มหาวิทยาลัยบูรพา

Thai Lexical Analysis

ศรายุทธ ฉายสุริยะ, เสรี ชิโนดม และธวัชชัย เอี่ยมไพโรจน์

คณะวิทยาการสารสนเทศ

มหาวิทยาลัยบูรพา

Version 1.0 January 2010

thaiengine

Content

- การประยุกต์ใช้ Thai Lexical Analysis ในงาน
  - การแยกแยะเว็ปไม่พึงประสงค์
  - Search Engine
- แนวทางในการปรับปรุงและใช้ประโยชน์ในอนาคต

งานชิ้นนี้เป็นความร่วมมือระหว่าง

คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา

และ

Nysiis solutions เจ้าของลิขสิทธิ์ ThaiEngine

เว็บไม่พึงประสงค์

ในที่นี้ “เว็บไม่พึงประสงค์” หมายถึง

- เว็บที่มีเนื้อหาที่น่าจะเป็นเป็นภัยสังคม เช่น เนื้อหาเรื่องเพศ ลามก อนาจาร สนับสนุนการฆ่าตัวตาย เป็นต้น
- เว็บที่ลักษณะเป็นแหล่งชุมชน(เสมือน) ของบุคคลที่น่าจะเป็นภัยต่อสังคม หรือมีลักษณะที่ผิดศีลธรรม
- เว็บที่มีเนื้อหาผิดกฎหมาย เช่น มีเนื้อหาหมิ่นพระบรมเดชานุภาพ

การบริหารจัดการเว็บไม่พึงประสงค์ในปัจจุบัน

- ใช้นโยบายในการห้ามคนในองค์กรเข้าไปใช้เว็บเหล่านี้(ไม่สามารถใช้ระบบปิดกั้นได้)
- บางองค์กรที่มี **firewall** ที่มีประสิทธิภาพสูง หรือมี **internet filter gateway** ที่มีคุณภาพสูง
  - สามารถใช้ **blocked list** ที่มีมาจากผู้ให้บริการต่างประเทศได้ (ไม่สามารถแยกแยะเว็บภาษาไทยได้)
- เป็นที่น่าสังเกตว่า **proxy server/firewall** แบบ **open source** ที่เป็นทางผ่านไปใช้ **internet** ขององค์กรขนาดเล็กทั่วไป (เช่นตามโรงเรียน) นั้น ไม่สามารถใช้ **blocked list** ได้ หรือใช้ได้ยากมาก ที่สำคัญคือ ไม่มี **blocked list** ที่มีประสิทธิภาพเพียงพอ
- **ระบบปิดกั้นโดยภาครัฐยังทำงานได้ไม่ดี**

ระบบการปิดกั้นโดยภาครัฐยังไม่ได้ผล

Google ขายค่ะ ไม่ก้น ค้นหา การค้นหาขั้นสูง

ค้นหา: เว็บ หน้าที่เป็นภาษาไทย เว็บจากประเทศไทย

เว็บ แสดงตัวเลือก... ผลการค้นหา 1 - 10 จากประมาณ 32,000

Baldur Talk >>>> V.3 - Pramool.com
2 พ.ย. 2006 ... ปิ๊ก 060027803 ขายคะ อายุ 30 156/45 บางใหญ่ บางบัวทองคะ 32-26-35 สส 156/45 โมกสดได้ไม่ใส่ก้น ไม่แตกปาก 1000/1 ครั้ง มีห้องคะ ...
bbs.pramool.com/webboard/view.php3?katoo=r831253... - แคช - ใกล้เคียง -

ขาย .... คะ เข้ามาชม กานก่อนตื้ย ราคา 700-1500 - 5 โพสต์ - 15 ส.ค. 2009
ขาย ID ออ เวล 45 คะ ((ชลบุรี)) ดูข้างในก่อนได้คะ - 10 โพสต์ - 13 มิ.ย. 2009
~ขายคะ โทรมาคุยกันก่อนนะคะ~, ตกลงราคาได้คะ - 6 โพสต์ - 28 ก.พ. 2007
ทำไม ไม่แจ้งคะ แจ้งสักนิดก่อนขายคะ - 21 โพสต์ - 11 พ.ย. 2006
ผลการค้นหาเพิ่มเติมจาก bbs.pramool.com »

เพื่อน MSN - หาเพื่อน MSN อยากมีเพื่อนคุย หาแฟน หาคนรู้ใจ หาคน ...
115040, soom, หญิง, 24, ขายคะตอนนี้ อวบ ผิวสองสี พอใช้ไม่ขี้เหร่ (ไม่มีรูป) ... 11503 29, พิษณุโลก หาเพื่อนคุย สาระๆ อ่านก่อนแอดนะคะ ...
www.mahamodo.com/webboard.../webboard_msn_main.aspx - แคช - ใกล้เคียง -

แน่ใจหรือไม่ ก่อนจะใช้ปากทำรักให้ผู้หญิง
30 พ.ค. 2009 ... แน่ใจหรือไม่ ก่อนจะใช้ปากทำรักให้ผู้หญิง ... และอุ้มอยากคะ อยากขา ท่านตื่มคะ เพื่อสวิงแข็ง แรงไม่ตก ได้หลายยก กาแฟเยี่ยมสำหรับชาย ...
leelacheewit.com/webboard/question.asp?gid=6631 - แคช -

sl1000/ไม่จำกัดครั้ง อมสดทุกอย่างไม่ก้น รับงานแถวอมตะนครชลบุรี
1 โพสต์ - 1 ผู้เขียน
sl1000/ไม่จำกัดครั้ง อมสดทุกอย่างไม่ก้น รับงานแถวอมตะนครชลบุรี. ... นักศึกษาไซด์ไลน์, เด็ก สาว ไซด์ไลน์, นักศึกษาขายบริการ, หาสาวไซด์ไลน์, sideline, สยามไซด์ไลน์ ... สนใจเมลหรือโทรมานะคะ 080-6450272 sidelinesex@hotmail.com ...
forum.siamsideline.com/index.php?topic=4331.0 -

sl1000/ไม่จำกัดครั้ง อมสดทุกอย่างไม่ก้น รับงานแถวอมตะนครชลบุรี
3 โพสต์ - 3 ผู้เขียน
sl1000/ไม่จำกัดครั้ง อมสดทุกอย่างไม่ก้น รับงานแถวอมตะนครชลบุรี. ... นักศึกษาไซด์ไลน์, เด็ก สาว ไซด์ไลน์, นักศึกษาขายบริการ, หาสาวไซด์ไลน์, sideline, สยามไซด์ไลน์ ... โทร 080-6450272 รับงาน9.00-16.00 ทุกวันคะ ...

ขายค	
ขายคอนโด	6,020,000 ผลการ
ขายค่ะ มีรูป	30,900,000 ผลการ
ขายค่ะ ร้อนเงิน	10,700,000 ผลการ
ขายคอนโดมือสอง	449,000 ผลการ
ขายคอมพิวเตอร์	3,420,000 ผลการ
ขายค่ะ 1500	1,020,000 ผลการ
ขายค่ะ มีห้อง	16,300,000 ผลการ
ขายค่ะ อมสด	1,460,000 ผลการ
ขายคอนแทคเลนส์	563,000 ผลการ
ขายค่ะ หาค่าเทอม	322,000 ผลการ

thaiengine

ผลร้าย ?

- จากการที่เราควบคุมไม่ได้ ทำให้เกิดผลร้ายทางสังคมมากมายจนปรากฏข่าวร้ายในสื่อต่างๆ บ่อยครั้ง

ข้อสังเกต 1

ก่อนถึงข้อเสนอแนะ มีข้อเท็จจริงบางข้อที่เกี่ยวข้อง

- **มันเป็นไปได้ยาก**ที่จะใช้มนุษย์ ค้นหาเว็บปริมาณมากๆ (หลายล้าน) เพื่อแยะแยะเนื้อหาที่ต้องการในเวลาเท่าทันต่อเหตุการณ์ **แต่มันเป็นไปได้**ที่จะใช้ คอมพิวเตอร์ช่วยหาเป้าหมายที่น่าสงสัย และใกล้เคียงจริงๆ
- ลองพิจารณาง่ายๆ เด็กและเยาวชนเข้าถึงข้อมูลเหล่านี้ด้วยวิธีใด เราก็ใช้วิธีเดียวกัน กำหนดเป้า (เบื้องต้น)ในการหาเว็บเป้าหมาย
  - คนส่วนใหญ่ใช้ public search engine เช่น google, yahoo ในการค้นหาข้อมูลใน internet ทั่วไป
- สมมุติว่าถ้าต้องการซื้อบริการทางเพศ สำหรับคนทั่วไป(ที่ไม่ใช่ search engineer)
  - อาจเริ่มต้นด้วย keyword ง่ายๆ เช่น “ขายค่ะ” หรือ “ขายค่ะ นักศึกษา” ก็จะพบ “เว็บเป้าหมาย” ปะปนอยู่จำนวนหนึ่ง(ประมาณ 1-2 รายการในหน้าแรกๆ) และ**เกือบทั้งหมด**ไม่ถูกปิดกั้น
  - หลังจากนั้นส่วนใหญ่ก็จะหา keyword ที่มีประสิทธิภาพมากขึ้นได้ เช่น “แตกปาก อมสด ไม่ก้น” เป็นต้น ซึ่งเว็บเป้าหมายที่พบก็จะมากขึ้น(อาจถึง 10 รายการในหน้าแรกๆ)
  - ดังนั้น การค้นหาโดยคนทั่วไปโดยใช้ search engine ปกติเป็นเครื่องมือสามารถทำได้ง่ายๆ ไม่ยุ่งยากมากนัก นอกจากนี้ จากการทดลองติดต่อดู พบว่าสามารถติดต่อซื้อบริการได้จริง เป็นจำนวนมาก

ข้อสังเกต 2

- ปริมาณข้อมูลที่อยู่ในข่ายที่เราสนใจ คือข้อมูลในประเทศไทย (อาจรวมถึงข้อมูลภาษาไทยจากต่างๆ ประเทศด้วยก็ได้) นั้น เล็กกว่าปริมาณข้อมูลในโลกทั้งหมด
- กลุ่มผู้ใช้ที่เราสนใจมากอันดับแรกๆ คือคนที่เราต้องการปกป้องคือกลุ่มเด็กและเยาวชน นั้น ก็มีขนาดเล็กกว่าคนไทยทั้งหมด และส่วนใหญ่ใช้ internet จากบางสถานที่เท่านั้น
- ดังนั้น ปริมาณข้อมูลที่เราต้องจัดการจริง **มีไม่มากเกินไป** และสามารถ**จัดลำดับความสำคัญ**ในการบริหารจัดการได้

ข้อเสนอ

1. สร้างระบบเพื่อค้นหาเว็บไม่พึงประสงค์ โดยใช้เนื้อหาในการแยกแยะเว็บเป้าหมาย แล้วสร้าง **blocked-list** ขึ้นมา

2. เมื่อได้ **blocked list** แล้ว ก็จัดการกระจายข้อมูลไปตาม **router, proxy server, firewall** ต่างๆ เพื่อปิดกั้นเครือข่ายกลุ่มผู้ใช้ **internet** เป้าหมาย

การแยกแยะโดยใช้ “เนื้อหา” - 1

- ผู้เขียนขอยกตัวอย่างข้อความที่ใช้ในการเสนอขายบริการทางเพศ

  ขายค่ะ สส. 32-25-33 อายุ 22 นน 45 ราคา 1200/1 1700/2
- จะเห็นว่ามีรูปแบบที่มนุษย์รับรู้ได้ง่าย แต่สามารถสลับถ้อยคำหรือแทรกวรรค หรือ ใช้ นน = นน. = น้ำหนัก
- แม้ว่าเป็นเรื่องยากสำหรับโปรแกรมทั่วไป แต่ก็เป็นเรื่องที่สามารถทำได้สำหรับระบบที่มีความสามารถด้วยภาษาศาสตร์อย่าง ThaiEngine ซึ่งสามารถประมวลผลคำไทยได้หลายแบบ
- การย้าย server ไปที่อื่นๆ ก็ยังคงไม่ทำให้เนื้อหาเปลี่ยน การเปลี่ยนเนื้อหาทำได้ยากและมีต้นทุนสูง
- การสื่อสารระหว่าง ผู้ขายบริการ กับผู้ใช้บริการ ก็ยังคงต้องใช้ภาษาไทย-อังกฤษเป็นหลัก ไม่สามารถสร้างภาษาใหม่ขึ้นมาสื่อสารกันได้

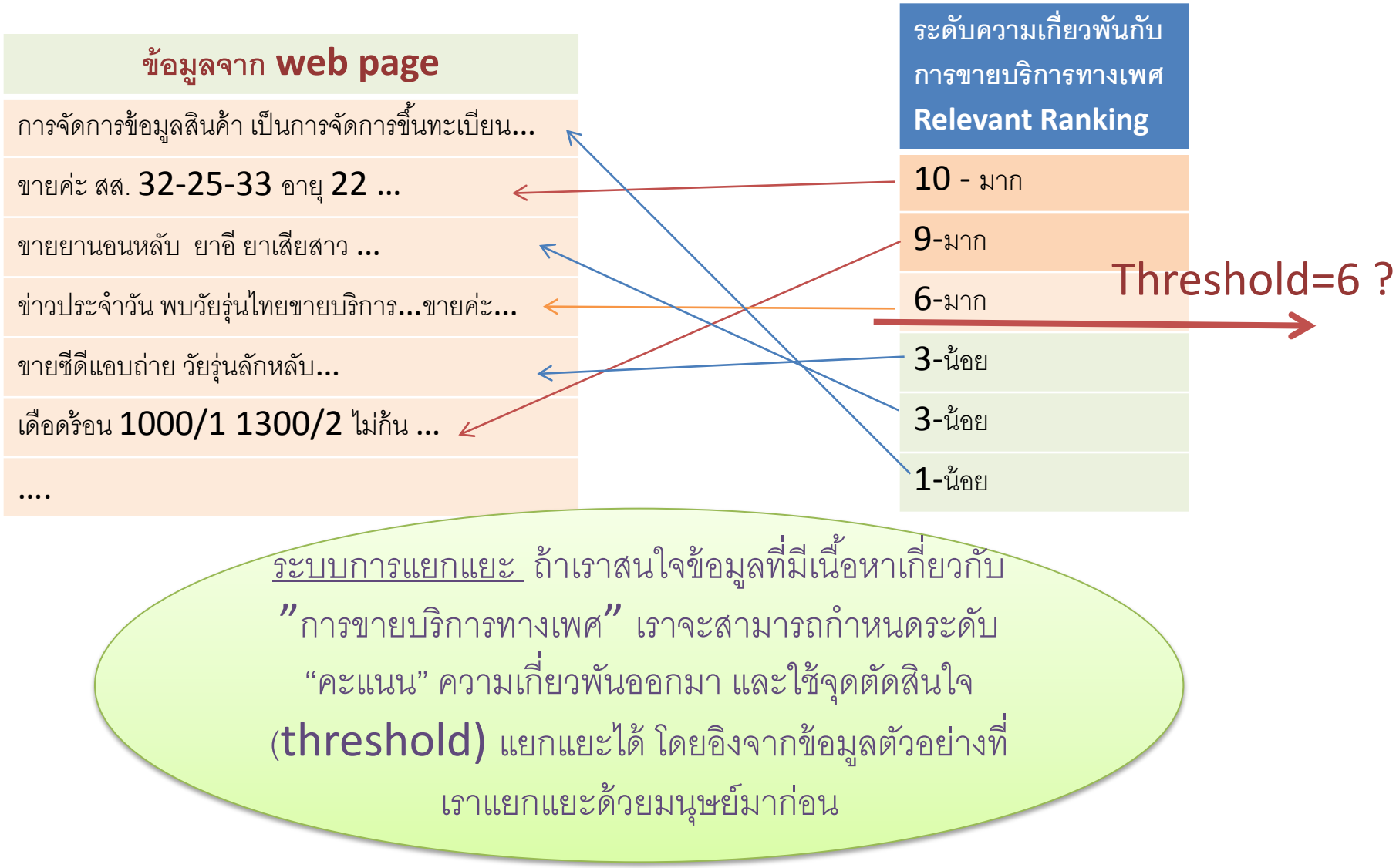
การแยกแยะโดยใช้ “เนื้อหา” - 2

- นอกจากนี้ระบบยังสามารถตรวจสอบรูปแบบการออกเสียงที่บางคนชอบใช้กัน เช่น

  อยาก หาเงินใช้ ครายว่างแอดมาคุยกานน๊ะ พันห้าร้อยบาท ภายนาย สาม ชม กี่ครั้งก็ได้ค่ะ สนจัยโท ศูนย์แปดสี่ หกสี่แปด หกเจ็ดสองเจ็ด พักอยู่ที่ลาดพร้าวค่ะ ***atlantise****หนึ่งสอง****@windowslive.com*** *กรุงเทพมหานคร*

- ระบบที่เราใช้สามารถตรวจสอบความเหมือนในเชิงการออกเสียงที่เรียกว่า “พ้องเสียง” ได้ด้วย เช่น จัย=ใจ คราย=ใคร
- หรือการใช้เสียงอ่านแทนตัวเลข เช่น ศูนย์แปดสี่ หกสี่แปด หกเจ็ดสองเจ็ด = 084-6486727

Clustering โดยใช้ Relevant Ranking Score



คะแนนความเกี่ยวพัน

Basic Relevant ranking score

เราสามารถสร้างคะแนนความเกี่ยวพันจาก

- จำนวนคำที่ตรงกับ keyword
  - ตรงแบบ สลับคำ จะมีคะแนนน้อยกว่า ตรงกันเป๊ะ เช่น ถ้า keyword คือ “ไม่อมสด” เว็บที่มีข้อความ “สดใส...ไม่กินลูกอม” จะมีคะแนนน้อยกว่าเว็บที่มีข้อความ “...ไม่ก้น ไม่อมสด...”
- จำนวนคำที่มีความหมายเหมือน(synonym)กับ keyword
  - เช่น กำหนดให้ “อม” = “โม้ก” เป็นต้น(ThaiEngine มีระบบการตรวจสอบ synonym)
- จำนวนคำที่มีเสียงพ้องกับ keyword
  - เช่น “ก้น” = “กน” = “gon” (ThaiEngine มีระบบตรวจสอบการพ้องเสียงไทย-อังกฤษ)

คะแนนความเกี่ยวพันรวม

สร้างคะแนนความเกี่ยวพันโดยใช้ชุดของ keyword แบบ positive และ negative

- Positive keyword หมายถึง keyword ที่สามารถดึงเว็บเป้าหมายได้ดี
- Negative keyword หมายถึง keyword ที่สามารถตัดเว็บที่ไม่ใช่เป้าหมายออกจากผลที่ได้จาก positive keyword

Set of Positive Keyword

Keyword	Threshold	Weight
ขายค่ะ	20	2
ไม่ก้น	20	3
เดือดร้อน	10	1

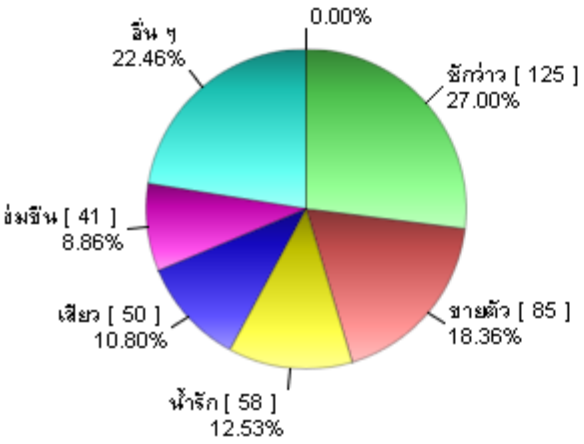
ใช้ในการค้นหาเป้า

Set of Negative Keyword

Keyword	Threshold	Weight
บทสัมภาษณ์	20	2
รัฐมนตรี	20	2

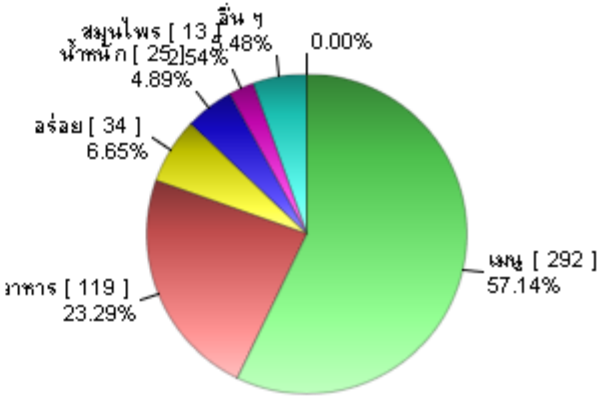
ใช้ตัดสิ่งที่ไม่ใช่ออกไป

คะแนนรวม = ผลรวมของคะแนนที่ได้จาก ทุกๆ positive keywords ที่เกิน threshold **ลบ** ผลรวมของคะแนนที่ได้จาก ทุกๆ negative keywords ที่เกิน threshold



Related Keyword UnRelated Keyword

Related Keyword	UnRelated Keyword
ลงลิ้น	บำรุงผิว
ข่มขืน	อาหาร
ขายหี	พระเครื่อง
น้ำเงี่ยน	ฟุตบอล
น้ำรัก	อิสลาม
ขายตัว	ล้อแม็กซ์
ชักว่าว	สมุนไพร
เงี่ยน	อร่อย
เสียว	บอล



Related Match All : 463

UnRelated Match All : 511

ค่าเบี่ยงเบน Related	15
ค่าเบี่ยงเบน UnRelated	10
Threshold	15

GO

Score :: 1835

VIEWING A THREAD - ข่มขืน

...โพสต์ได้) -> เล่าเรื่องเสียว -> VIEW THREAD YOU ...(LOGON | REGISTER) ข่มขืน JUMP TO PAGE : 1 NOW ...โพสต์ได้) -> เล่าเรื่องเสียว MESSAGE FORMAT FLAT ...:53 (#7425) SUBJECT: ข่มขืน VETERAN POSTS: 295 ...แต่งงานกันใหม่ๆ ก็ร่วมรัก...ตัว...ตัว...ตัว...ตัว...ลง...ตัว...ขืน...น้ำ...น้ำ

Point	Pass
701	

แหล่งที่มา :: www.thaix.org/bbs/forums/thread-view.asp?tid=1275&posts=10

VIEWING A THREAD - ข่มขืน

...โพสต์ได้) -> เล่าเรื่องเสียว -> VIEW THREAD YOU ...(LOGON | REGISTER) ข่มขืน JUMP TO PAGE : 1 NOW ...โพสต์ได้) -> เล่าเรื่องเสียว MESSAGE FORMAT FLAT ...:53 (#7425) SUBJECT: ข่มขืน VETERAN POSTS: 294 ...แต่งงานกันใหม่ๆ ก็ร่วมรัก...ตัว...ตัว...ตัว...ตัว...ลง...ตัว...ขืน...น้ำ...น้ำ

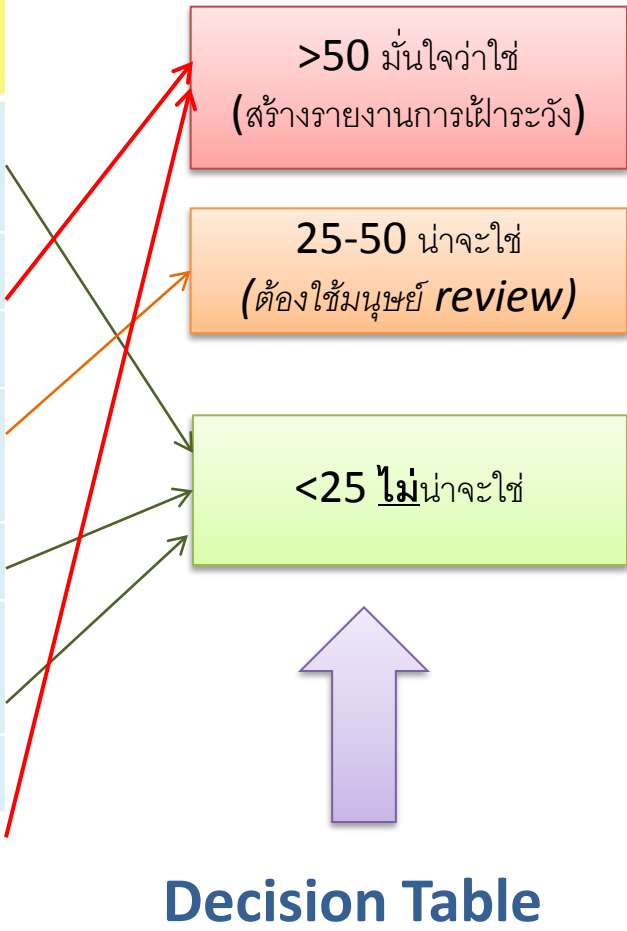
Point	Pass
685	

แหล่งที่มา :: www.thaix.org/bbs/forums/thread-view.asp?tid=1275&posts=10&start=1

VIEWING A THREAD - ข่มขืน

การตัดสินใจ

ข้อมูลจาก web page	Positive score	Negative score	คะแนนรวม
การจัดการข้อมูลสินค้า เป็นการจัดการขึ้นทะเบียน...	5	-3	2
ขายค่ะ สส. 32-25-33 อายุ 22 ...	61	0	61
ขายยานอนหลับ ยาอี ยาเสียสาว ...	28	0	28
ข่าวประจำวัน พบวัยรุ่นไทยขายบริการ... ขายค่ะ...	38	-16	22
ขายซีดีแอบถ่าย วัยรุ่นลักหลับ...	23	0	23
เดือดร้อน 1000/1 1300/2 ไม่ก้น ...	52	0	52
....			



Decision Table

วัฏจักรการแยกแยะ

C

Keyword Tool

search

Searcher
(IR System)

B

Indexer

Blocked list

ตัวสร้าง
Blocked list

update

Clustering
Profile

Improvement
Cycle

เว็บ
เป้าหมาย

A

เว็บอื่นๆ

น่าจะเป็นเว็บไม่พึง
ประสงค์ แน่ใจ

ยืนยัน

เว็บต้องสงสัย

Clustering
Agent

ใช้มนุษย์
พิจารณา

อาจเป็นเว็บ ไม่พึง
ประสงค์
(ไม่แน่ใจ)

D

E

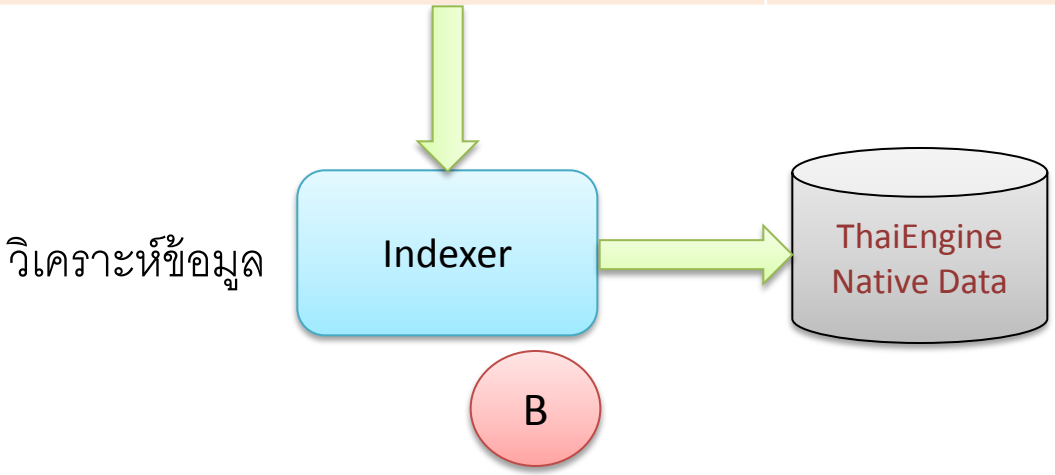
A,B แยกแยะข้อมูลเริ่มต้นโดยมนุษย์ (เพื่อสร้างบทเรียนเริ่มต้นให้ระบบ)

A

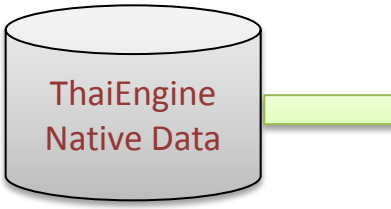
แยกแยะ

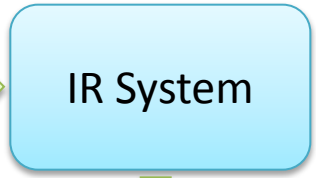
โดยมนุษย์

ข้อมูลเริ่มต้น	เกี่ยวพันหรือไม่
การจัดการข้อมูลสินค้า เป็นการจัดการขึ้นทะเบียน...	N
ขายค่ะ สส. 32-25-33 อายุ 22 ...	Y
ขายยานอนหลับ ยาอี ยาเสียสาว ...	N
ข่าวประจำวัน พบวัยรุ่นไทยขายบริการ...ขายค่ะ...	N
ขายซีดีแอบถ่าย วัยรุ่นลักหลับ...	N
เดือดร้อน 1000/1 1300/2 ไม่ก้น ...	Y
....	

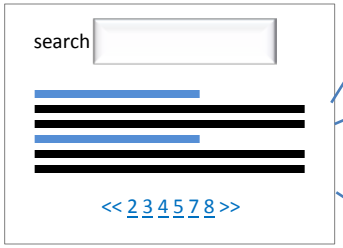


C-การใช้ Thai Search Engine เป็น keyword tool ในการสร้าง Keyword





Keyword Tool



หาเงินใช้ กี่ครั้งก็ได้

Positive - Keyword ที่สามารถดึงข้อมูลที่ “ใช่” ได้อย่างมีประสิทธิภาพ

รัฐมนตรี สัมภาษณ์

Negative - Keyword ที่สามารถดึงข้อมูลที่ “ไม่ใช่” ได้อย่างมีประสิทธิภาพ

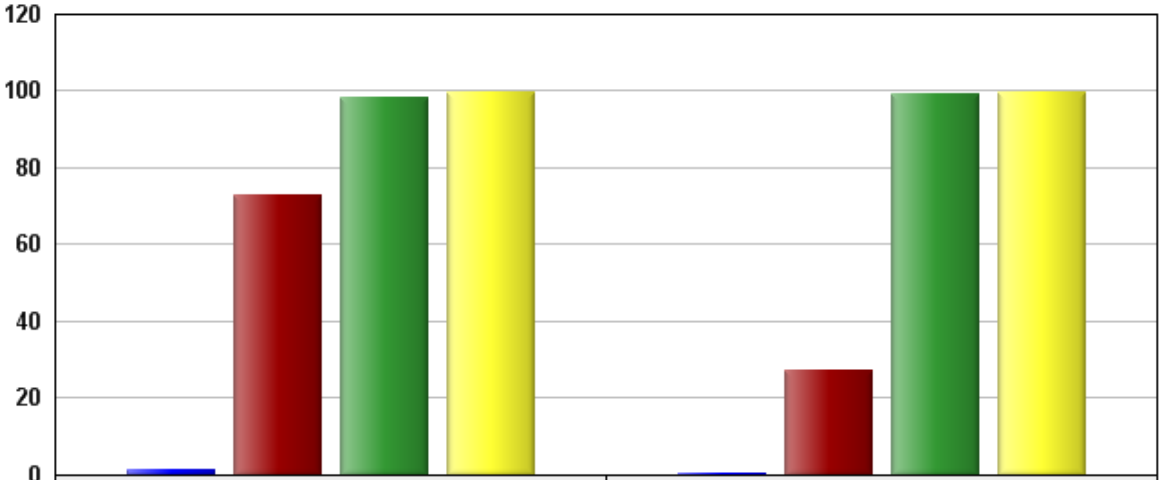
หาเงินใช้

Reject Keyword ที่ดึงข้อมูลทั้งที่ใช่และไม่ใช่ได้ใกล้เคียงกัน

อยาก หาเงินใช้ คราวว่างแอดมาคุยกานน๊ะ พันห้าร้อยบาท ภายนาย สาม ชม กี่ครั้งก็ได้ค่ะ สนจัยโท ศูนย์แปดสี่ หกสี่แปด หกเจ็ดสองเจ็ด พักอยู่ที่ลาดพร้าว ค่ะ ***atlantise****หนึ่งสอง****@windowslive.com*** *กรุงเทพมหานคร*

Find ลงลิ้น SEARCH

Threshold 10



	sex_board	sex_story
Match	16(1.6%)	6(0.6%)
Contribution	72.73%	27.27%
Not Match	984(98.4%)	994(99.4%)
Total	1000(100%)	1000(100%)

จำนวนทั้งหมด 57

VIEWING A THREAD - ข่มขืน

▶...ใจมาก รู้สึกว่าตัวเองนั่งลงกับโซฟาเมื่อไหร่ไม่รู้ ...ไม่มีสติเขาให้ดิฉันนอนลงน้ำตายังคงไหลไ ม่หยุด ...แขนดิฉันไว้เบาๆ โน้มตัวลงหอมแก้มดิฉัน มือเขา...เองแต่เขาก็คว้าแขนดิฉันลงข้างกายหน้าอกของเขา ...ไม่ทันระวังจึงถูกประกบจูบ ลิ้นของเขาพุงวาบเข้ามากระดิกสะกิดลิ้นของดิฉัน ...ลง...ลิ้น...ลิ้น ลง...ลง...ลง...ลง...ลิ้น...ลง...ลง...ลง...ลิ้น...ลิ้น...ลิ้น...ลิ้น

Point	Pass
76	

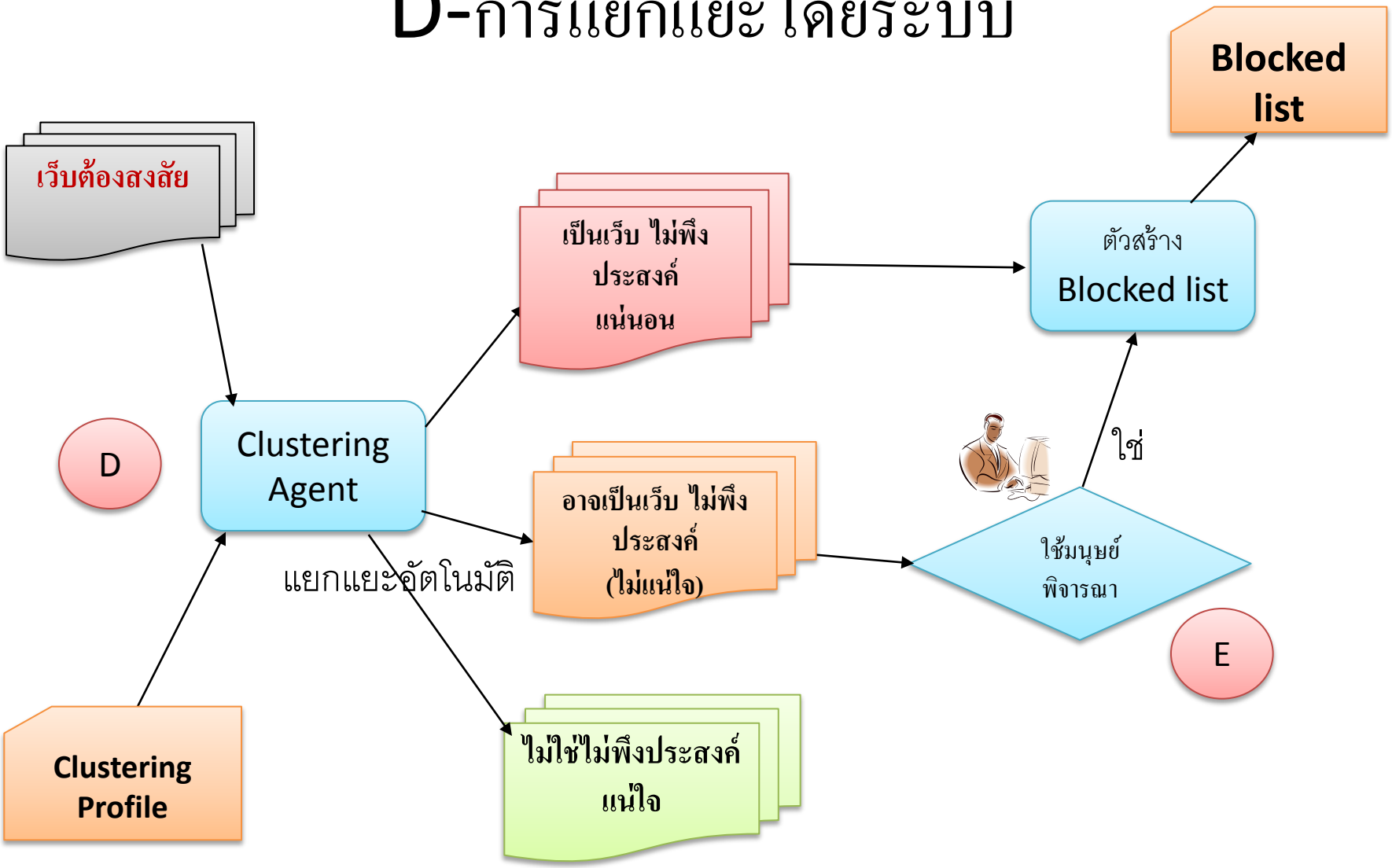
แหล่งที่มา :: www.thaix.org/bbs/forums/thread-view.asp?tid=1275&posts=10

VIEWING A THREAD - ข่มขืน

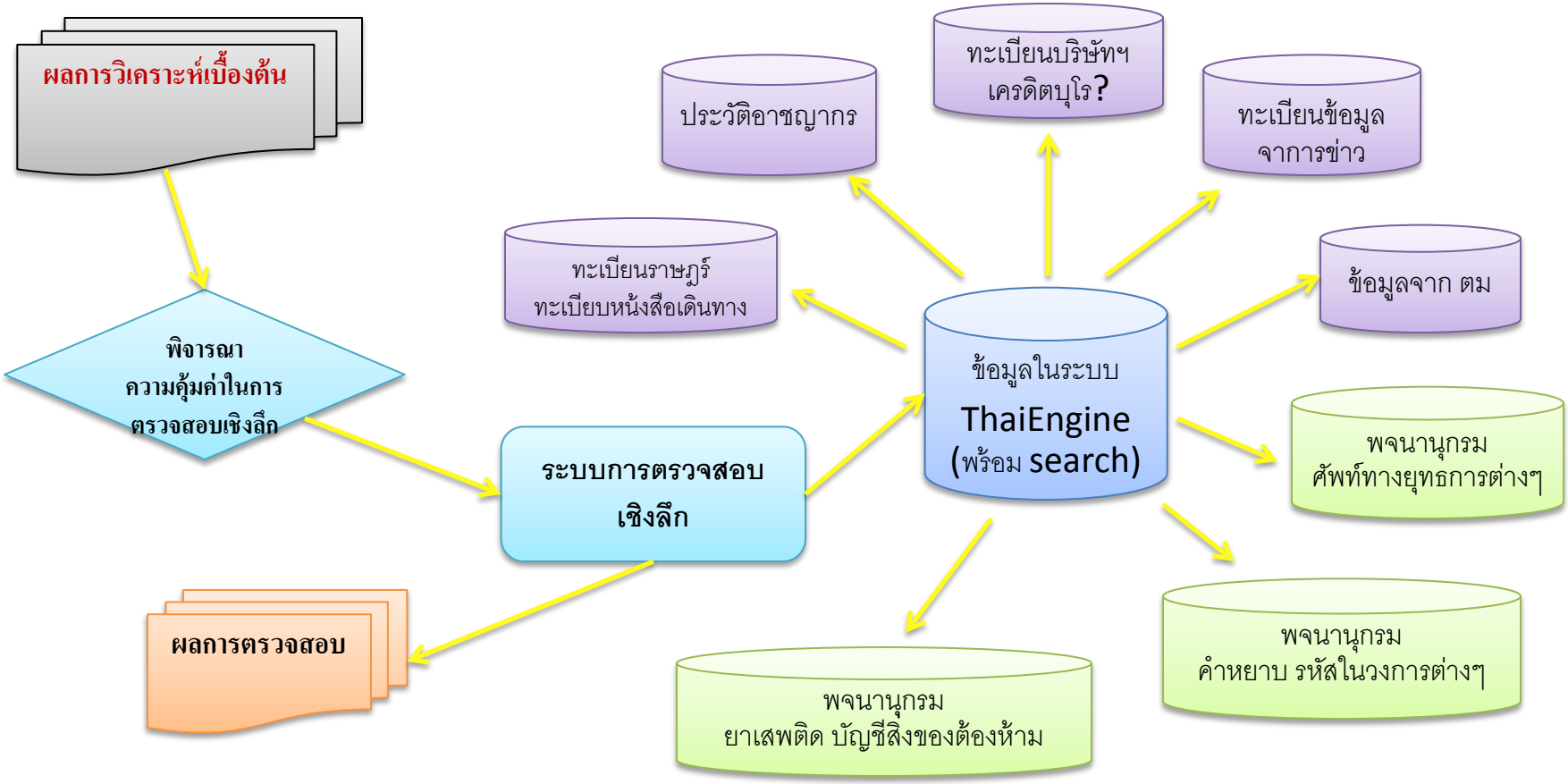
▶...ใจมาก รู้สึกว่าตัวเองนั่งลงกับโซฟาเมื่อไหร่ไม่รู้ ...ไม่มีสติเขาให้ดิฉันนอนลงน้ำตายังคงไหลไ ม่หยุด ...แขนดิฉันไว้เบาๆ โน้มตัวลงหอมแก้มดิฉัน มือเขา...เองแต่เขาก็คว้าแขนดิฉันลงข้างกายหน้าอกของเขา ...ไม่ทันระวังจึงถูกประกบจูบ ลิ้นของเขาพุงวาบเข้ามากระดิกสะกิดลิ้นของดิฉัน ...ลง...ลิ้น...ลิ้น

Point	Pass
76	

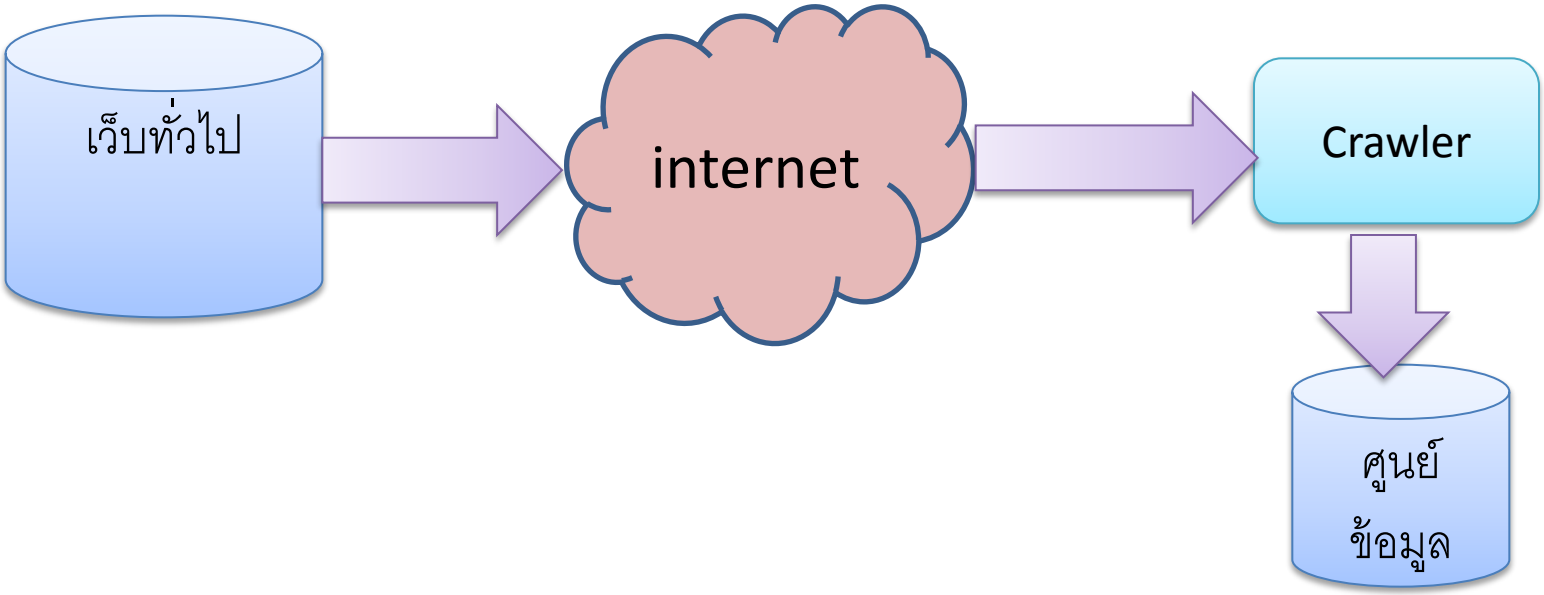
D-การแยกแยะโดยระบบ



การวิเคราะห์เชิงลึก



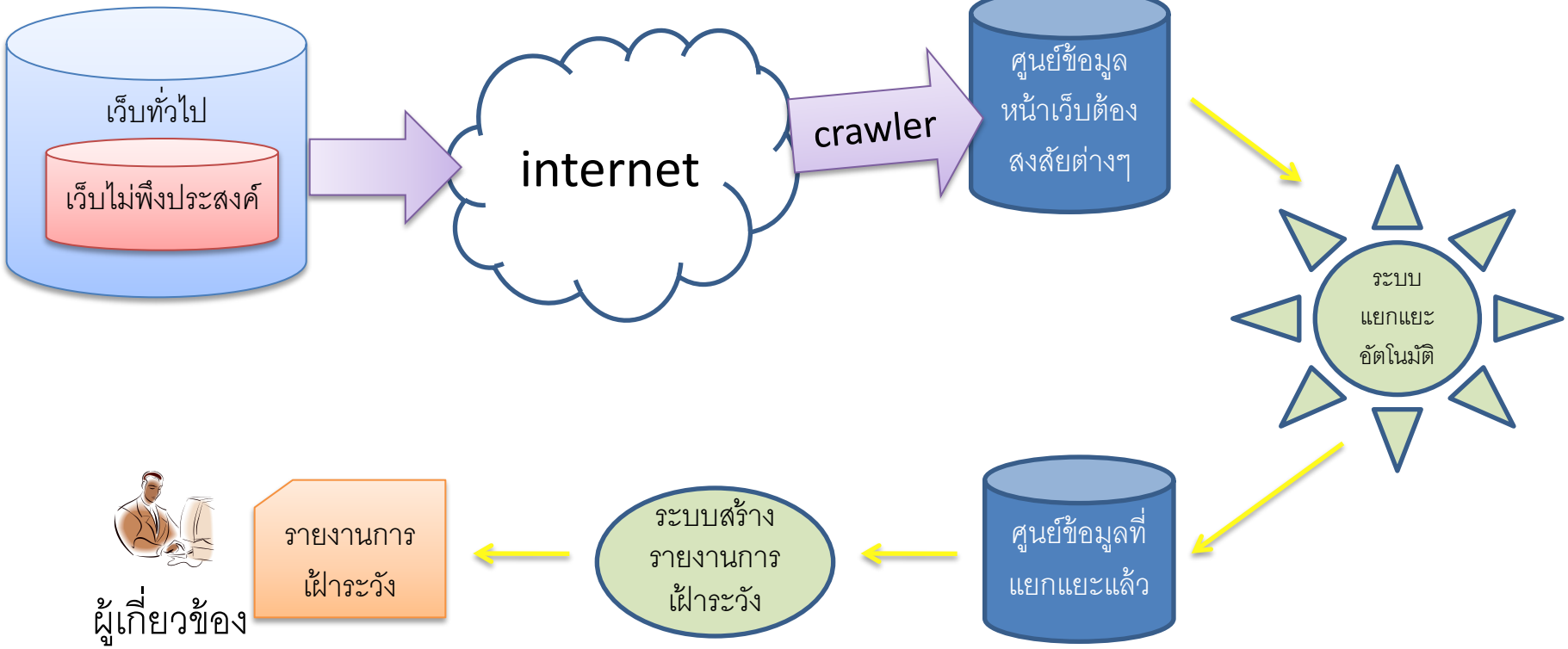
Spider หรือ crawler



- เป็นระบบที่ใช้ในการดึงข้อมูลใน internet มาเก็บไว้ภายใน (เพื่อทำการวิเคราะห์และใช้ประโยชน์อื่นๆ ต่อไป)
- ปกติเราสามารถใช้ crawler ที่เป็น open source ในการทำงานทั่วไปได้ (เช่น httrack หรือ wget) ในกรณีนี้เราควรสร้าง crawler ขึ้นมาเอง

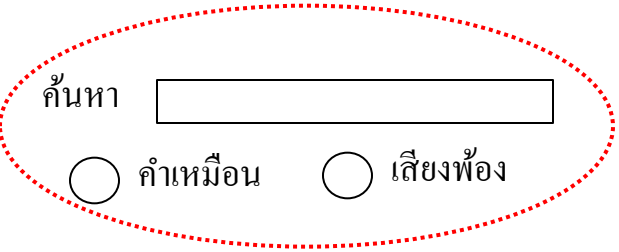
การใช้งาน แบบการเฝ้าระวัง

ในบางกรณี เราอาจไม่ต้องการ blocked list แต่ต้องการให้ผลออกมาในรูปของ รายการการเฝ้าระวังแทน

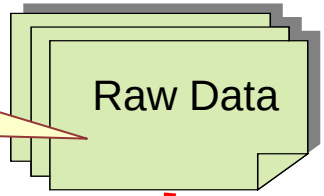


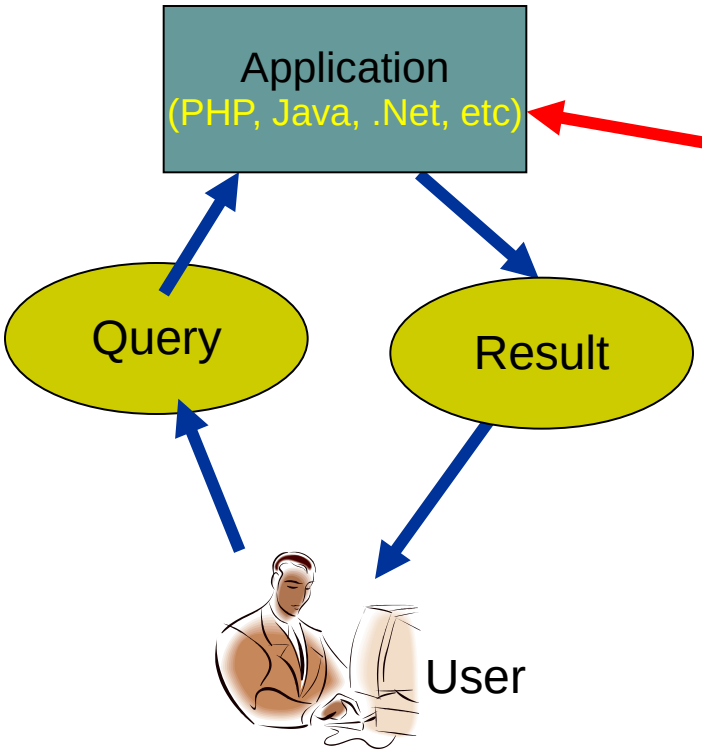
thaiengine

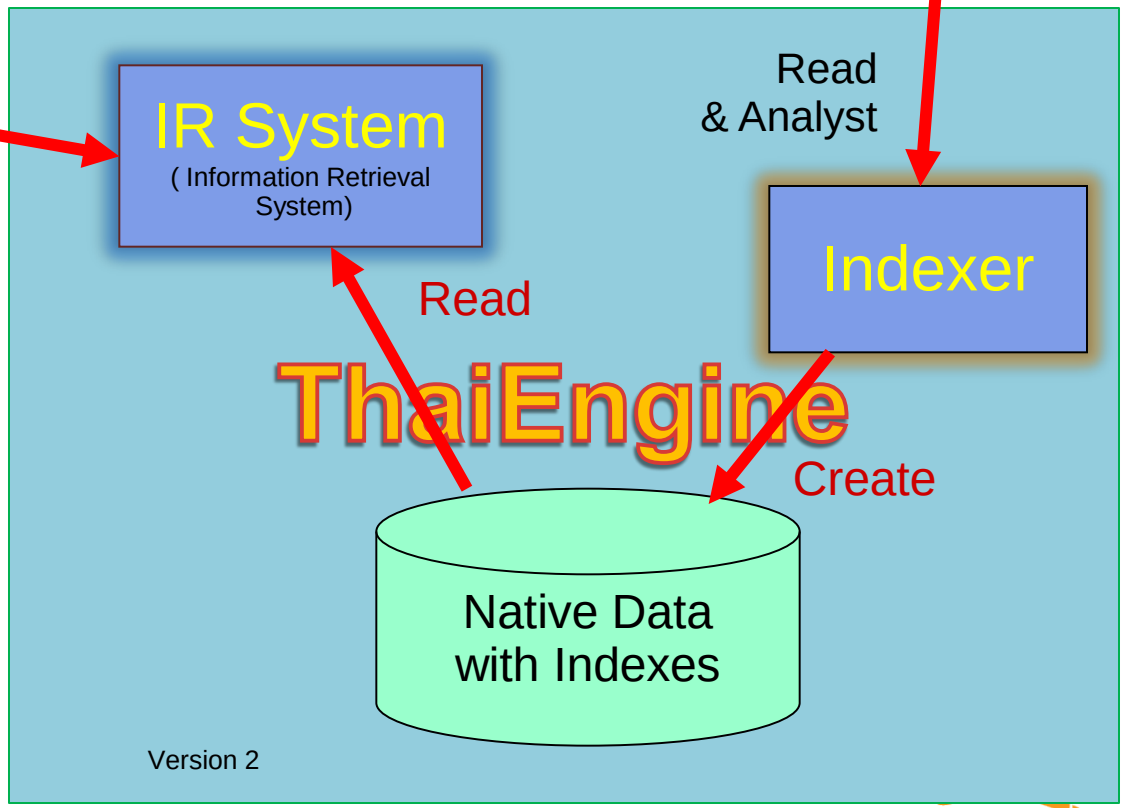
Search Engine Architecture



Such as
DBMS Data (MySQL, Oracle, SQLServer),
***HTML File**, Text files, etc.*







Nov 4th, 2008

thaiengine

Full-text search

- ทำงานได้ดีกับข้อมูลที่ได้จากระบบฐานข้อมูลโดยเฉพาะ
- เป็นระบบ 2 ภาษาคือ ไทย-อังกฤษ ใช้พยางค์เป็นหน่วยย่อยที่สุดในการประมวลผล การออกแบบใช้สมมุติฐานดังนี้
  - ข้อมูลที่ใช้อักขระไทย*(ก ข ค ...)* อาจเป็นได้ทั้งภาษาไทย ภาษาอังกฤษ*(เช่น คอมพิวเตอร์ นิวยอร์ค)* หรือภาษาอื่น*(เช่น เส้าหลิน ลี่ผี)* ก็ได้
  - ข้อมูลที่ใช้อักขระลาติน *(a b c ...)* อาจเป็นภาษาอังกฤษ ภาษาอเมริกัน ภาษาไทย*(เช่น **sukhumwit, baimai**)* หรือภาษาอื่นๆ ก็ได้
  - คำ/พยางค์ไทย ที่เราพบในข้อมูล อาจพบอยู่ในพจนานุกรมไทย หรือไม่ก็ได้
  - คำ/พยางค์ภาษาอังกฤษ อาจพบหรือไม่พบในพจนานุกรมอังกฤษ/อเมริกันก็ได้

ประมวลผลโดยใช้พยางค์ทั้งไทย-อังกฤษ

- มีระบบการตัดพยางค์ เช่น ตัดพยางค์ - > ตัด-พ-ยางค์ หรือ samudprakan➔sa-mud-pra-kan
  - จะเห็นว่า ThaiEngine สามารถตัดพยางค์ภาษาอังกฤษได้ด้วย
- สามารถค้นหาข้อมูลต่างๆ ได้อย่างเป็นธรรมชาติ เช่น
  - หา “ใบไม้” จะพบ “ใบไม้” , “ไม้ใบ”, “ไม้xxxใบ” แต่จะไม่พบ “ใบ้ไม้”
  - หา “samud” จะพบ “samudprakan” “samudsongkram”
  - สามารถบีบให้เอาผลลัพธ์เฉพาะที่เป็น exact match ได้ เช่น จากตัวอย่าง “ใบไม้” สามารถระบุว่าไม่เอา “ไม้ใบ”, “ไม้xxxใบ” ได้
- จะเห็นว่าทำงานได้ดีกับ คำที่ไม่พบในพจนานุกรม(ไทย-อังกฤษ)ซึ่งมีสัดส่วนสูงมากในข้อมูลภาษาไทยในปัจจุบัน

การค้นหาแบบพ้องเสียง

- Thai Engine มีระบบการแปลงพยางค์ทั้งภาษาไทยและภาษาอังกฤษ ไปเป็นรหัสเสียงแบบ phonetic (ซึ่งเป็นงานวิจัยเฉพาะของผู้ผลิต Thai Engine) ผู้ใช้จึงสามารถค้นหาแบบพ้องเสียงได้หลายรูปแบบ ที่ไม่เคยปรากฏมาก่อนในวงการ IT ไทย เช่น
  - ค้น “สมศักด์” จะพบ “สมสัก” “สมศักด์” “สมศักดิ์” “somsak” “ซมซัก”
  - ค้น “ใบไม้” จะพบ “ใบไม้” “ไบไม” “บัยมัย” “bimi” “baimai” “baimi” “bi mai”
  - ค้น “สมชาย” จะพบ “สมชาย” “somchai”

หมายเหตุ ThaiEngine ใช้ text-to-phonetic ไม่ใช่ soundex

คำเหมือน (Synonym)

- คำเหมือน หมายถึง กลุ่มของ “วลี” ที่มีความหมายเหมือนกัน เช่น
  - สถานีตำรวจ = สน.
  - โรงเรียน = รร.
  - ไทย = Thai
  - บ้าน = เรือน, house, home
- ผู้ใช้สามารถค้นหาโดยให้ระบบนำเอาข้อมูลที่เป็นคำพ้องออกมาด้วยได้
- Thai Engine มีฐานข้อมูลคำเหมือนต่างๆ ไว้ให้ผู้ใช้เลือก เช่น คำพ้องจากพจนานุกรม อังกฤษ-อังกฤษ ไทย-ไทย ไทย-อังกฤษ
- ผู้ใช้สามารถเพิ่มเติมข้อมูลคำพ้องภายในองค์กรเข้าไปได้เอง

ThaiEngine ไม่ได้ใช้ stem algorithm

ระบบการค้นหาแบบส่วนหนึ่งของพยางค์/คำ

- ในข้อมูลบางอย่าง เช่น หมายเลขโทรศัพท์(0816996269, 029504381) หรือ รหัสสินค้า เช่น (ABC9383, 94883048) ซึ่งถือเป็น 1 พยางค์ ปกติแล้วผู้ใช้จะไม่สามารถกรอกรายละเอียดได้ทุกตัวอักษร
  - เป็นประโยชน์กับ call center ที่ต้องรับเรื่องทางโทรศัพท์
- Thai Engine มีระบบการทำ index แบบพิเศษซึ่งสามารถทำสร้าง index แบบส่วนหนึ่งของพยางค์ได้ เช่น
  - นำเอาตัวอักษรใดๆ ที่อยู่ติดกัน ตั้งแต่ 4 ตัวอักษร ถึง 7 ตัวอักษรไปทำ index ด้วย
  - ซึ่งในกรณี 0816996269 ผู้ใช้จะสามารถค้นพบข้อมูลด้วย keyword ดังนี้ 0816, 8169, 1699, …, 6269, 08169, 81699, …, 6996269

ข้อมูลที่ไม่ใช่ text

- ในการใช้งานทั่วไปบางกรณีเราจำเป็นต้องใช้เงื่อนไขในการค้นหาที่ไม่ใช่ text ควบคู่ไปกับ full text search เช่น
  - ต้องการหาหนังสือที่มีคำว่า “ภาษาไทย” และมีราคาต่ำกว่า 200 บาท
  - ต้องการภาพยนตร์เรื่อง “จดหมายรัก” ที่ฉายระหว่าง 13:00 ถึง 18:00 วันที่ 1 กุมภาพันธ์ 2550
- จะเห็นว่าใน full text search ทั่วไปจะตรวจสอบเงื่อนไขด้านราคา หรือวัน-เวลาไม่ได้
- Thai Engine สามารถทำงานได้กับการค้นหาดังกล่าว โดยที่มีชนิดของข้อมูลที่ support หลายอย่าง เช่น
  - numeric, Date, Date-time, time เป็นต้น

Relevance Ranking (การเดาใจ)

- ในการคาดเดาว่า ผลลัพธ์รายการใดน่าจะตรงกับความต้องการของผู้ใช้มากที่สุดนั้น Thai Engine เรียกค่าความตรงใจนี้ว่า relevance ranking โดยที่ใช้ในการเรียงลำดับผลลัพธ์ออกมา เพื่อให้รายการที่มี relevance ranking สูง(ตรงใจมาก) จะออกมาในลำดับต้นๆ
- เจ้าของระบบสามารถกำหนดสูตรของ relevance ranking ได้ ดังนี้
- ในการคำนวณ relevance ranking จะใช้ค่าต่อไปนี้เป็น parameter ในการคำนวน
  - จำนวนพยางค์ที่ตรงกับ search string(P1)
  - จำนวนคู่ของพยางค์ที่ตรงกับ search string(P2)
  - จำนวนคู่ของพยางค์ที่ตรงและไม่สลับลำดับ ที่ตรงกับ search string(P3)
  - ความเก่า-ใหม่ของข้อมูล (P4)
  - ระดับความสำคัญของ column(P0)
- สามารถปรับสูตรในการคำนวณได้ เช่น

  Relevance Ranking = P0 * (2*P4+4*P3 + 2*P2 + P1)

สำหรับผู้ที่สนใจใช้ vector space model เราก็สามารถปรับแต่งระบบให้ support ได้เช่นกัน

GEO-Index & GEO Ranking

- สำหรับข้อมูลที่มี พิกัดทางภูมิศาสตร์รวมอยู่ด้วย
- ระบบสามารถ เพิ่มเงื่อนไขเกี่ยวกับระยะทางเข้าไปควบคู่กับ keyword ทั่วไปด้วย

  ค้นหา ร้านข้าวมันไก่ ที่อยู่ไม่เกิน 5 กม จากจุดที่กำหนด
- ส่งผลการค้นหาไปแสดงในระบบแผนที่
- ใช้ระยะใกล้ไกลจากจุดที่กำหนด ปรับแต่ง relevance ranking ได้

thaiengine

แนวทางดำเนินการในอนาคต

- สร้างระบบที่ “เข้าใจ” ความหมายที่อยู่ใน ข้อความภาษาไทย
  - สร้าง “ระบบความจำของมนุษย์”
  - นำ “ความจำของมนุษย์” ดังกล่าวไปใช้ในการแยกแยะข้อความที่มีลักษณะกำกวม
  - จัดทำระบบวิเคราะห์ประโยคภาษาไทย เพื่อให้สามารถ แยกแยะ knowledge ออกมาจากข้อความได้
  - ปรับระบบ search engine ให้สามารถ index ด้วย knowledge
  - ปรับปรุง API ของระบบให้สามารถนำไปใช้งานง่ายขึ้น

thaiengine

ระบบที่เลียนแบบความจำของมนุษย์

- จากหลักการของ lexical semantic เราน่าจะเริ่มต้นจากการนำแนวคิด 2 แบบเริ่มมาจัดทำขึ้น
  - แบบจำลองของ Collins และ Quillian(1972) คือสร้าง network ซึ่งมี node ต่างๆ และมี link ระหว่าง node ต่างๆ โดยรวมๆ แล้วมีลักษณะเหมาะกับความรู้ในลักษณะที่เป็น encyclopedia คือเป็น long term memory
  - แบบจำลองของ Miller และ Johnson-Laird(1976) คือ สร้างความสัมพันธ์ระหว่าง ภาษา และ Concept ต่างๆ โดยตัวอย่างของความสัมพันธ์ก็เช่น “isa”, “has-a” “is-in” เป็นต้น

ต้นทุนใน พจนานุกรม

- สิ่งที่มีอยู่แล้ว
  - wordnet dictionary จะเห็นว่าเค้ามีฐานข้อมูลของ “is-a” “has-a” “is-in” อยู่เป็นแสนรายการแล้ว (ภาษาอังกฤษ)
  - พจนานุกรมไทย ฉบับราชบัณฑิตฯ และ Lexitron
  - ข้อมูลจาก อนุกรมวิทาน และ ข้อมูลอื่นๆ ที่จัดหมวดหมู่ของสิ่งต่างๆ
- นำข้อมูลเหล่านี้ ทำเป็นภาษาไทย และอยู่ในรูป XML

การต่อสู้กับความกำกวมในภาษา

- ตัวอย่างของความกำกวม
  - แดงนั่งตากลมอยู่หน้าบ้าน
  - เรือโคลงเพราะโคลงเรือ
- มนุษย์แยกแยะความกำกวมโดยการพิจารณาความหมายของความกำกวมในทุกๆ แบบที่เป็นไปได้ แล้วเลือกแบบที่สอดคล้องกับสถานะการณ์ – ความเป็นไปได้มากที่สุด
- ถ้าเรามีวัตถุประสงค์เพียงแค่ ดึงเอา “ข้อเท็จจริงบางอย่าง” ไปทำ Index ใน search engine หรือแยกแยะว่า “เกี่ยวพัน” กับเนื้อหาที่เราสนใจหรือไม่ ความกำกวมเหล่านี้ ก็บริหารจัดการได้ง่ายขึ้น

การ index ด้วย Knowledge

- ประโยค “กลางดึกวันที่ 13 กพ.นี้ พบศพผู้เสียชีวิตที่พงหญ้าหน้าร้านสะดวกซื้อ”
- ปัจจุบันใน ThaiEngine จะ Index ทุกพยางค์ เช่น

  กลาง-ดึก-วัน-ที่-13-กพ-.-นี้-พบ...
- จะเห็นว่าปัญหาใหญ่คือเรามีปริมาณพยางค์จำนวนมากที่ต้องบริหารจัดการ
- อาจเป็นประโยชน์มาก หากระบบจะสร้าง knowledge ขึ้นมาเพียงว่า
  - ประธาน: มนุษย์ (ซึ่งเป็นสิ่งมีชีวิต-สัตว์เลี้ยงลูกด้วยนม-...)
  - กริยา: เสียชีวิต
  - วันเวลา: 13-กพ-53 ระหว่าง 21:00 ถึง 24:00
- ซึ่งสามารถใช้ควบคู่กับ Index ด้วยระบบเดิมได้

thaiengine

ปรับปรุง API ใน ThaiEngine

- เรียกใช้ผ่าน CGI
- มีภาษาที่เป็น procedural language ผสมกับ คำสั่ง search ที่คล้าย SQL เช่น

  SEARCH *"ใบไม้"* ***IN*** *(Name,Address)*
  AND *"อาหาร"* ***IN*** *(Category)*
  SELECT *ID,Name,Address,sysPageRank*
  FROM *YellowPage*
  WHERE *Province="กรุงเทพ"*
  ORDER BY *sysPageRank* ***DESC***

- หรือ

  Create procedure sp_bp
  @phase text
  ;
  Search @phase in (name,address)
  Select …
  If @phase = "xxx" than

- **…**

thaiengine

Q & A