บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 เทคนิคการค้นหาข้อมูลบนฐานข้อมูลเชิงสัมพันธ์

การค้นคืนสารสนเทศเป็นเทคโนโลยีที่ใช้ในการเพิ่มประสิทธิภาพการทำงานของระบบค้นหาข้อมูล และจัดลำดับความสำคัญหรือความน่าจะเป็นของข้อมูลที่ผู้ใช้ต้องการตามค่าน้ำหนักที่คำนวณได้จากมากไปหาน้อย โดยเดิมทีการค้นคืนสารสนเทศจะใช้ได้กับเอกสารประเภทข้อความ (Text) เท่านั้น แต่ปัจจุบันนี้ ถ้ากล่าวถึงการค้นคืนสารสนเทศอาจจะหมายความร่วมไปถึงเอกสารที่อยู่บนเครือข่ายอินเทอร์เน็ต หรือเอกสารประเภทอื่นๆ ด้วย เช่น รูปภาพ เสียง วิดีโอ เป็นต้น

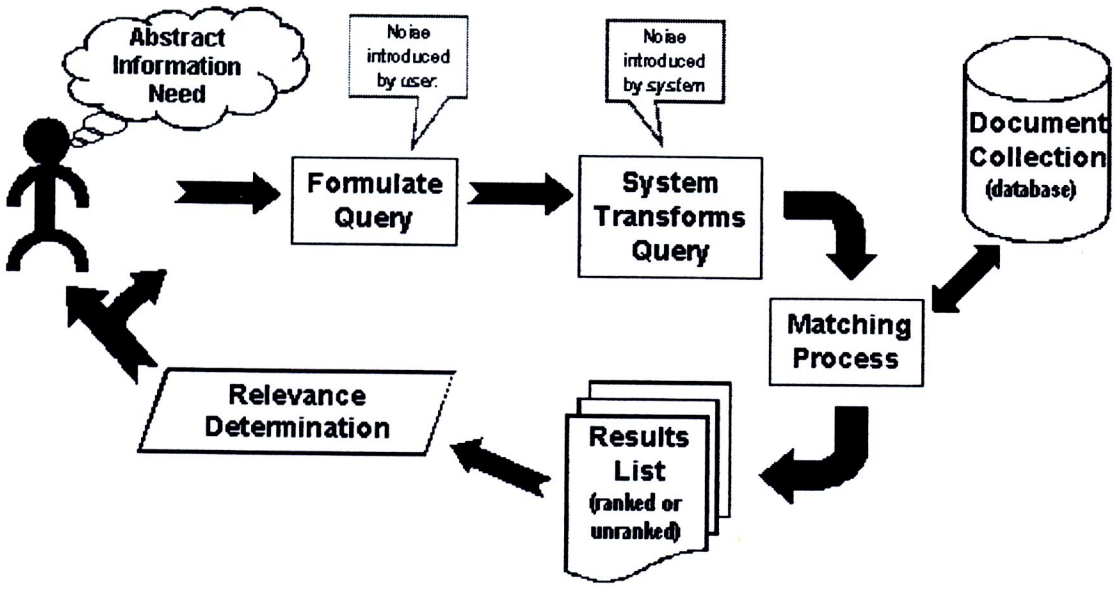
ในปัจจุบันเทคนิคการค้นหาข้อมูลบนฐานข้อมูลเชิงสัมพันธ์โดยเรียงลำดับความสำคัญของข้อมูลจากมากไปหาน้อย จะนำเอาคำหลัก (Keyword) ที่ต้องการค้นหาไปเปรียบเทียบกับ Pattern ข้อมูลที่ได้จัดเตรียมไว้ แล้วนำผลลัพธ์ที่ได้จากการค้นหามาจัดลำดับความสำคัญหรือความน่าจะเป็นของข้อมูลที่ผู้ใช้ต้องการ ซึ่งสามารถคำนวณได้จากสูตรดังต่อไปนี้

$$W_{ij} = IDF_i * TF_{ij}$$

$$IDF_i = \log_2(n / DF_i)$$

โดยที่

- W_{ij} คือ ค่าน้ำหนักของแต่ละคำที่ค้นหาเจอ
- TF_{ij} คือ จำนวนความถี่ของแต่ละคำ ที่ค้นหาเจอในแต่ละ Term
- DF_i คือ จำนวนความถี่ของคำทั้งหมดที่ปรากฏในทุกๆ เอกสาร
- IDF_i คือ ค่าที่ได้จากการคำ Log ฐานสองของจำนวนคำทั้งหมดที่ค้นพบหารด้วย จำนวนความถี่ของคำทั้งหมดที่ปรากฏในทุกๆ เอกสาร
- n คือ จำนวนคำทั้งหมดที่ปรากฏในทุกๆ Term รวมกัน



ภาพ 2.1 แสดงหลักการทำงานของการค้นคืนสารสนเทศ

(อ้างอิงจาก : http://web.missouri.edu/~heink/7301-fs2004/inforetrieval/irmodel.html)

จากภาพ 2.1 สามารถอธิบายหลักการทำงานของเทคนิคการค้นหาข้อมูลบนฐานข้อมูลเชิงสัมพันธ์ ได้ตามขั้นตอนการทำงาน ดังต่อไปนี้

- Formulate Query คือกระบวนการรับ Keyword ที่ผู้ใช้ต้องการค้นหา
- System Transforms Query คือกระบวนการประมวลผล Keyword ที่ได้รับจากผู้ใช้
- ตัดคำที่ไม่มีความหมายออก เช่น a, an, the เป็นต้น
- แยก Keyword ที่รับเข้ามาออกเป็นคำๆ
- Matching Process คือกระบวนการนำคำที่ได้จากกระบวนการ System Transforms Query ไปเปรียบเทียบกับคำต่างๆ ที่อยู่ในเอกสารในฐานข้อมูล หากเอกสารใดๆ มีคำที่มีความหมายตรงกับคำที่นำไปเปรียบเทียบให้ Return เอกสารนั้นออกมา
- Results List คือกระบวนการจัดลำดับความสำคัญหรือความน่าจะเป็นของเอกสารที่ผู้ใช้ต้องการ
- Relevance Determination คือกระบวนการแสดงผลข้อมูลต่างๆ ที่ผ่านการประมวลผลให้ผู้ใช้งาน

2.2 International Components for Uni-Code (ICU) (อ้างอิงจาก : http://site.icu-project.org)

ICU คือเทคโนโลยีที่นำมาใช้ในการแยกคำต่างๆ ออกจากกันโดยรับอินพุต (Input) เป็นข้อความ เทคโนโลยีดังกล่าวได้มีการใช้งานอย่างแพร่หลาย อาทิ เช่น ผลิตภัณฑ์ของ Google, Apple, Apache, IBM เป็นต้น ซึ่งในปัจจุบันเทคโนโลยี ICU ได้กลายเป็นเทคโนโลยีพื้นฐานของโปรแกรมภาษา C, C++ และ Java โดยพัฒนาภายใต้มาตรฐานของ Open Source License

2.2.1 หลักการทำงานของ ICU

2.2.1.1 กำหนดข้อความที่เราต้องการแยกคำออก ในเทคโนโลยี ICU เราสามารถกำหนดบทความได้โดยไม่จำกัดความยาวของบทความ แต่ไม่ควรใช้บทความที่มีความยาวมากเกินไป เช่น บทความที่มีความยาวหลายๆ หน้ากระดาษ A4 หรือหนังสือทั้งเล่ม เป็นต้น เพราะจะทำให้เสียเวลาในการประมวลผลค่อนข้างมาก

2.2.1.2 กำหนดรูปแบบของภาษาที่ต้องการใช้แยกคำ เช่น US, TH เป็นต้น

2.2.1.3 กำหนดสัญลักษณ์ที่ใช้แบ่งคำ เช่น , - เป็นต้น

2.2.1.4 การเข้ารหัส เทคโนโลยี ICU จะนำเอาบทความที่รับเข้ามาไปเข้ารหัสให้อยู่ในรูปแบบยูนิโค๊ด (Uni-Code) แล้วนำไปเปรียบเทียบกับข้อมูลที่เก็บไว้ในตารางเปรียบเทียบคำของ ICU ชื่อปัจจุบันตารางที่ใช้ในการเปรียบข้อมูลของ ICU ที่นิยมใช้งานและมีความแม่นยำที่สุด จะเป็นเทคโนโลยี ICU ที่ถูกพัฒนาโดยบริษัท IBM

2.2.1.5 ประมวลผลข้อมูลที่รับเข้ามา โดยจะได้เอาท์พุทเป็นคำต่างๆ ที่ถูกคั่นด้วยเครื่องเครื่องหมายที่เรากำหนด

2.2.2 ขอบเขตความสามารถของเทคโนโลยี ICU

2.2.2.1 ภาษาที่สามารถใช้เทคโนโลยี ICU ได้ จำต้องเป็นภาษาที่อยู่ภายใต้มาตรฐานของรูปแบบยูนิโค๊ด (Uni-Code) เท่านั้น

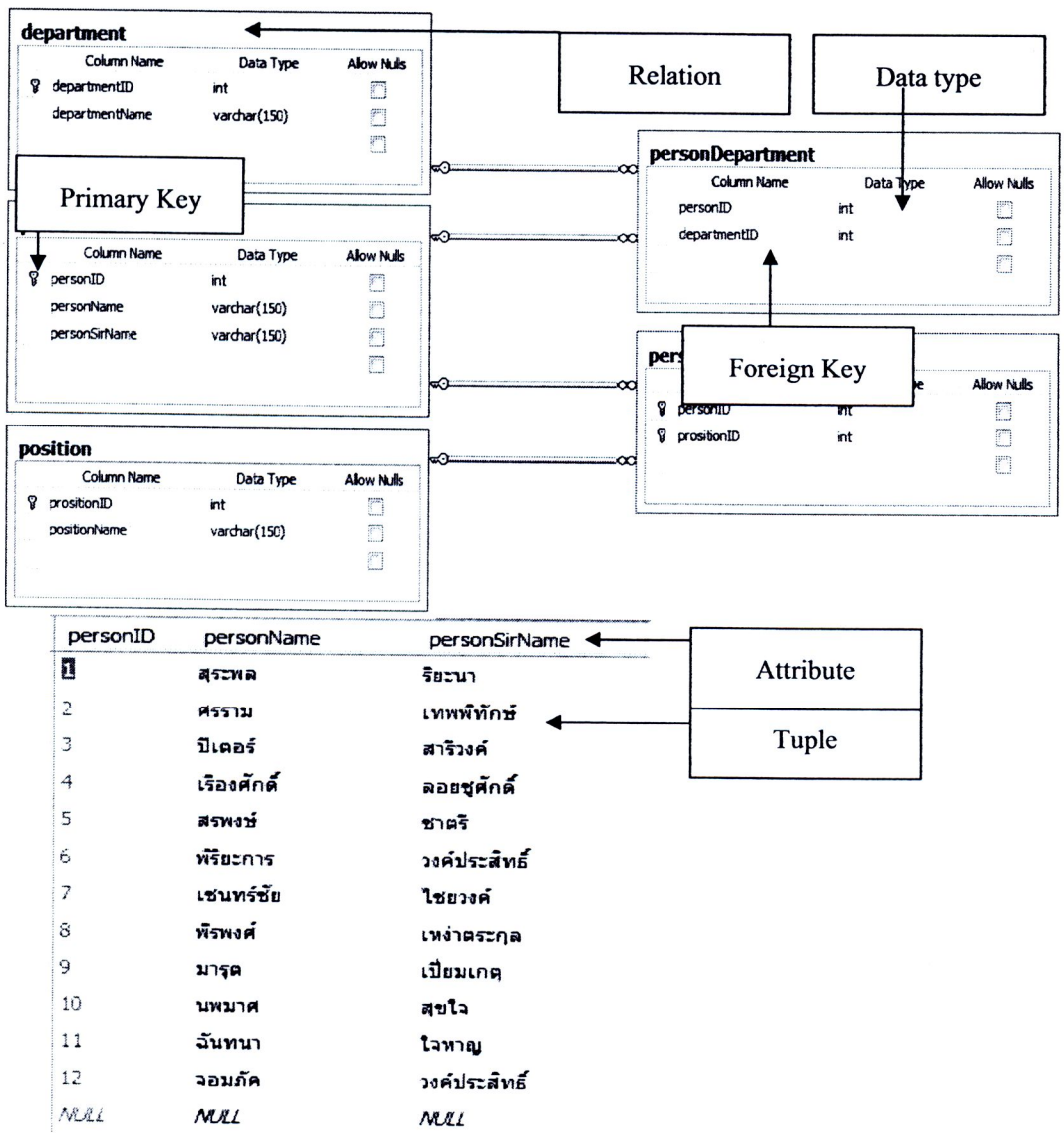
2.2.2.2 ภาษาที่เทคโนโลยี ICU รองรับจะต้องเป็นภาษาที่เป็นไปตามอนุสัญญาและมาตรฐานของภูมิภาคหรือประเทศเท่านั้น

2.2.2.3 การจัดรูปแบบของวันที่ และรูปแบบอัตราเงินจะเป็นไปตามข้อมูลพื้นที่ของภาษาที่เลือกใช้งาน

2.2.2.4 การคำนวณวันที่และเวลาจะขึ้นอยู่กับรูปแบบปฏิทินของภาษาที่เลือก ในกรณีที่ไม่มีรูปแบบของปฏิทินตามรูปแบบภาษาที่เลือก หรือไม่ได้กำหนดรูปแบบปฏิทิน ICU จะเลือกใช้ปฏิทินในรูปแบบคริสต์ศักราชอัตโนมัติ

2.3 ระบบฐานข้อมูลเชิงสัมพันธ์ (Relational Database)

ระบบฐานข้อมูลเชิงสัมพันธ์เป็นฐานข้อมูลที่ใช้โมเดลเชิงสัมพันธ์ (Relational Database Model) คือ ภายในระบบฐานข้อมูลจะประกอบด้วย ตาราง (Relation) ในแต่ละตารางแบ่งออกเป็น 2 ส่วนคือ แถว (Tuple) และคอลัมน์ (Attribute) โดยแต่คอลัมน์ ก็จะมีชนิดข้อมูล (Data type) และคีย์ (Key) เป็นตัวกำหนดรูปแบบของข้อมูลที่จัดเก็บข้อมูล หลักการทำงานของระบบฐานข้อมูลเชิงสัมพันธ์ จะอาศัยหลักการทางคณิตศาสตร์ในเรื่องของ Set เพื่อใช้สำหรับการบริหารจัดการข้อมูลต่างๆ ที่อยู่ภายใต้ฐานข้อมูลเชิงสัมพันธ์



ภาพ 2.2 แสดงส่วนประกอบของฐานข้อมูลเชิงสัมพันธ์

2.4 ภาษา SQL (Structured Query Language)

SQL เป็นภาษามาตรฐานที่ใช้ติดต่อกับระบบฐานข้อมูลเชิงสัมพันธ์ (Relational Database Management System) หรือ RDBMS ซึ่ง ANSI ได้ประกาศออกมาอย่างเป็นทางการ ดังนั้น ผู้ที่ทำงานกับฐานข้อมูลในปัจจุบันจำเป็นต้องรู้ เนื่องจากระบบฐานข้อมูลที่มีอยู่ในปัจจุบันเกือบทั้งหมดเป็นระบบฐานข้อมูลแบบ RDBMS SQL สามารถแบ่งคำสั่งออกเป็น 4 กลุ่ม คือ

2.4.1 Data Manipulate (DML) เป็นคำสั่งจัดการข้อมูล ได้แก่ Insert, Update, Delete, Rollback, Commit

2.4.2 Data Definition (DDL) เป็นคำสั่งจัดการกับไฟล์ในฐานข้อมูล ได้แก่ Create, Alter, Drop

2.4.3 Query เป็นคำสั่งการเรียกดูข้อมูล คือ Select

2.4.4 Data Control เป็นคำสั่งจัดการความปลอดภัย

2.5 การค้นหาข้อมูลใน RDBMS โดยทั่วไป

การค้นหาข้อมูลบนฐานข้อมูล RDBMS จะกระทำได้ผ่านภาษา SQL (Structured Query Language) โดยใช้คำสั่งในกลุ่มของ Query เป็นคำสั่งการเรียกดูข้อมูล คือ Select

2.5.1 Like เป็นคำหลักที่ใช้สำหรับการระบุเงื่อนไขการเลือกข้อมูลในตาราง (Relation) โดยทำการค้นหาข้อความที่ระบุภายในคอลัมน์ (Attribute) ที่กำหนด

```
SELECT Attribute FROM Relation WHERE LIKE Attribute '%Keyword%'
```

โดยภายใต้คำสั่ง Like เราสามารถกำหนดรูปแบบสัญลักษณ์ที่ใช้แทนตัวอักษรต่างๆ นอกเหนือจากที่ผู้ใช้กำหนดได้ คือ

2.5.1.1 % ใช้แทนอักษรตัวอะไรก็ได้และมีกี่ตัวก็ได้

2.5.1.2 _ ใช้แทนตัวอักษรหนึ่งตัวตัวอะไรก็ได้

แต่ปัญหาหลักในการค้นหาข้อมูลแบบนี้คือ RDBMS จะ คืนค่า(Return) ทั้งหมดที่ค้นหาเจอออกมา โดยไม่ได้คำนึงถึงระดับความสำคัญของข้อมูลที่ผู้ใช้ต้องการ

2.5.2 Full Text Search

Full Text Search คือเทคโนโลยีที่ใช้ในการค้นหาข้อมูลในระบบฐานข้อมูลเชิงสัมพันธ์อีกเทคโนโลยีหนึ่ง ที่ถูกพัฒนามาเพื่อใช้ในการค้นหาข้อมูลเพื่อตอบสนองความต้องการของผู้ใช้ข้อมูล แต่เทคโนโลยีนี้ไม่ได้มีอยู่ในระบบฐานข้อมูลเชิงสัมพันธ์ทุกระบบ จะมีบางระบบฐานข้อมูลเท่านั้นที่สามารถใช้เทคโนโลยีนี้ได้ เช่น Oracle, MSSQL MySQL เป็นต้น

แต่ปัจจุบันเทคโนโลยีเหล่านี้ยังมีปัญหาด้านการใช้งานอยู่ คือใช้ระยะเวลาในการประมวลผลข้อมูลค่อนข้างนาน ราคาสูง และเทคโนโลยีนี้ก็ยังไม่ได้เป็นมาตรฐานของระบบฐานข้อมูลเชิงสัมพันธ์ ทำให้รูปแบบการใช้งาน Full Text Search ของฐานข้อมูลเชิงสัมพันธ์แต่ละผลิตภัณฑ์ก็มีรูปแบบในการใช้งานที่แตกต่างกัน ทำให้ยากต่อการจะศึกษา และนำมาใช้งานจริง

2.6 ตัวอย่างการใช้งานเทคโนโลยีบนฐานข้อมูล MSSQL ของ Microsoft

2.6.1 Contains

```
SELECT Attribute FROM Relation WHERE CONTAINS(Attribute, N'Keyword')
```

การค้นหาข้อมูลโดยใช้คำสั่ง Contains จะป้อนคำหลัก (Keyword) ที่ต้องการค้นหาเป็นคำๆ เท่านั้น ไม่สามารถป้อนเป็นข้อความได้ และรูปแบบการค้นหาข้อมูลของคำสั่ง Contains ระบบฐานข้อมูลจะค้นหาเฉพาะคำที่มีความหมายตรงกับคำหลัก (Keyword) เท่านั้น ตัวอย่างเช่น ต้องการหาคำว่า Song แต่ในฐานข้อมูลไม่มีข้อมูลคำว่า Song แต่มีคำว่า Songs ผลลัพธ์ที่ได้ คือ **ไม่พบข้อมูลที่ต้องการค้นหา** แต่หากต้องการให้ค้นหาข้อมูลโดยพิจารณาคำที่มีความหมายใกล้เคียงคำหลัก (Keyword) ด้วย ให้ใช้คำสั่ง **FORMSOF** ช่วยในการค้นหา ตัวอย่างเช่น

```
SELECT Attribute FROM Relation
WHERE CONTAINS(Attribute, N' FORMSOF (INFLECTIONAL, Keyword) ');
```

2.6.2 Freetext

```
SELECT Attribute FROM Relation WHERE FREETEXT(Attribute, N'Keyword');
```

การค้นหาข้อมูลโดยใช้คำสั่ง Freetext จะมีรูปแบบในการค้นหาข้อมูลที่คล้ายๆ กับ คำสั่ง Contains แต่สามารถป้อนคำหลัก (Keyword) เป็นข้อความได้

2.6.3 Containstable

```
SELECT Attribute
FROM Relation
INNER JOIN CONTAINSTABLE(Relation, Attribute 'Keyword') AS KEY_TBL
ON Relation. Attribute = KEY_TBL.[KEY]
```

การค้นหาข้อมูลโดยใช้คำสั่ง Containstable จะมีรูปแบบในการค้นหาข้อมูลที่คล้ายๆ กับ คำสั่ง Contains แต่ในคำสั่ง Containstable จะมีการทำจัดลำดับความสำคัญของข้อมูล (Ranking) ที่ค้นหาพบด้วย โดยจะเรียงลำดับความสำคัญของข้อมูลตามคอลัมน์ (Attribute) ที่ถูกกำหนดให้เป็น KEY_TBL

2.7 งานวิจัยที่เกี่ยวข้อง

สิทธิโชค ปัญญาฤกษ์ชัย และศิพาณีนุชิตประสิทธิ์ชัย (2552) ศึกษาเรื่อง ระบบการค้นคืนสารสนเทศโดยใช้เทคนิค N-Gram มีวัตถุประส่งค์ในการศึกษา เพื่อพัฒนาระบบการค้นคืนสารสนเทศโดยใช้เทคนิค N-Gram เพื่อให้ได้ข้อมูลสารสนเทศที่ตรงตามความต้องการของผู้ใช้งาน และได้ทำการประเมินประสิทธิภาพการใช้ค่าความแม่นยำPrecision และค่าความถูกต้อง Recall ในการค้นคืน โดยผล ได้จากการประเมินประสิทธิภาพของระบบ พบว่ามีค่าเฉลี่ยรวมอยู่ในระ ดับดี และระบบสามารถนำไปใช้ งานได้จริง

วิไลพร เลิศมหาเกียรติ (2553) ศึกษาเรื่อง การค้นคืนสารสนเทศแบบสหความสัมพันธ์ตามหมวดหมู่ระบบทศนิยมดิวอี้ โดยวัตถุประสงค์เพื่อศึกษา วิเคราะห์ข้อมูลแบบสหความสัมพันธ์และพัฒนาขั้นตอนวิธีการค้นคืนสารสนเทศแบบสหความสัมพันธ์ ภายใต้กรอบความรู้การจัดหมวดหมู่ระบบทศนิยมดิวอี้แบบสหความสัมพันธ์ (DDC-MR) โดยวิเคราะห์ความสัมพันธ์ของคำสำคัญในเอกสาร และคำค้นของผู้ใช้กับกรอบความรู้ระบบทศนิยมดิวอี้ คำนวณน้ำหนักและหาสัดส่วนความสัมพันธ์ของเนื้อหาในแต่ละหมวด เพื่อนำมาแสดงเป็นสัดส่วนของความสัมพันธ์สำหรับใช้เปรียบเทียบความคล้ายคลึงระหว่างคำสำคัญในเอกสาร และในส่วนคำค้นของผู้ใช้ ซึ่งการนำเทคนิคการวิเคราะห์หมวดหมู่แบบสหความสัมพันธ์ จะทำให้สามารถนำเสนอเอกสารที่มีรูปแบบสหความสัมพันธ์ของเนื้อหาที่ตรงกัน หรือมีความคล้ายคลึงกับรูปแบบสหความสัมพันธ์ของคำค้นของผู้ใช้ และสามารถใช้คำค้นในภาษาใด ภาษาหนึ่ง เพื่อค้นหาเอกสารจากฐานข้อมูลสากลที่มีการ

จัดหมวดหมู่ตามระบบทศนิยมดิวอี้ได้ และได้มีการทดสอบโดยมีค่าความแม่นยำ (Precision) เท่ากับ 0.74

นิพนธ์ เจริญกิจการ (2544) ได้นำเสนองานวิจัยระบบการค้นคืนเอกสารภาษาไทยด้วยเทคนิคขั้นสูง ระยะที่ 2 โดยมีวัตถุประสงค์เพื่อศึกษาฐานสารสนเทศภาษาไทย และคุณสมบัติต่าง ๆ ของฐานสารสนเทศไทยาและวิธีการค้นคืนภาษาไทยด้วยเทคนิคขั้นสูง การวิจัยนี้จะเน้นไปที่การค้นคืนในแบบแนวความคิด เทคนิคที่ได้รับการยอมรับอย่างกว้าง ขวางเช่น vector space และ latent semantic indexing เป็นต้น

นพดล หมื่นโพ (2548) ได้ศึกษาเรื่องการเพิ่มประสิทธิภาพระบบค้นคืนข้อสนเทศภาษาไทยด้วยนามวลีผันแปรและออนโทโลจี มีวัตถุประสงค์ในการศึกษา เพื่อศึกษาลักษณะและปัญหาในการวิเคราะห์โครงสร้างนามวลีภาษาไทยและการใช้คำอ้างอิงในภาษาไทย และพัฒนาเทคนิคเพื่อเพิ่มประสิทธิภาพการหาตัวแทนเอกสารและขยายคำค้นสำหรับระบบค้นค้นข้อสนเทศที่เหมาะสมกับเอกสารภาษาไทย ที่มีความถูกต้องสูงและนำไปใช้ได้อย่างมีประสิทธิภาพจี จากการศึกษานี้พบว่าการใช้นามวลีผันแปรร่วมกับนิพจน์ระบุนามเป็นตัวแทนเอกสารและใช้คำกับนิพจน์ระบุนามและนามวลีร่วมกับคำที่ได้จากการขยายคำค้นเป็นตัวแทนคำค้น ทำให้ประสิทธิภาพของระบบค้นคืนเอกสารดีขึ้น ค่าความถูกต้อง สูงขึ้นเฉลี่ย 103%

พงษ์ปัญญา จงจักรพันธ์ (2542) ได้ศึกษาเรื่องระบบค้นคืนข้อมูลสารสนเทศโดยใช้ภาษาธรรมชาติ มีวัตถุประสงค์ในการศึกษา เพื่อพัฒนาระบบค้นคืนสารสนเทศโดยใช้ภาษาธรรมชาติ โดยระบบจะทำการวิเคราะห์และแปลงคำร้องของผู้ใช้ด้วยวิธีการเก็บรวบรวมข้อมูลแบบ non-deterministic top-down โดยใช้รูปแบบไวยากรณ์เป็นตัวช่วยในการค้นหาความหมายร่วมกับพจนานุกรมฐานความรู้ และใช้ mSQL เป็นฐานข้อมูลในการศึกษาครั้งนี้