การใช้เทคนิคการระบุชื่อเฉพาะในการสกัดสารสนเทศเพื่อใช้สร้างฐานความรู้เฉพาะด้าน

อ. สุพัฒนวรี ทิพย์เจริญ [1,2,3]

บทคัดย่อ

เนื่องจากปริมาณข้อมูลที่มีแนวโน้มเพิ่มมากขึ้นส่งผลให้การสร้างฐานความรู้มีความจำเป็นอย่างยิ่งต่อองค์กร ทั้งนี้การสร้างฐานความรู้นั้นต้องอาศัยเทคนิคหลายด้าน อาทิเช่น การทำเหมืองข้อความ การประมวลผลภาษาธรรมชาติ การสืบค้นข้อมูล โดยเฉพาะการสกัดความรู้จากเอกสาร หรือแหล่งข้อมูลต่างๆ ปัจจุบันงานทางด้านการสกัดสารสนเทศเป็นงานที่ได้รับความสนใจ ยกตัวอย่างเช่น การสกัดสารสนเทศจาก เว็บไซต์ จดหมายอิเล็กทรอนิกส์ ห้องสมุดดิจิตัล ชีวการแพทย์ฯลฯ และมีงานวิจัยมากมายที่ศึกษาและพัฒนาเทคนิควิธีการสกัดสารสนเทศให้มีประสิทธิภาพ และตอบสนองต่อปริมาณสารสนเทศที่มากขึ้น วิธีการระบุชื่อเฉพาะนั้นเป็นวิธีการหนึ่งของการสกัดสารสนเทศที่สำคัญ ซึ่งช่วยในการระบุชนิดของคำและกลุ่มคำให้ถูกต้อง ในบทความนี้จะกล่าวถึงความสำคัญและหลักการของวิธีการระบุชื่อเฉพาะ ตลอดจนการประยุกต์ใช้กับงานด้านต่าง ๆ ทั้งนี้เพื่อเป็นแนวทางวิจัยต่อผู้ที่มีความสนใจทางด้านการทำเหมืองข้อความ และการสกัดสารสนเทศต่อไป

คำสำคัญ

การสกัดสารสนเทศ การระบุชื่อเฉพาะ การคัดเลือกคำสำคัญ การสร้างฐานความรู้

บทนำ

ในยุคแห่งสารสนเทศ ข้อมูลกลายเป็นแหล่งทรัพยากรที่สำคัญในการสร้างฐานความรู้ ทั้งนี้เพราะการจัดการ การประมวลผล และเพิ่มประสิทธิภาพข้อมูลจะนำไปสู่การพัฒนาขององค์กร และด้วยปริมาณของข้อมูลที่เพิ่มมากขึ้นจากหลายแหล่งข้อมูลทั้งในองค์กร และจากแหล่งอื่นๆ อาทิ เช่น เว็บไซต์ เอกสาร บทความต่างๆ ทำให้การรวบรวมข้อมูล และสกัดสารสนเทศจากข้อมูลทำได้ยากส่งผลให้การสร้างฐานความรู้นั้น จำเป็นต้องอาศัยเทคนิคหลายๆ ด้าน ได้แก่ การทำเหมืองข้อความ ซึ่งเป็นการค้นหาความรู้จากฐานข้อมูลเอกสารที่มีอยู่อย่างมหาศาลโดยอัตโนมัติ เพื่อค้นหารูปแบบแนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง หลักคณิตศาสตร์ ซึ่งผลที่ได้จากการทำเหมืองข้อความมีหลายรูปแบบ เช่น การสรุปเอกสารข้อความ การแบ่งประเภทเอกสารข้อความ การแบ่งกลุ่มเอกสารข้อความเป็นต้น การประมวลผลภาษาธรรมชาติที่ศึกษาปัญหาในการประมวลผลและใช้งานภาษาศาสตร์ ทั้งนี้เพื่อให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์ได้ ตลอดจนการสืบค้นข้อมูล และการสกัดความรู้จากเอกสารได้อย่างถูกต้องและ ตรงตามความต้องการ (Wikipedia, 2008)

[1] ห้องปฏิบัติการวิจัยชีวสารสนเทศ (Bioinformatics Research Laboratory) คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

[2] ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

[3] อาจารย์ประจำภาควิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยฟาร์อีสเทอร์น

การสกัดสารสนเทศถือเป็นขั้นตอนสำคัญ เนื่องจากข้อมูลส่วนใหญ่มาจากหลายแหล่ง ซึ่งมีความแตกต่างกัน ทั้งที่เป็นข้อมูลแบบมีโครงสร้างและไม่มีโครงสร้างทำให้ยากต่อการรวบรวมและการเลือกใช้ข้อมูล (Yu-shan Changand Yun-Hsuan Sung, 2005) จึงได้มีงานวิจัยที่จะศึกษา และพัฒนาวิธีการสกัดสารสนเทศจากแหล่งข้อมูลต่างๆ นั้นได้อย่างมีประสิทธิภาพ โดยอาศัยเทคโนโลยีทางด้านคอมพิวเตอร์ที่มีความก้าวหน้าเข้ามาช่วยในการเข้าถึงข้อมูล ตลอดจนทำการอ่านและวิเคราะห์ข้อมูลแทนมนุษย์ โดยงานที่กำลังได้รับความสนใจได้แก่ การระบุชื่อเฉพาะ ซึ่งเป็นวิธีการหนึ่งของการสกัดสารสนเทศ การระบุชื่อเฉพาะมีประโยชน์อย่างมาก เพราะเป็นการช่วยระบุขอบเขต และชนิดของคำ ทำให้สามารถค้นหา และสืบค้นข้อมูลได้อย่างถูกต้อง ปัจจุบันมีการพัฒนาเทคนิคในการระบุชื่อเฉพาะมากมายเพื่อนำไปใช้งานเฉพาะด้านไม่ว่าจะเป็นการสกัดสารสนเทศจากห้องสมุดดิจิตอล อีเมล์ ชีวการแพทย์ เว็บไซต์ต่างๆ ในหลายๆ ภาษา ทั้งภาษาอังกฤษ ภาษาญี่ปุ่น ภาษาจีน รวมถึงภาษาไทย การระบุชื่อเฉพาะถูกนำไปใช้ในงานต่างๆ มากมาย เช่น การหาขอบเขตเพื่อทำการตัดคำ (การตัดคำเป็นปัญหาพื้นฐานในการเตรียมข้อมูล) การสร้างคำในพจนานุกรม เป็นต้น ดังนั้น การระบุชื่อเฉพาะจึงเป็นขั้นตอนสำคัญของการสกัดสารสนเทศ

แต่เนื่องจากความซับซ้อนทางด้านโครงสร้างทางภาษาทำให้การระบุชื่อเฉพาะยังคงเป็นงานที่มีความน่าสนใจ และ มีความท้าทายต่อการพัฒนาเทคนิคการระบุชื่อเฉพาะ ตลอดจนโปรแกรมการใช้งานเฉพาะด้านให้มีการระบุชื่อได้อย่างถูกต้องแม่นยำ เช่น งานวิจัยการระบุชื่อทางการเกษตรสำหรับภาษาไทย (Asanee, et. al, 2002) โปรแกรมการระบุชื่อเฉพาะของไทย ที่พัฒนาโดย ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติประเทศไทย (Human Language Technology National Electronics and Computer Technology Center, 2007) โปรแกรมโปรไมเนอร์ระบุชื่อ ยีน และ ดีเอ็นเอ (Daniel, et. al, 2005) โปรแกรมระบุชื่อเฉพาะที่พัฒนาโดยมหาวิทยาลัยเสตนฟอร์ต (Jenny Rose Finkel, 2007)

ทั้งนี้บทความนี้ต้องการนำเสนอถึงหลักการและความสำคัญของวิธีการระบุชื่อเฉพาะ เพื่อเป็นประโยชน์ต่องานวิจัยสำหรับนักวิจัยที่มีความสนใจในด้านการทำเหมืองข้อมูล รวมถึงการสกัดสารสนเทศเพื่อสร้างฐานความรู้เฉพาะด้าน โดยบทความนี้จะมีการนำเสนอถึงปัญหา หลักการ วิธีการสกัดสารสนเทศตลอดจนเทคนิคโมเดลของการระบุชื่อเฉพาะรวมถึงแนวทางในการพัฒนา ในหัวข้อถัดไป

การสกัดสารสนเทศ (Information Extraction)

การสกัดสารสนเทศ เป็นการสืบค้นข้อมูลอย่างหนึ่งที่มีจุดประสงค์ในการสกัดโครงสร้างข้อมูลอย่างอัตโนมัติ เช่น การจัดประเภท การหาความหมาย ใจความสำคัญจากเอกสารที่ไม่มีโครงสร้าง (Wikipedia, 2008) โดยการสกัดสารสนเทศจะทำการระบุ และดึงเอาคำที่สนใจออกจากประโยคหรือข้อมูล ซึ่งเป็นขั้นตอนที่สำคัญ และเป็นพื้นฐานสำหรับการนำไปประยุกต์ใช้ในการทำเหมืองข้อความ (Hercules A. and Edilson, 2007)

ความสำคัญของวิธีการสกัดสารสนเทศ คือ การ์ช่วยจัดการกับปริมาณข้อมูลสารสนเทศที่มีเพิ่มมากขึ้นในปัจจุบัน ยกตัวอย่างเช่น การสกัดความรู้จากการแปลงข้อมูลให้อยู่ในรูปของความสัมพันธ์ การใช้ Marking-Up โดย XML tags ทำให้ข้อมูลอยู่ในรูปแบบที่มีโครงสร้าง มีความหมายและสะดวกต่อการนำไปใช้ วิธีการการสกัดสารสนเทศที่ได้รับความนิยมมีอยู่ 3 วิธีได้แก่ วิธี Rule Learning Based วิธี Classification Based และวิธี Sequential Based

1. วิธีการสกัดสารสนเทศ (Methodologies of Information Extraction)

- Rule Learning Based : วีธีการนี้สามารถแบ่งได้เป็น 3 วิธี ได้แก่
  - o Dictionary-Based Method เป็นการใช้พจนานุกรมซึ่งประกอบด้วยคำ ในการระบุหาเอกสารที่เกี่ยวข้องจากข้อความได้
  - o Rule-Based Method วิธีการนี้มีความแตกต่างจาก Dictionary-Based Method คือจะไม่ใช้พจนานุกรมแต่จะใช้การสร้างกฎแทน วิธีการนี้ถูกใช้มากสำหรับเอกสารWebsite ซึ่งมีลักษณะแบบ Semi Structured ดังภาพที่ 1

Pattern					Action
Word	POS	Kind	Lookup	Name Entity	
;	;	Punctuation			
Patrick	NNP	Word	Person's first name	Person	<Speaker>
Stroh	NNP	Word			
,	,	Punctuation			
Assistant	NN	Word	Job title		
Professor	NN	Word			
,	,	Punctuation			
SDS	NNP	Word			

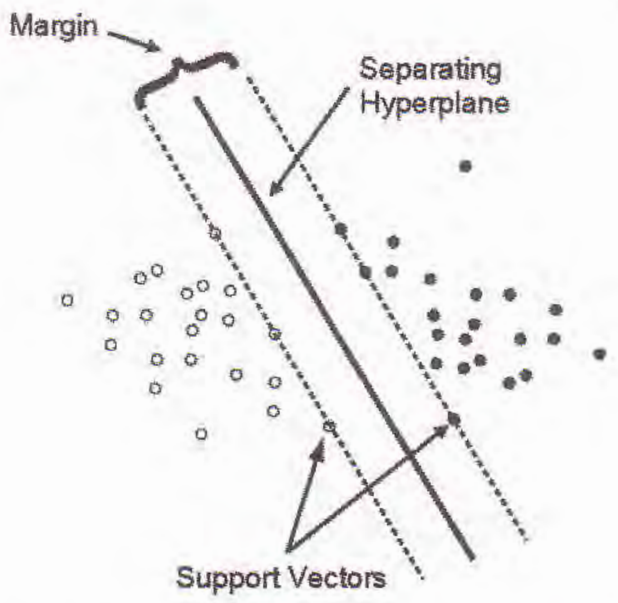
ภาพที่ 1 : Rule Based Method

o Wrapper Induction วิธีการนี้มุ่งเน้นทั้งเอกสารที่เป็นแบบ Structured and Semi Structured ซึ่งมีการทำงานเป็นการเรียนรู้แบบอัตโนมัติ โดยมีการใช้ training data set และ induction algorithm ในการสกัดข้อมูลที่ต้องการ ดังภาพที่ 2

```
D1: <B> Congo </B><I>242</I><BR>
D2: <B> Egypt </B><I>20</I><BR>
Rule: *'<B>' (*)  '</B>' (*) ' <I> '(*) '</I>'
Output: Country__Code
{Country@1}{AreaCode@2}
```

ภาพที่ 2 : Wrapper Induction

● Classification Based : วิธีการนี้เป็นวิธีที่ถูกเสนอขึ้นมาเพื่อใช้ในการแก้ปัญหาการแบ่งแยกประเภทข้อมูลแบบ Supervised machine learning ยกตัวอย่าง เช่น Support Vector Machine (SVM) Cortes and Vapnik (1995) เป็นเทคนิคที่ได้รับความนิยมและมีประสิทธิภาพเป็นอย่างมาก (David Meyer, 2007) มีงานวิจัยมากมายที่ใช้วิธี SVM ในการแบ่งแยกประเภทของเอกสาร SVM เป็น Linear function หลักการของ SVM คือ ทำการหาระยะการแบ่งประเภทที่ดีที่สุดของข้อมูล โดยอาศัย Hyper-plane ซึ่ง SVM รู้จักกันดีในปัญหาแบบ Optimization Problem ดังภาพที่ 3



ภาพที่ 3 : Classification ด้วย hyper plane

Linear Function ของ SVM คือ $f(x) = w^T x + b,$

f(x) : เป็นฟังก์ชันในการแบ่งแยกข้อมูล

w^T *: เป็น Parameter ของโมเดล*

x : คือ ข้อมูล

b : คือ ค่าคงที่

สมการของ Hyper plane (เส้นแบ่งกลุ่ม) ได้แก่ $x^T w + w_0 = 0$

โดยที่ w และ w_0 *คือ Parameter ของโมเดล*

SVM ให้ผลลัพธ์ที่มีความถูกต้อง แม่นยำ (Recall & Precision) ในอัตราที่สูง จึงได้รับความนิยมในการนำไปใช้ในการแยกประเภทเอกสารต่างๆ อยู่เสมอ

● Sequential Based : วิธีการนี้เป็นการสกัดสารสนเทศด้วยการกำหนด และ ระบุชนิดของคำ (Sequential Labeling) นั่นคือ เอกสาร ประกอบด้วย คำ โดยจะมีการระบุชนิดของคำ (Part- Of-Speech) ให้เหมาะสม ซึ่งวิธีการนี้จะอาศัยความรู้ทางด้านภาษา (Natural language processing) เข้ามาช่วยพิจารณาด้วย เช่น ประโยคในเอกสาร

"Pierre Vinken will join the board as a nonexecutive director Nov.29."สามารถที่จะใช้ POS ในการระบุหน้าที่ของคำ (Token) ได้ดังนี้

[NNP Pierre] [NNP Vinken] [MD will] [VB join]
[DT the] [NN board] [IN as] [DT a] [JJ nonexecu-
tive] [NN director] [NNP Nov.] [CD 29] [. .]

โดยวิธีการนี้จะอาศัยเงื่อนไขความน่าจะเป็น (Conditional Probability) ในการทำนายโอกาสความเป็นไปได้ที่ให้ค่าสูงสุดของหน้าที่ของคำ ดังสมการ

$$\mathbf{y}^* = argmax_y\, p(\mathbf{y}|\mathbf{x})$$

p คือความน่าจะเป็นที่สูงที่สุดที่ข้อมูล x จะมีชนิดเป็น y

ในขณะที่ x หมายถึง ประโยค (Sequence of words) และ y แทนรูปแบบในการระบุหน้าที่และชนิดของคำที่ต้องการทำนาย (Label Sequences)วิธีการนี้ต่างจาก Rule Based และ Classification Based โดย Sequence labeling นั้นมีเรื่องของความสัมพันธ์เข้ามาเกี่ยวข้อง เน้นการปรับปรุงความถูกต้องของการสกัดสารสนเทศ มีเทคนิคมากมายที่ถูกสร้างขึ้นเพื่อ สนับสนุนงาน Sequence labeling ได้แก่ Hidden Markov Model, Maximum Entropy Markov Model และ Conditional Random Fields

2. ขั้นตอนการสกัดสารสนเทศ (Stages of Information Extraction)

การสกัดสารสนเทศ สามารถแบ่งได้ออกเป็น 2 ขั้นตอน คือ การเรียนรู้ (Training) และ การสกัดความรู้จากข้อมูลที่เป็นเอกสารประกอบด้วยข้อความต่างๆ (Extraction)

การเรียนรู้ (Training) โมเดล จะถูกสร้างจากการเรียนรู้การอ่าน และ การจำจากเอกสารข้อมูลที่นำมาเรียนรู้

การสกัดความรู้จากข้อมูลที่เป็นเอกสาร (Extraction) ส่วนประกอบของคำที่ต้องการจะถูกระบุและสกัดจากเอกสารโดยใช้ลักษณะของการเรียนรู้จากโมเดล (Learned Model) เมื่อข้อมูลถูกกำหนด (Annotated) ข้อมูลจะอยู่ในรูปของข้อมูลที่มีความเฉพาะเจาะจง และมีโครงสร้าง (Metadata)

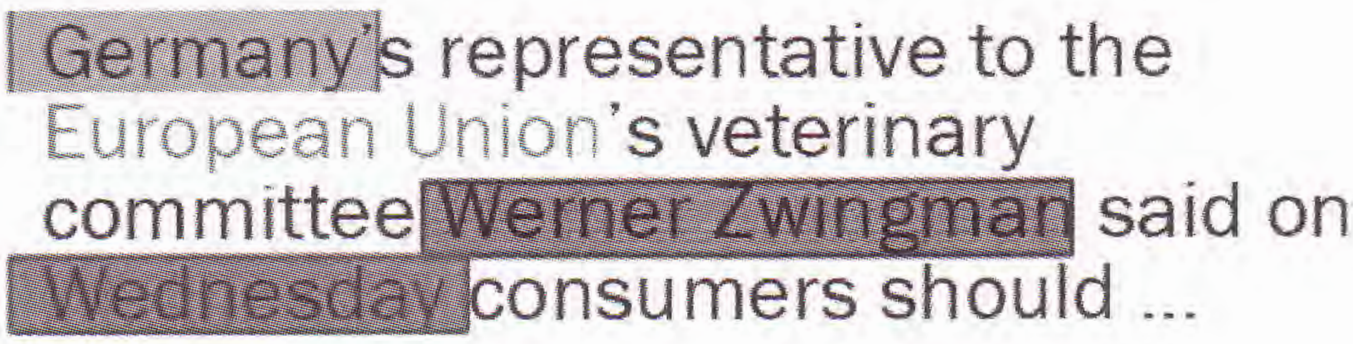
ในปัจจุบันมีการสร้างโมเดลสำหรับการสกัดสารสนเทศมากมาย และได้ถูกพัฒนาขึ้นมาเป็นโปรแกรมเพื่อใช้งานเฉพาะด้านมากมาย เช่น การใช้ POS tagging การหาความสัมพันธ์ของข้อมูลจากคลังข้อมูล (Terminology Extraction) การระบุความสัมพันธ์ระหว่างสิ่งที่น่าสนใจ (Relation Extraction) เช่น ความสัมพันธ์ระหว่างคนและองค์กร (Bill works for IBM.)รวมถึง การระบุชื่อเฉพาะของสิ่งที่สนใจได้อย่างถูกต้อง (Named Entity Recognition) ไม่ว่าจะเป็นชื่อบุคคล สถานที่ องค์กร ฯลฯ ดังจะกล่าวในหัวข้อถัดไป

3. การระบุชื่อเฉพาะ(Named Entity Recognition)

การระบุชื่อเฉพาะ เป็นการสกัดสารสนเทศอย่างหนึ่งที่ใช้วิธีการ Sequence labeling ในการกำหนดชื่อเฉพาะ เป็นงานที่ถูกนำไปประยุกต์ใช้อย่างกว้างขวางในศาสตร์หลายแขนง เช่น ทางการเกษตร การสกัดข้อมูลส่วนบุคคลจากเอกสาร หรือแม้แต่ด้านชีวการแพทย์ เป็นต้น

3.1 นิยามของการระบุชื่อเฉพาะ

เป็นหนึ่งในวิธีการสกัดสารสนเทศ ในลักษณะของการหา และ ระบุ ตลอดจนแยกประเภท ส่วนประกอบของข้อความให้อยู่ในประเภทที่ต้องการ เช่น ชื่อของแต่ละบุคคล องค์กร สถานที่ การบ่งบอกถึงเวลา ฯลฯ (Wikipedia, 2008) ยกตัวอย่างดังภาพที่ 4



ภาพที่ 4 : ตัวอย่างการระบุชื่อเฉพาะ (Jenny Rose , 2007)

โดยที่ Germany แทน Location, European Union แทน Organization Werner Zwingman แทนชื่อคนและ Wednesday บอกว่าเป็นวันพุธ

การระบุชื่อเฉพาะถูกนำไปใช้กับงานหลายๆ ด้าน เช่น

งานทั่วๆไป : การระบุถึงวัน และ เวลามาตรวัดต่างๆ(จำนวนร้อยละ หน่วยของเงิน น้ำหนัก) ที่อยู่จาก Email Address

งานเฉพาะด้าน : การระบุถึงชื่อของพืชในการเกษตร ชื่อของยา หรือโรค ตลอดจน ลักษณะอาการทางการแพทย์

3.2 ปัญหาของการระบุชื่อเฉพาะ (Named Entity Recognition Problem)

การระบุชื่อเฉพาะนั้นเป็นวิธีที่มีความน่าสนใจ และเป็นปัญหาที่มีความท้าทาย อันเนื่องมาจากปัญหาดังต่อไปนี้ (Hamish Cunningham, Kalina Bontcheva, 2007)

o ความกำกวมในการระบุชนิดของชื่อเฉพาะ เช่น คำว่า May อาจจะหมายถึง ชื่อคน หรือ หมายถึงเดือนพฤษภาคม Washington อาจจะหมายถึงชื่อคน หรือสถานที่

o ความหลากหลายของการระบุชื่อเฉพาะ เช่น Marry Kate Miss Marry Kate หมายถึงคนๆเดียวกัน

ดังนั้นการระบุชื่อเฉพาะจึงเป็นปัญหาที่ต้องใช้ความเชี่ยวชาญทางโครงสร้างของภาษาเข้ามาช่วย การระบุชื่อเฉพาะส่วนใหญ่ ที่ใช้ในการสกัดสารสนเทศจากเอกสารได้แก่ ใคร (Person) เมื่อไร (Time) ที่ไหน (Location) และสิ่งไหน (Organization) เช่น

"***คณบดีคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยฟาร์อีสเทอร์น*** ได้เดินทางไปร่วมประชุมวิชาการที่ ***ประเทศออสเตรเลีย*** เมื่อ***วันเสาร์ที่ 23***"

3.3 โมเดลที่ใช้ในการระบุชื่อเฉพาะ (Named Entity Recognition Models) แบ่งได้เป็น 2 วิธีการใหญ่ๆ ด้วยกันคือ

1.) Rule-Based : ประกอบด้วยโมเดลต่างๆ คือ

- การกำหนด Regular Expression โดย ผู้เชี่ยวชาญ
- การกำหนดด้วย พจนานุกรม Dictionary-Based
- การกำหนดด้วยไวยากรณ์ Linguistics Grammar or Heuristic Rule

ข้อดีของวิธีการนี้คือ มีความถูกต้องสูงเพราะมีผู้เชี่ยวชาญในการระบุหน้าที่ที่ชัดเจนไม่จำเป็นต้องใช้ความซับซ้อนของคอมพิวเตอร์เข้ามาช่วยมากนัก แต่ข้อเสียของวิธีการนี้ คือ ต้องมีผู้เชี่ยวชาญทางด้านนั้นๆ อย่างแท้จริง และยากที่จะมีพจนานุกรมหรือออกแบบกฎได้อย่างถูกต้องสมบูรณ์ นอกจากนี้ยังใช้เวลาค่อนข้างนานในกรณีที่เป็นงานด้านที่มีความฉพาะเจาะจง

2.) Machine-Learning : เป็นวิธีการรู้จำด้วยเครื่องคอมพิวเตอร์ ประกอบด้วยโมเดลต่างๆ ดังนี้

- Hidden Markov Model (HMM) (Bikel et al., 1997)
- Maximum Entropy Markov Model (MEMM) (Andrew Bortwick et al., 1998)
- Decision tree (S. Baluja et al., 1999)
- Support Vector Machine (Nigel Collier and Koichi Takeuchi, 2002)
- Conditional Random Fields (Andrew McCallum, 2003)
- HMM และ MEMM เป็น โมเดลที่ใช้มากในงานด้าน Sequence Labeling โดยใช้หลักการของความน่าจะเป็นในการทำนาย แต่โมเดลเหล่านี้ จะทำให้เกิดปัญหา Label Bias Problem, Decision tree เป็นการใช้หลักการตัดสินใจแบบต้นไม้ และ SVM เป็นโมเดล Linear Function ที่ใช้สูตรการคำนวณทางคณิตศาสตร์ในการแก้ปัญหา ส่วนวิธีสุดท้ายเป็นวิธีที่กำลังได้รับความสนใจอย่างมากในขณะนี้ มีงานวิจัยมากมายที่นำเอาโมเดล CRFs ไปใช้ประยุกต์ เช่น ในปัญหาการระบุชื่อเฉพาะในภาษาไทยก็มีกลุ่มนักวิจัยได้นำเอาโมเดลนี้ไปประยุกต์ใช้ ในส่วนของงานวิจัยทางด้านชีวการแพทย์ (Biomedical) ก็ได้นำเอาโมเดลนี้ไปช่วยวิเคราะห์ความหมายและหน้าที่ของยีน และดีเอ็นเอ ซึ่งผลลัพธ์ที่ได้ให้ประสิทธิภาพดี วิธีการนี้ใช้หลักการความน่าจะเป็นเช่นเดียวกับ HMM และ MEMM แต่ด้วยเนื่องจาก CRFs เป็นโมเดลกราฟแบบ Undirected Graph จึงทำให้ไม่เกิดปัญหา Label Bias Problem เช่นเดียวกับ HMM และ MEMM

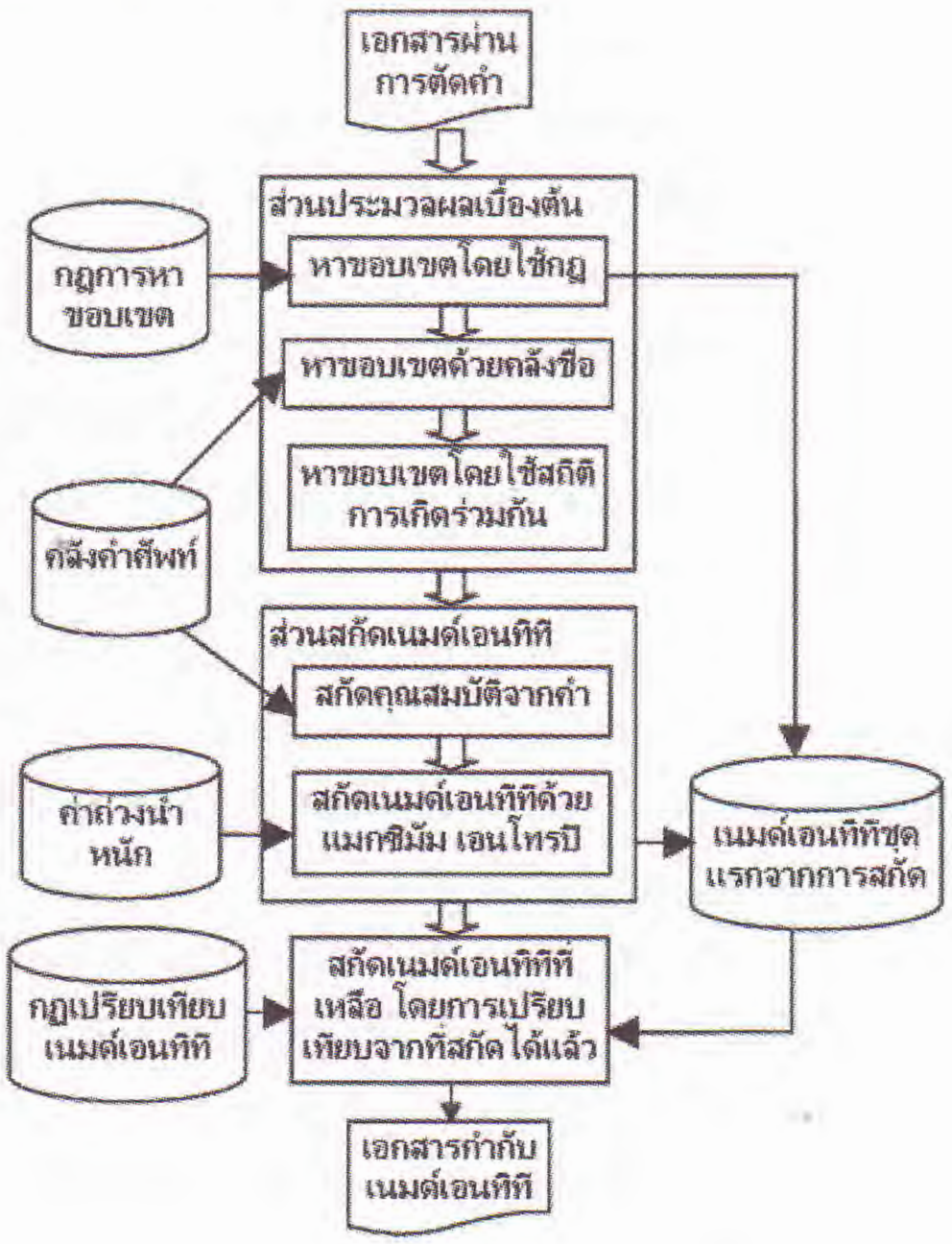
ข้อดีของวิธีการนี้ คือ ลดระยะเวลาลงเนื่องจากใช้คอมพิวเตอร์เข้ามาช่วย สามารถช่วยสร้างความรู้ได้จากคลังข้อมูลที่มีอยู่ ข้อจำกัดของวิธีการนี้ คือ ต้องอาศัยคลังข้อมูลขนาดใหญ่ในการเรียนรู้สร้างโมเดล อีกทั้งต้องการทรัพยากรและความซับซ้อนทางการคำนวณของคอมพิวเตอร์ และประสิทธิภาพของระบบนั้นขึ้นอยู่กับเอกสารที่นำมาเรียนรู้ (Training)

3.4 โครงสร้างของการระบุชื่อเฉพาะ (Architecture of Named Entity Recognition)

กระบวนการเรียนรู้ประกอบด้วยขั้นตอนดังนี้ (หัชทัย และ อัศนีย์, 2546)

1) เตรียมคลังเอกสารเพื่อฝึกฝนระบบ โดยการตัดคำและกำกับชื่อเฉพาะด้วยมือ จากนั้นสกัดเวกเตอร์คุณสมบัติให้กับแต่ละคำในคลังเอกสารฝึกฝนระบบ

2) ฝึกฝนระบบ โดยใช้โมเดลต่างๆ ผลลัพธ์ที่ได้จากขั้นตอนนี้ คือค่าถ่วงน้ำหนักของแต่ละฟังก์ชันคุณสมบัติ



ภาพที่ 5 : กระบวนการระบุชื่อเฉพาะ

กระบวนการสกัดเนมด์เอนทิที ประกอบด้วย 2 ขั้นตอนหลัก

1) การประมวลผลเบื้องต้น เป็นการหาขอบเขตของการระบุชื่อ โดยใช้ข้อมูลจากกฎ คลังชื่อ และจากการคำนวณทางสถิติตามลำดับ

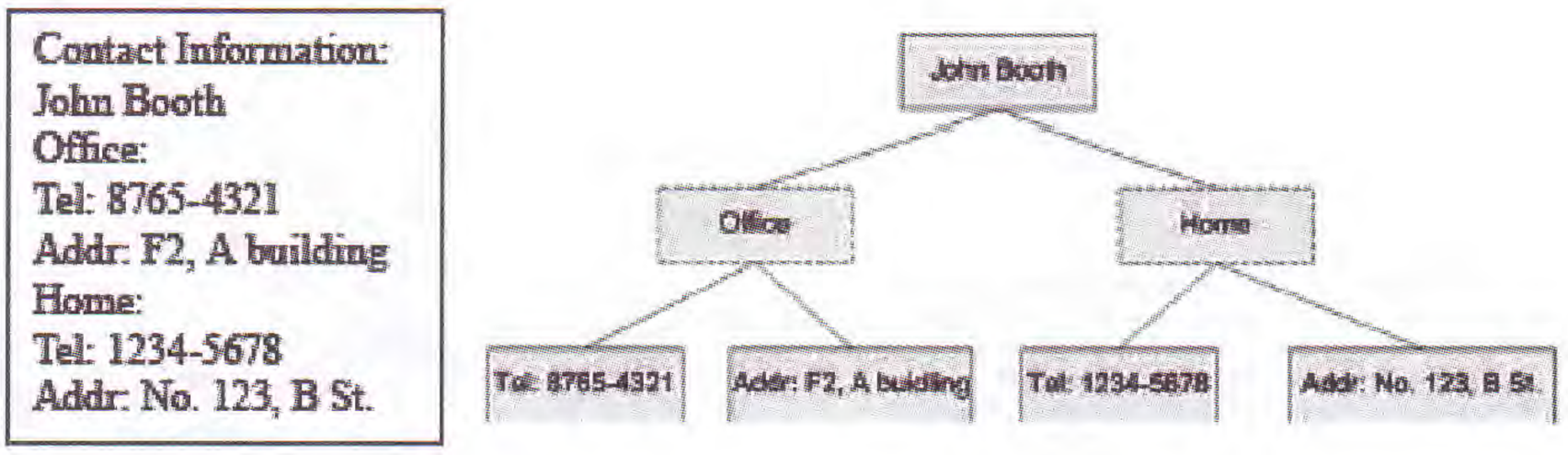
2) สกัดชื่อเฉพาะ โดยสกัดคุณสมบัติจากคำที่สนใจ และจากบริบทข้างเคียง แล้วใช้ค่าถ่วงน้ำหนักของแต่ละคุณสมบัติที่ได้จากขั้นตอนการเรียนรู้มาคำนวณความน่าจะเป็นของชื่อเฉพาะในแต่ละประเภท หลังจากนั้นชื่อเฉพาะที่สกัดได้ จะถูกนำมาเปรียบเทียบกับคำในเอกสารนั้น เพื่อสกัดชื่อเฉพาะที่อาจเหลืออยู่ในเอกสารต่อไป

4. การประยุกต์ใช้งานการระบุชื่อเฉพาะ (Named Entity Recognition Applications)

ในส่วนนี้จะนำเสนองานด้านต่างๆ ที่ได้มีการนำเอาการระบุชื่อเฉพาะไปใช้

4.1 ห้องสมุดดิจิตัล (Information Extraction in Digital Libraries : DL)

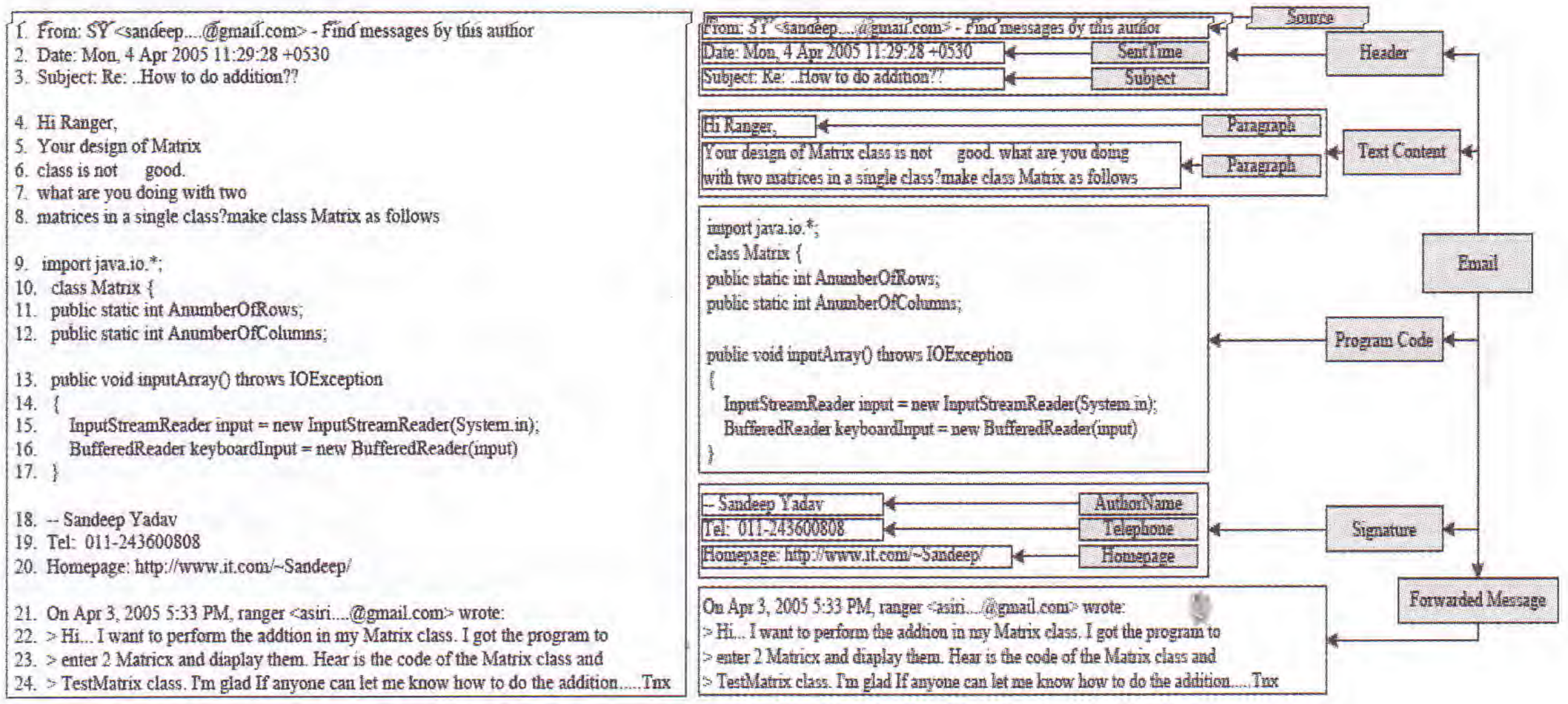
เป็นการจัดโครงสร้างข้อมูลสำหรับช่วยในการหาของผู้ใช้เกี่ยวกับเอกสารและรูปภาพ ซึ่งนักวิทยาศาสตร์และบรรณารักษ์ในปัจจุบันใช้เวลา และแรงงานค่อนข้างมากในการจัดรูปแบบโครงสร้างเอกสารเพื่อให้ง่ายต่อการค้นหาและเรียกใช้ดังนั้นเทคนิคการสกัดความรู้ จึงช่วยให้สามารถจัดการข้อมูลได้อย่างอัตโนมัติ ยกตัวอย่างเช่น Citeseer ซึ่งเป็นเว็บไซต์ที่รวบรวมเอางานบทความทางวิชาการตลอดจนงานวิจัยที่ได้รับการตีพิมพ์มากกว่า 700,000 เอกสารมาจัดการโดยใช้โมเดล เพื่อสร้างเป็นฐานข้อมูลของเอกสารที่มีโครงสร้าง โดยเข้าถึงข้อมูลด้วยการ query คำต่างๆ ที่ต้องการโดย Citeseer ทำการ tag ข้อมูล ได้แก่ Title, Author, Affiliation เป็นต้น โดย โมเดลที่ใช้ได้แก่ SVM ยกตัวอย่างเอกสารและการสกัดข้อมูล ดังภาพที่ 6



ภาพที่ 6 : การสกัดข้อมูลจากเอกสาร

4.2 อีเมล์ (Information Extraction from E-mails)

ปัจจุบันได้มีการสกัดสารสนเทศจาก E-mails ซึ่งโดยเฉลี่ยแล้วคนจะได้รับเมล์อย่างน้อยวันละ 40 ถึง 50 ฉบับต่อวันทำให้จำเป็นต้องใช้เทคนิคของการทำเหมืองข้อความเข้ามาช่วยในการกรอง และสกัดข้อมูลเพื่อสามารถนำไปวิเคราะห์ได้ ดังภาพที่ 7



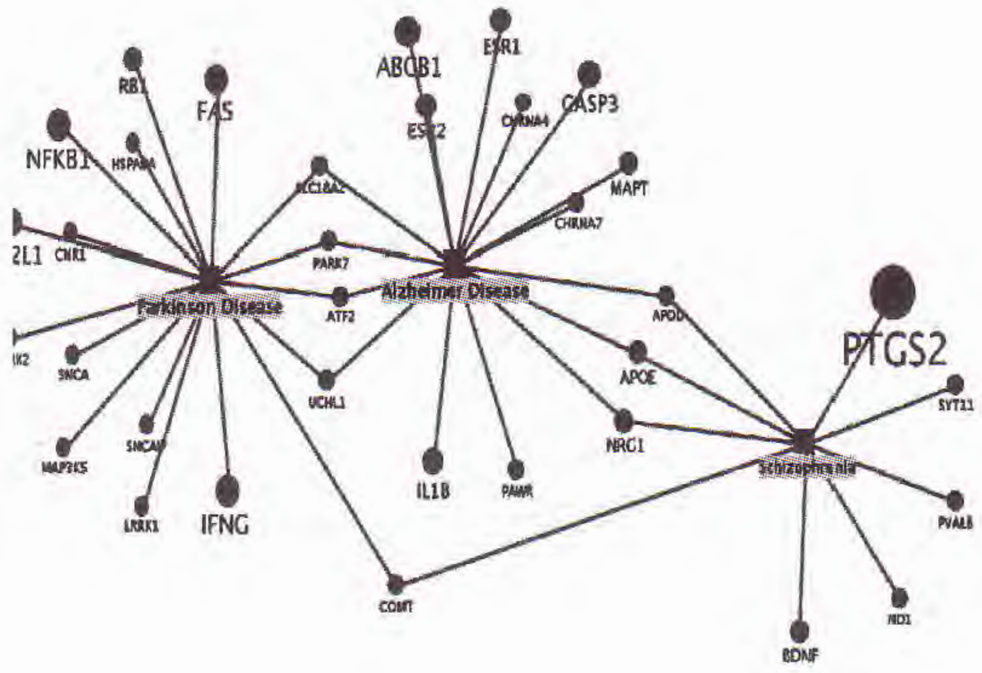
ภาพที่ 7 : การสกัดสารสนเทศจากอีเมล์

4.3 การสร้างฐานข้อมูลเฉพาะทางชีวการแพทย์ (Bio Medical Named Entity Recognition)

มีการสร้างฐานความรู้โดยการหาความสัมพันธ์ระหว่างคำจากเอกสาร ยกตัวอย่าง เช่น ทางชีวการแพทย์ ได้มีการสร้างโปรแกมที่ทำการอ่านเอกสารแล้วทำการระบุคำที่บอกถึงยีน (GENE) ดีเอ็นเอ (DNA) อาร์เอ็นเอ (RNA) แล้วทำการหาความสัมพันธ์ของยีน และโรค โดยนำเสนอออกมาในรูปแบบของกราฟ ดังภาพที่ 8

BACKGROUND AND PURPOSE: The collagen alpha2(I) gene (COL1A2) on chromosome 7q22.1, a positional and functional candidate for intracranial aneurysm (IA), was extensively screened for susceptibility in Japanese IA patients. METHODS: Twenty-one single nucleotide polymorphisms (SNPs) of COL1A2 were genotyped in genomic DNA from 260 IA patients (including 115 familial cases) (mean age, 59.9 years) and 293 controls (mean age, 61.6 years). Differences in allelic and genotypic frequencies between the patients and controls were evaluated with the chi(2) test. Circular dichroism spectrometry was monitored with collagen-related peptides that mimic triple-helical models of type I collagen with Ala-459 and Pro-459 to estimate the conformation and stability of alterations. RESULTS: Significant genotypic association in the dominant model was observed between an exonic SNP of COL1A2 and familial IA patients (chi(2)=11.08; df=1; P=0.00087; odds ratio=3.19; 95% CI, 2.22 to 6.50). This SNP induces Ala to Pro substitution at amino acid 459, located on a triple-helical domain. Circular dichroism spectra showed that the Pro-459 peptide had a higher thermal stability than the Ala-459 peptide. CONCLUSIONS: The variant of COL1A2 could be a genetic risk factor for IA patients with family history.

state, location, gene, type



ภาพที่ 8 : การสร้างฐานข้อมูลเฉพาะทางชีวการแพทย์

จากตัวอย่างเป็นเพียงงานบางส่วนที่นำเอาเทคนิคการระบุชื่อเฉพาะมาใช้ในการสกัดสารสนเทศซึ่งยังมีงานอีกมากที่ไม่ได้กล่าวถึงในที่นี้

สรุป

บทความนี้ได้นำเสนอหลักการและประโยชน์ของการระบุชื่อเฉพาะในการสกัดสารสนเทศ ในงานด้านต่างๆ ซึ่งในปัจจุบันยังคงมีงานวิจัยมากมายที่คิดโมเดล เพื่อปรับให้เหมาะกับงานทางด้านต่างๆที่มีความเฉพาะเจาะจงมากขึ้น โดยแนวโน้มของงานวิจัยส่วนใหญ่ต่อไป จะเน้นในเรื่องของการหาความสัมพันธ์จากความหมายของสิ่งต่างๆ ที่สนใจจากแหล่งข้อมูลต่างๆ ทั้งจากเอกสาร เว็บไซต์ ฯลฯ และสามารถนำวิธีการสกัดสารเทศไปใช้ในการทำนาย หรือสร้างฐานความรู้ในองค์กรธุรกิจ หาความสัมพันธ์ทางด้านการตลาด เพื่อใช้สนับสนุนการตัดสินใจได้ โดยผู้เขียนบทความนี้หวังเป็นอย่างยิ่งว่าบทความนี้ จะช่วยให้ผู้อ่านหรือนักวิจัยมองเห็นภาพกระบวนการการทำงานของการสกัดสารสนเทศด้วยการระบุชื่อเฉพาะ และสามารถนำหลักการไปประยุกต์ใช้ในด้านที่สนใจเพื่อให้เกิดประโยชน์และสร้างสรรค์ งานวิจัยอื่นๆ ต่อไป

เอกสารอ้างอิง

หัชทัย ชาญเลขา และ อัศนีย์ ก่อตระกูล. (2546). **การสกัดนิพจน์ระบุนามในภาษาไทย โดยใช้แมกซิมัมเอนโทรปี โมเดล และอิงความรู้**. กรุงเทพฯ : ภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์.

Asanee K., Nigel C., Koichi T., Kenji O.,Mukda S., Hutchatai C., and Kitsana W., (2001) Collaboration on Named Entity Discovery in Thai Agricultural Texts, EIGHT International workshop on **Academic Information Networks and Systems (WAINS 8),** Tokyo, Japan.

Baluja, S. , Barto, A.G. , Boese, K. D., Boyan, J., Buntine, W., Carson, T., Caruana, R., Davies, S., Dean,T. , Dietterich, T.G., Hazlehurst, Impagliazzo, R., Jagota, A. K., Kim, K.E., McGovern A. , et al., (2000). **Statistical Machine Learning for Large-Scale Optimization.** n.p. : n.p.

Bikel et al. (1997). Named entity recognition using an HMM-based chunk tagger, Proceedings of **the 40th Annual Meeting on Association for Computational Linguistics,** Philadelphia, Pennsylvania.

Borthwick, Andrew. (1998). **A Maximum Entropy Approach to Named Entity Recogniton.** A dissertation submitted in partial fullment of the requirements for the degree if Doctor of Philosophy Computer Science. New York University.

Chang, Yu-shan. and Sung, Yun-Hsuan. (2005). **Applying Name Entity Recognition to Informal Text, cs224n.** n.p. : Department of Computer Science and Department of Electrical Engineering, Stanford University.

Cortes, C. and Vapnik, V. N. (1995). Support vector networks. **Machine Learning.** 20, p. 1-25.

Collier, Nigel and Takeuchi, Koichi. (2002). Use of Support Vector Machines in Extended Named Entity Recognition, In **proceeding of the 6th Coference on Natural Language Learning 2002.**

Cunningham, Hamish and Bontcheva, Kalina. (2007). **Recent Advances in Natural Language Processding International Conference, RANLP - 2007.** Borovets, Bulgaria September 27-29, 2007.

Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. (2005) ProMiner: rule-based protein and gene entity recognition. **BMC Bioinformatics.** 6 (Suppl1) (S14).

Finkel, Jenny Rose. (March 9, 2007). **Named Entity Recognition and the Stanford NER Software.** Stanford University.

Hercules, A. and Edilson, F. (2007). **Emerging Technologies of Text Mining: Techniques and Applications, Information science reference.** New York : Hershey.

McCallum,Andrew (2003). **Efficiently inducing features of conditional random fields.** USA : Computer Science Department University of Massachusetts.

Meyer, David. (October 19, 2007). **Support Vector Machines**. Austria : Technische University at Wien.

Wikipedia Foundation, Inc. (2008). **Information Extraction, NLP.** Retrieved July 18, 2008, from: http://en.wikipedia.org/wiki/Information__extraction

Wikipedia Foundation, Inc. (2008). Named Entity Recognition. Retrieved July 18, 2008, from: http://en.wikipedia.org/wiki/Named__Entity__Recognition