรชนีแบบ post – coordinate หรือการจัดทำดรรชนีอัตโนมัติ

ดังนั้นศัพท์สัมพันธ์มีบทบาทสำคัญ 2 ขั้นตอน คือ

1. **ขั้นตอนการจัดเก็บ / จัดแฟ้มของเอกสาร**

ขั้นตอนนี้ "ศัพท์สัมพันธ์" เป็นคู่มือของผู้จัดทำดรรชนีในระบบภาษาควบคุม กล่าวคือ ผู้จัดทำดรรชนีจะวิเคราะห์เนื้อหาของเอกสาร แล้วกำหนดแนวคิดซึ่งเป็นประเด็นสำคัญของเอกสาร โดยอาศัยศัพท์สัมพันธ์เป็นคู่มือในการคัดเลือกคำและควบคุมความคงที่ของการใช้คำทุกครั้งที่กำหนดคำแทนเนื้อหาสาระให้กับเอกสารที่มีเนื้อหาเดียวกัน เพื่อให้เกิดประสิทธิภาพและประสิทธิผลในการค้นคืนต่อไป

2. **ขั้นตอนการค้นคืนสารสนเทศ**

ผู้ใช้บริการสารสนเทศ จะใช้ศัพท์สัมพันธ์เป็นคู่มือค้นหาคำที่มีความหมายของเรื่องที่ต้องการสืบค้น โดยมีกระบวนการสืบค้น ซึ่งใช้หลักการเดียวกันกับขั้นตอนการจัดแฟ้มของเอกสาร

8.2.2 องค์ประกอบของศัพท์สัมพันธ์

ศัพท์สัมพันธ์มีลักษณะเป็นคำศัพท์ควบคุม ประกอบด้วย คำศัพท์เป็นชุด ๆ แต่ละชุดมีส่วนประกอบทั่วไป ดังนี้

1. คำหลักของชุดเรียกว่า Descriptor
2. คำที่มีความหมายพ้องกัน เป็นคำเหมือน หรือมีความหมายใกล้เคียงกับคำหลักนำไปใช้ในระบบ เพื่อแทนคำหลักได้ จะใช้การโยงรายการว่า "USE"
3. คำที่มีความหมายพ้องกัน หรือมีความหมายใกล้เคียงกับคำหลัก แต่ระบบไม่ใช้ และกำหนดให้ใช้คำหลักแทน จะใช้การโยงรายการว่า "Used for"
4. ถ้าคำหลักมีความหมายไม่กระจ่างชัด จำเป็นต้องมีคำอธิบายคำหลัก (descriptor) ให้ชัดเจน ส่วนคำอธิบายนี้ เรียกว่า ข้อความอธิบายคำหลัก (scope note)
5. คำที่มีความสัมพันธ์กับคำหลัก ได้แก่
   1) คำที่เป็นต้นสกุลของคำหลัก (TT = Top term)
   2) คำที่มีความหมายกว้างกว่าคำหลัก (BT = Broader term)
   3) คำที่มีความหมายแคบกว่าคำหลัก (NT = Narrower term)
   4) คำที่มีความหมายเกี่ยวข้องกับคำหลัก (RT = Related term)

ในบัญชีศัพท์สัมพันธ์มีสัญลักษณ์ที่ใช้ต่อไปนี้

SN	=	Scope note	ใช้นำหน้าข้อความอธิบายคำศัพท์
U	=	Use	ใช้นำหน้าคำศัพท์ที่กำหนดให้ใช้
UF	=	Used for	ใช้นำหน้าคำศัพท์ที่ไม่ใช้
TT	=	Top term	ใช้นำหน้าคำศัพท์ต้นสกุล หรือคำศัพท์ที่กว้างที่สุด หรือคำรวม
BT	=	Broader term	ใช้นำหน้าคำศัพท์ที่มีความหมายกว้างกว่าคำหลัก
NT	=	Narrower term	ใช้นำหน้าคำศัพท์ที่มีความหมายแคบกว่าคำหลัก
RT	=	Related term	ใช้นำหน้าคำศัพท์ที่มีความหมายเกี่ยวข้องกับคำหลัก

ตัวอย่าง

Abortion counseling

SN: Professional advice and assistant to individuals considering induced abortion ; includes counseling both before and after the abortion
TT: Guidance
BT: Counseling
RT: Contraception
Contraceptive devices
Family planning behavior

ABSTRACTING & INDEXING SERVICES

SN: The preparation and dissemination of indexes and abstracts of currently published documents
UF: Indexing services
RT: Abstracting
Abstracts
Abstracts & indexing services

INFORMATION SERVICE

TT:	Services
UF:	Bibliographical services
	Documentation centres
	Information centers
	Information centres
	Library information services
NT:	Abstracting & indexing services
	Data banks
	Libraries
	Media services
	Translation services
RT:	Information
	Information dissemination
	Information exchange
	Information science
	Information systems
	Statistical services
	University & college libraries

ในการจัดทำดรรชนีโดยใช้ศัพท์สัมพันธ์ ผู้จัดทำดรรชนีไม่จำเป็นต้องทราบรายละเอียดวิธีการสร้างศัพท์สัมพันธ์ แต่สิ่งสำคัญที่ควรกระทำ คือ การอ่านเนื้อหาสาระในเอกสารที่ต้องการจัดทำดรรชนีให้เข้าใจและสามารถดึงคำ ซึ่งเป็นคำสำคัญในเนื้อหาของเอกสารมาจัดทำดรรชนีได้อย่างครบถ้วน หากใช้คอมพิวเตอร์ช่วยจัดทำดรรชนี นั่นคือ การดึงคำทุกคำที่เป็น go lists ออกมา หมายความว่า คำสำคัญทุกคำนั้นมีลักษณะเป็นศัพท์อิสระ แต่เป็นคำศัพท์ที่มีการควบคุม เพื่อให้ดรรชนีมีมาตรฐาน ถ้าผู้ใช้บริการค้นคืนสารสนเทศออกมาได้ แม้ใช้คำศัพท์ไม่ตรงกับที่คำศัพท์ควบคุม แต่ระบบคอมพิวเตอร์สามารถช่วยแสดงศัพท์สัมพันธ์ โดยโยงคำศัพท์ที่ไม่ใช้ไปยังคำศัพท์ควบคุมที่บังคับใช้ได้ และสามารถดึงคำศัพท์อื่น ๆ ซึ่งไม่ใช่คำศัพท์ควบคุมออกมา เพื่อเป็นสื่อช่วยการค้นคืนและการเข้าถึงสารสนเทศได้

IS 307 171