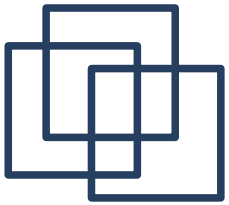
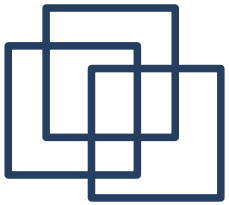


Search Engine ภาษาไทย

วิรัช ศรเลิศล้ำวาณิช

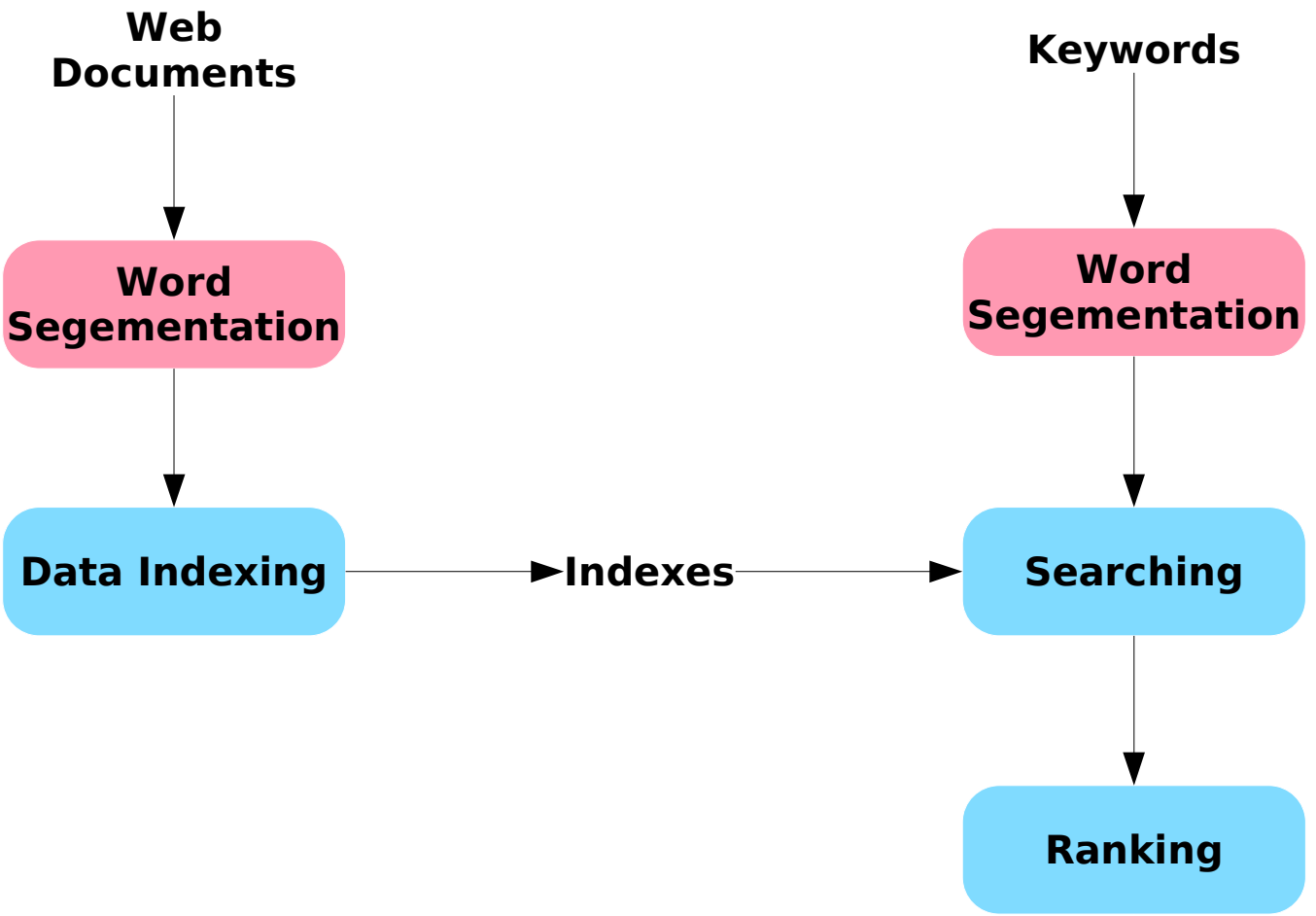
Thai Computational Linguistics Laboratory (TCL), NICT
virach@tcllab.org

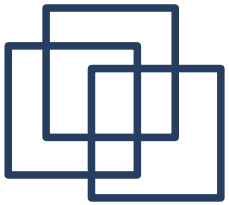


Search Engine สำหรับภาษาอังกฤษ

- Lemmatization
  - work:- work, working, works, worked
  - go:- go, going, goes, went, gone
  - Windows:- WINDOWS, Windows
- Scoring สำหรับการจัดลำดับบทความ
  - ความสำคัญของคำในบทความ เช่น คำที่ปรากฏใน Title, คำสำคัญ (key word), ความถี่ของคำในไฟล์ เป็นต้น
  - ความนิยมของบทความ เช่น จำนวน click ที่เข้าแวะชม, จำนวน link จากคำที่ใช้สืบค้น, page rank, เป็นต้น
- สร้าง Inverted Index File โดยอาศัยโปรแกรม Database

Dictionary-based Search Engine --Architecture--





ความยากในการตัดคำ

- **ความกำกวม**

แบบนอก	->	แบบ | นอก	<> แบ | บน | อก
มีที่นา	->	มี | ที่นา	<> มี | ที่ | นา
ขอบอกขอบใจ	->	ขอบอก | ขอบใจ	<> ขอ | บอก | ขอบใจ
ขนมอบกรอบ	->	ขนม | อบ | กรอบ	<> ขน | มอบ | กรอบ
ร้านข้าวซอยลำดวน	->	ร้าน | ข้าวซอย | ลำดวน	<> ร้าน | ข้าว | ซอย | ลำดวน

- **คำที่ไม่ปรากฏในพจนานุกรม**

นาตาลี	->	นา | ตา | ลี
อยุธยาอะลิอันซ์ซีพี	->	อยุธยา | อะลิอันซ์ | ซี | พี
กาลิเลโอ	->	กา | ลิเล | โอ

- **การจัดเก็บข้อมูลในพจนานุกรม**

  อาทิ บัญชีคำ, หน้าที่ของคำ, เป็นต้น

Search Engine สำหรับภาษาไทย

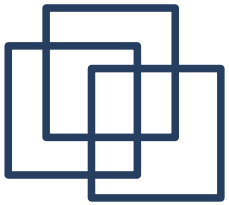
1. | ประวัติ | วัด | มหาธาตุวรวิหาร | และ | วัด | อารามหลวง |
2. | ประวัติ | วัด | มหาธาตุ | วรวิหาร | และ | วัด | อาราม | หลวง |

1. มหาธาตุ — หาไม่เจอ ? ⇐ Partial matching method
2. มหาธาตุวรวิหาร — หาไม่เจอ ? ⇐ Word segmentation + Concatenation (AND)

มหาธาตุ AND วรวิหาร

- เพิ่ม Index โดยอาศัยวิธีการ SIString (Semi-infinite string)
- ตัดคำให้ย่อยที่สุด

Precision ⇨	สถิติคลาดเคลื่อน ⇨	Ranking ไม่ตรง
Recall ⇨	Coverage ⇨	หาไม่พบ



Search Engine
อาศัยความน่าจะเป็นของคำ

ครอบ
ใน
ที่
ห้อง
ทำ
⋮

ครัว ⇨ “ครัว”

ความหลากหลายของคำที่อยู่ข้างเคียง

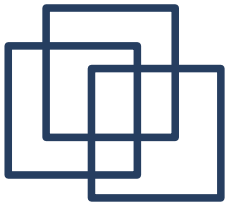
ค + รอบ ⇨ “ครอบ”

ครอบ + ครัว ⇨ “ครอบครัว”

ความบ่อยของการอยู่เคียงข้างกันของอักขระ

Ranking ⇨

- เปรียบเทียบความน่าจะเป็นของอันดับคำ
- Weight ตามค่าความสำคัญของคำ (key word, title, ...) และความถี่ (term frequency)



ความยากในการสืบค้นกรณีภาษาไทย

ครัว

...การสมรสหมู่โดยสหพันธ์ครอบ**ครัว**เพื่อความสามัคคี...

...เปลี่ยนเป็นห้องน้ำ...ห้อง**ครัว**...ห้องรับแขก...

ประชา

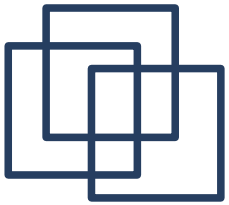
...กอง**ประชา**สัมพันธ์การสื่อสารแห่งประเทศไทย...

...นายแพทย์**ประชา** เป็นประธานคณะกรรมการ...

ธาตุ

...ประวัติวัดมหา**ธาตุ**วรวิหาร...

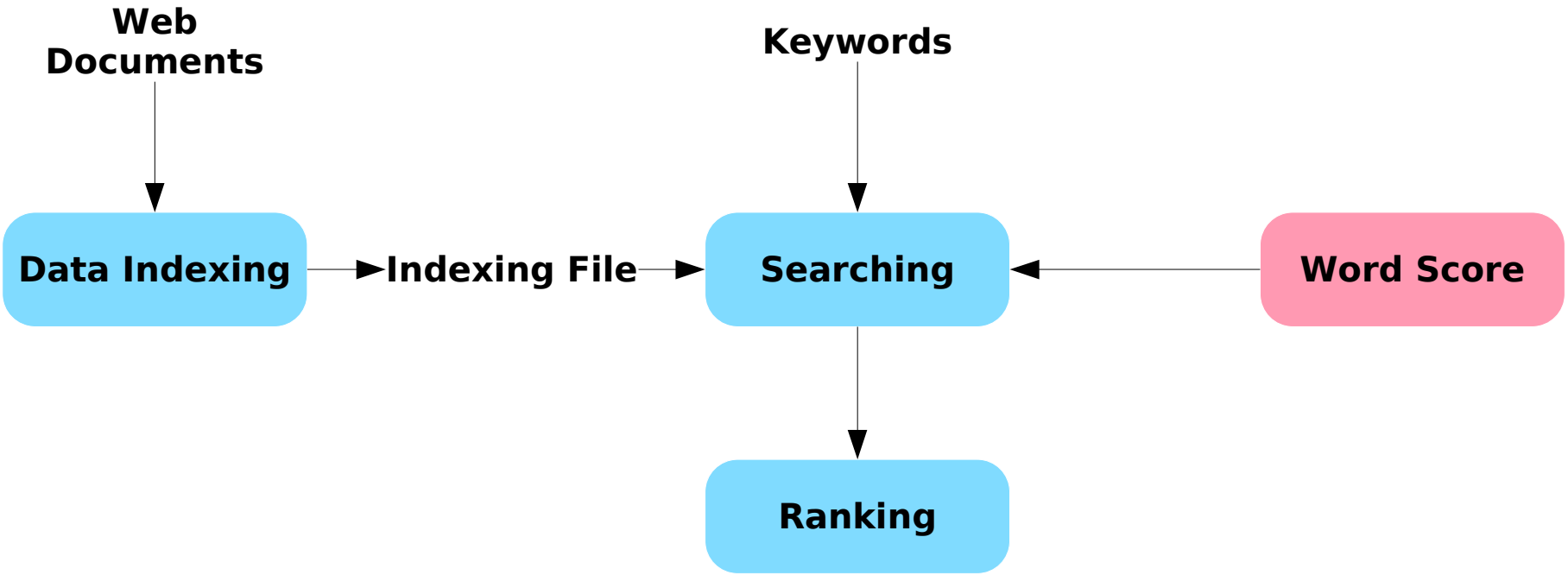
...โปรแกรมช่วยสอนเคมีเบื้องต้น และตาราง**ธาตุ**...

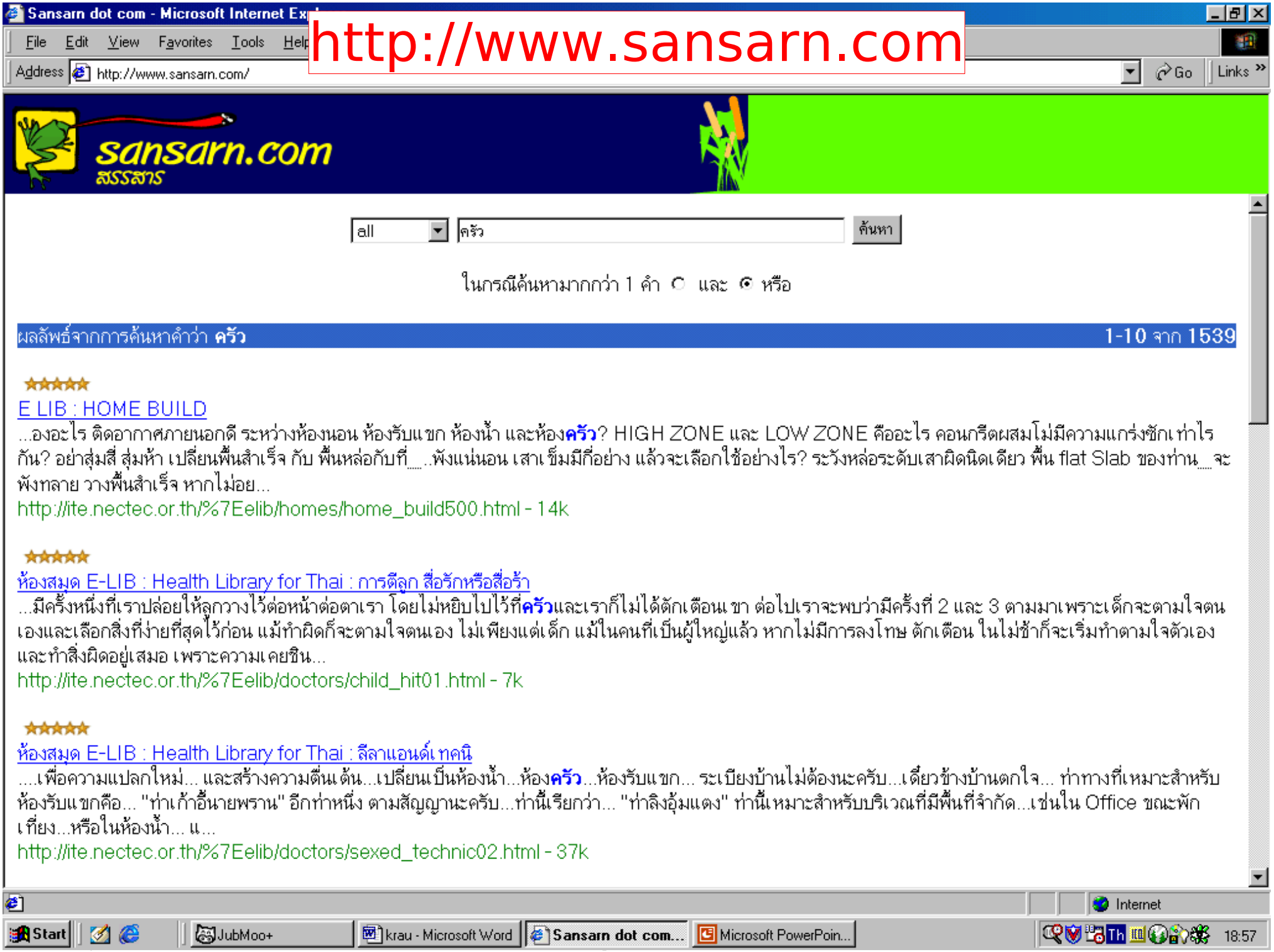


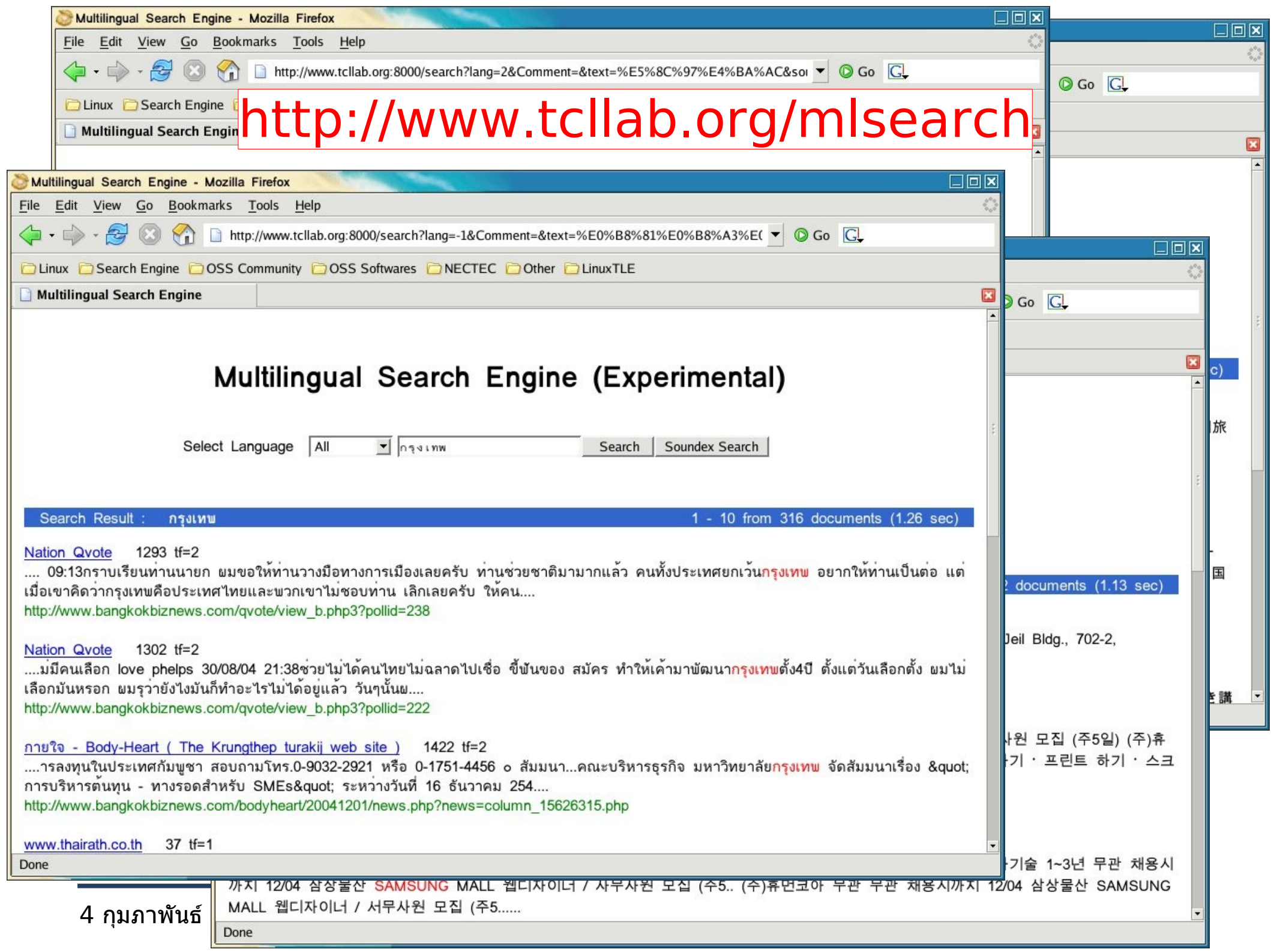
Search Engine ที่ไม่อาศัยพจนานุกรม

- To overcome the limitation of vocabulary for making index
- To deal with the out-of-vocabulary problem
- To avoid the incomplete word segmentation result
- To avoid multiple search in case of phrase search
- To make it extensible for multi-lingual search

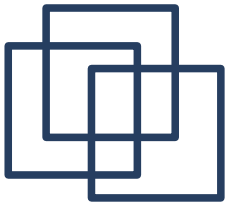
Dictionary-less Search Engine --Architecture--






Multilingual Search Engine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.tcllab.org:8000/search?lang=2&Comment=&text=%E5%8C%97%E4%BA%AC&so
Go
Linux Search Engine
Multilingual Search Engin
http://www.tcllab.org/mlsearch
Multilingual Search Engine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.tcllab.org:8000/search?lang=-1&Comment=&text=%E0%B8%81%E0%B8%A3%E(
Go
Linux Search Engine OSS Community OSS Softwares NECTEC Other LinuxTLE
Multilingual Search Engine
Multilingual Search Engine (Experimental)
Select Language All กรุงเทพ Search Soundex Search
Search Result : กรุงเทพ 1 - 10 from 316 documents (1.26 sec)
Nation Qvote 1293 tf=2
.... 09:13กราบเรียนท่านนายก ผมขอให้ท่านวางมือทางการเมืองเลยครับ ท่านช่วยชาติมามากแล้ว คนทั้งประเทศยกเว้นกรุงเทพ อยากให้ท่านเป็นต่อ แต่เมื่อเขาคิดว่ากรุงเทพคือประเทศไทยและพวกเขาไม่ชอบท่าน เลิกเลยครับ ให้คน....
http://www.bangkokbiznews.com/qvote/view_b.php3?pollid=238
Nation Qvote 1302 tf=2
....มีคนเลือก love phelps 30/08/04 21:38ช่วยไม่ได้คนไทยไม่ฉลาดไปเชื่อ ขี้ฟันของ สมัคร ทำให้เค้ามาพัฒนากรุงเทพตั้ง4ปี ตั้งแต่วันเลือกตั้ง ผมไม่เลือกมันหรอก ผมรู้ว่ายังไงมันก็ทำอะไรไม่ได้อยู่แล้ว วันๆนั้นผ....
http://www.bangkokbiznews.com/qvote/view_b.php3?pollid=222
กายใจ - Body-Heart (The Krungthep turakij web site) 1422 tf=2
....ารลงทุนในประเทศกัมพูชา สอบถามโทร.0-9032-2921 หรือ 0-1751-4456 o สัมมนา...คณะบริหารธุรกิจ มหาวิทยาลัยกรุงเทพ จัดสัมมนาเรื่อง " การบริหารต้นทุน - ทางรอดสำหรับ SMEs" ระหว่างวันที่ 16 ธันวาคม 254....
http://www.bangkokbiznews.com/bodyheart/20041201/news.php?news=column_15626315.php
www.thairath.co.th 37 tf=1
Done
documents (1.13 sec)
Jeil Bldg., 702-2,
MALL 웹디자이너 / 서무사원 모집 (주5.....
Done

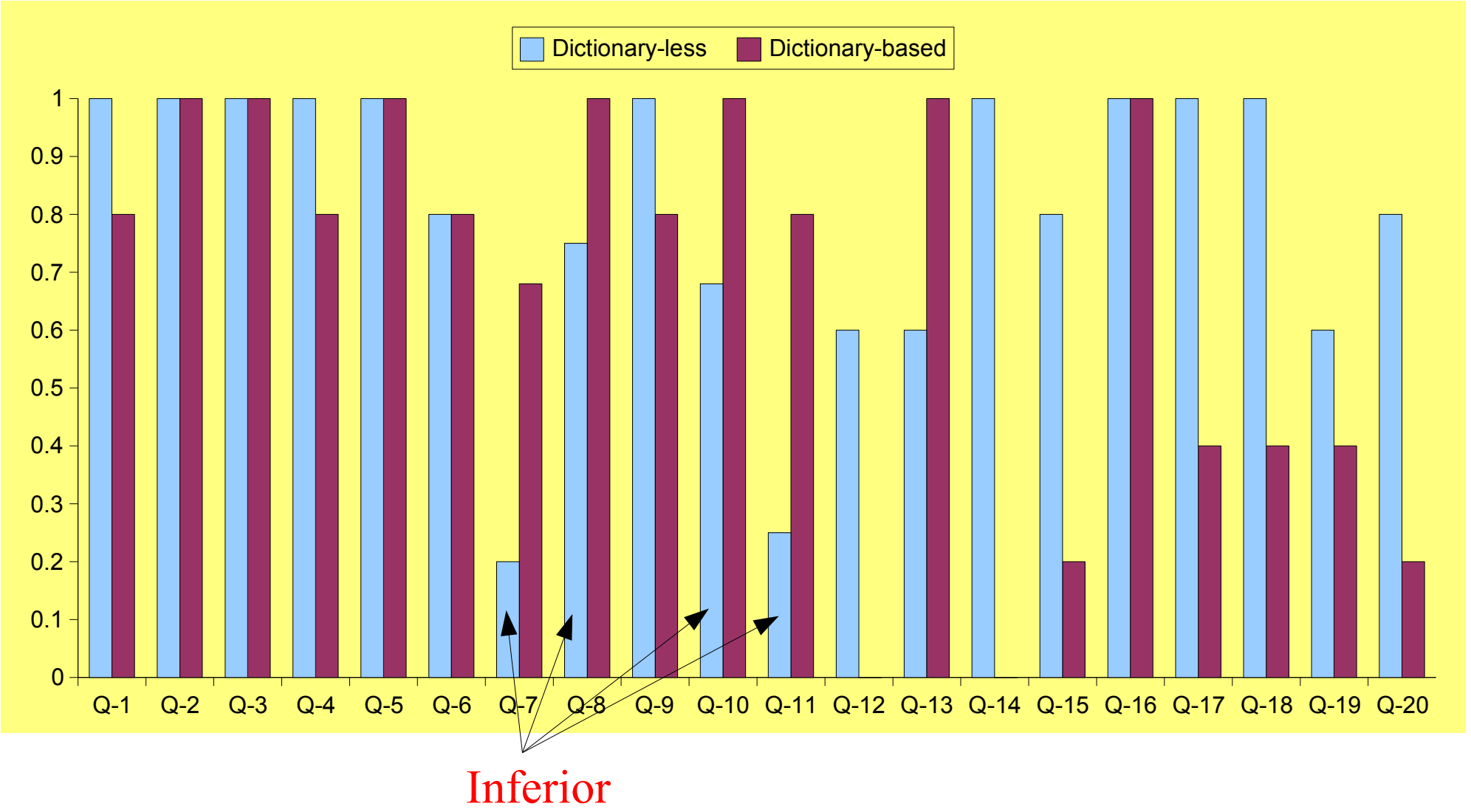
4 กุมภาพันธ์



การประเมินผล

- Evaluate top 10 results of 20 queries by 5 evaluators
- Relevant if 3 out of 5 evaluators agree on each result
- Satisfaction on the result for each query is the evarage on the relevant

Dictionary-based VS Dictionary-less --Evaluation--



List of Queries

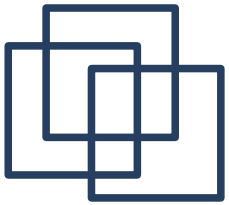
Uncommon and large

	Unsegmented queries	*Segmented queries*
1	บริจาค, สืนามิ	บริจาค, สืนามิ
2	เส้นตาย, ซัดดม	เส้นตาย, ซัดดม
3	มนัส, โชว์ตัว, จีน	มนัส, โชว์|ตัว, จีน
4	ผลกระทบ, ราคาน้ำมันแพง	ผลก|ระ|ทบ, ราคา|น้ำมัน|แพง
5	ไข้หวัดนก	ไข้หวัด|นก
6	ทุจริต, การเลือกตั้ง	ทุจริต, การ|เลือกตั้ง
7	จับกุม, ผู้ก่อการร้าย, ภาคใต้	จับกุม, ผู้ก่อการร้าย, ภาค|ใต้
8	นโยบาย, แก้ไข, ปัญหายาเสพติด	นโยบาย, แก้ไข, ปัญหา|ยา|เสพติด
9	ทดลอง, ลดค่าทางด่วน	ทดลอง, ลด|ค่า|ทางด่วน

	Unsegmented queries	*Segmented queries*
11	ลงทุน, ในพม่า	ลงทุน, ใน|พม่า
12	ส่งเสริม, การท่องเที่ยว, ไทย	ส่งเสริม, การ|ท่องเที่ยว, ไทย
13	เลือกตั้ง, ประธานาธิบดี, ปาเลสไตน์	เลือกตั้ง, ประธานาธิบดี, ปา|เลสไตน์
14	เลือกตั้ง, ผู้ว่า, กทม.	เลือกตั้ง, ผู้|ว่า, กทม|.
15	สินค้าไทย, การส่งออก	สินค้า|ไทย, การ|ส่ง|ออก
16	แปรรูปรัฐวิสาหกิจ	แปรรูป|รัฐวิสาหกิจ
17	อุ้ม, นายสมชาย	อุ้ม, นาย|สม|ชาย
18	ฉลองปีใหม่	ฉลอง|ปี|ใหม่
19	พรทิพย์, ลาออก	พร|ทิพย์, ลา|ออก
20	สวนสนุก	สวน|สนุก

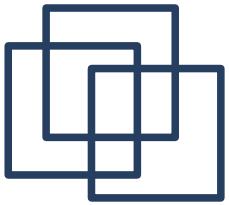
Common and small

=> Dictionary-based approach is good at searching uncommon words but not the common words.



การประเมินผล

- Dictionary-based gets inferior results in case of excessive segmentation to be common and small words i.e. การส่งออก (การ|ส่ง|ออก), but not in the case of being uncommon and large words i.e. ผลกระทบ (ผลก|ระ|ทบ)



Web language engineering for Open collaborative archiving (WLE-OCA)

Motivation:

- Establishing Thailand-Japan collaborative archiving on broadband Internet (NECTEC-MIC 45M bps high speed link)
- Collaboration between TCL-Thailand and NUT-Japan to support Language Observatory Project and Asian Language Resource Network
- Real-time monitoring population of web document

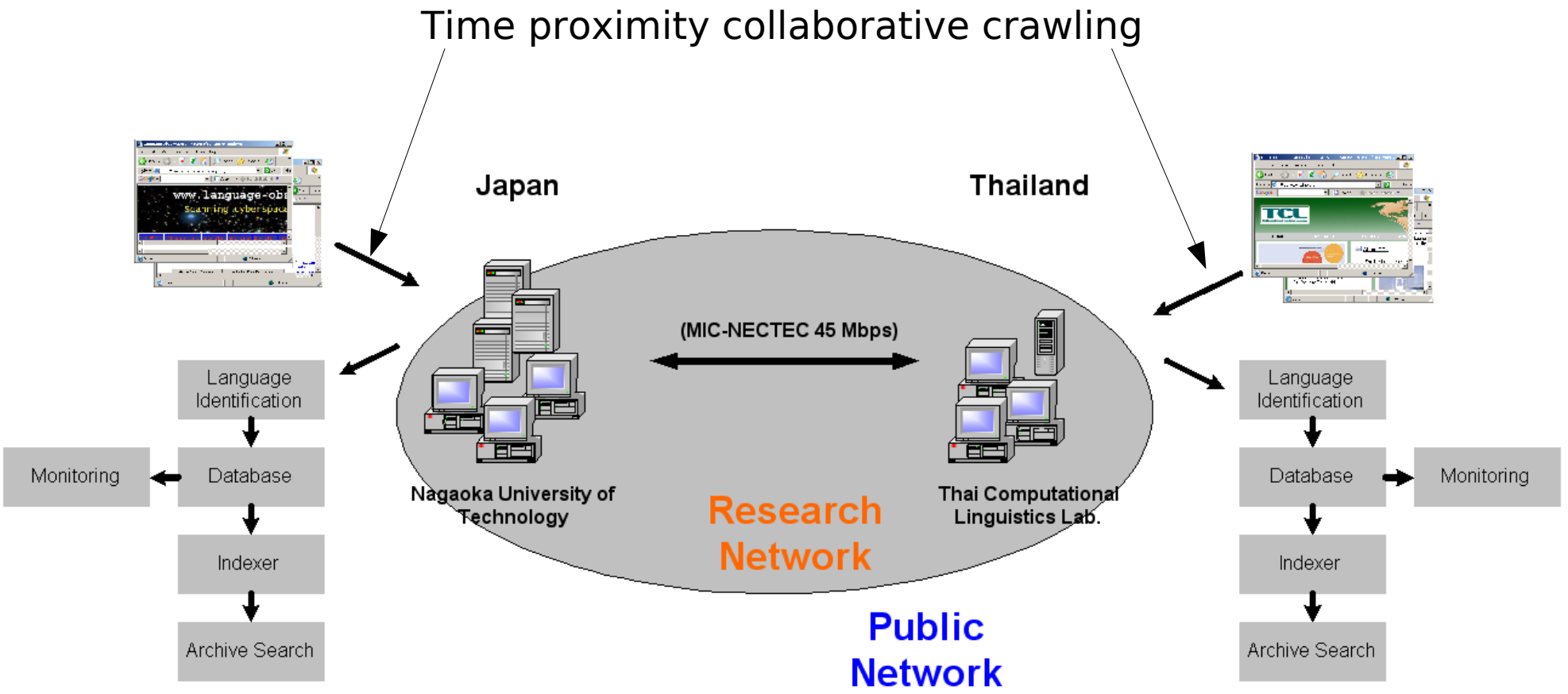
Web language engineering for Open collaborative archiving (WLE-OCA)

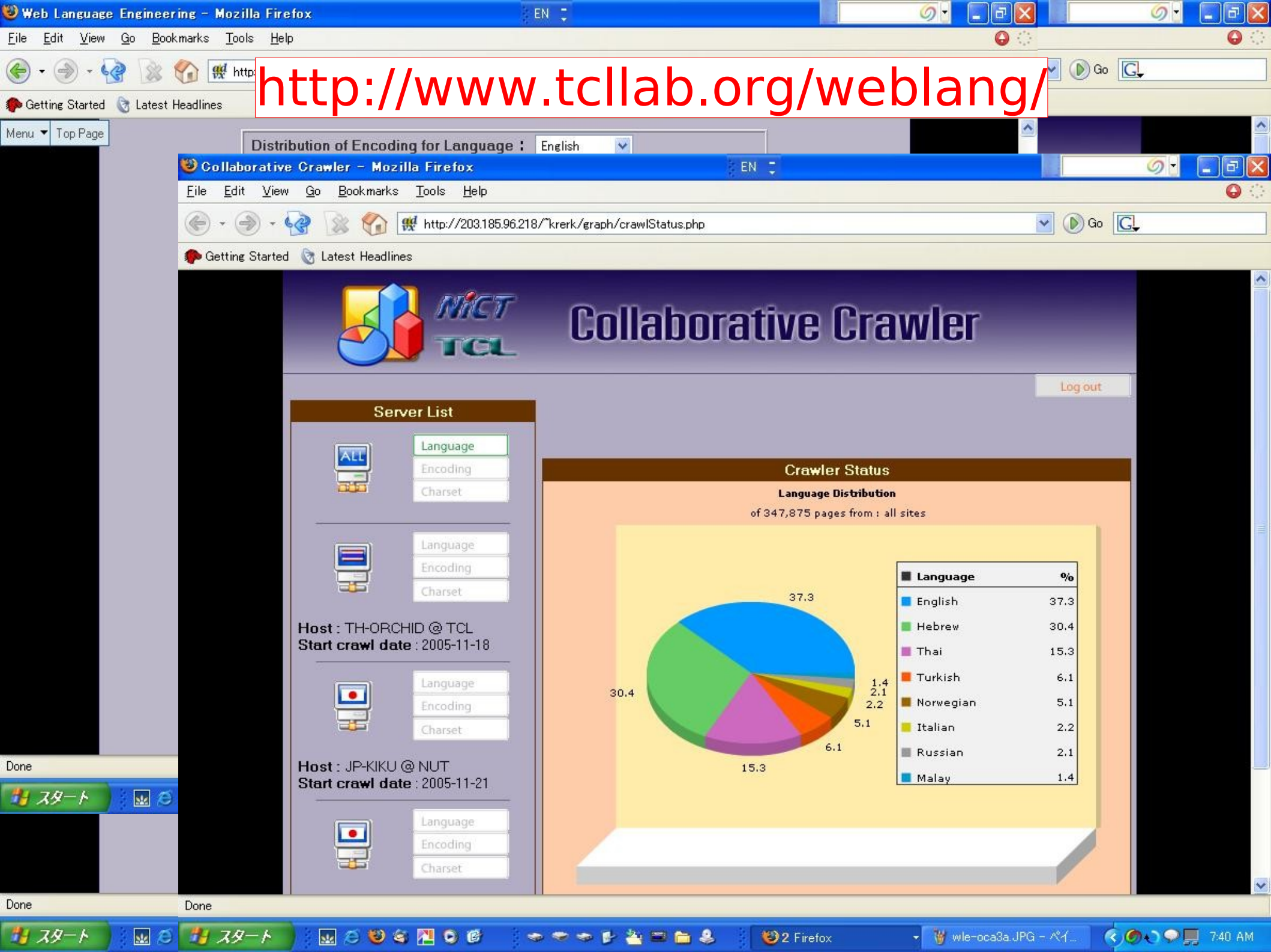
Goal:

- Archiving web document at 1m pages/day/site
- Evaluating Multi-lingual and Cross-lingual search engine

Web language engineering for Open collaborative archiving (WLE-OCA)

Architecture:




Web Language Engineering - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.tcllab.org/weblang/
Getting Started Latest Headlines
Menu Top Page
Distribution of Encoding for Language : English
Collaborative Crawler - Mozilla Firefox
http://203.185.96.218/~krerk/graph/crawlStatus.php
Collaborative Crawler
NiCT TCL
Log out
Server List
ALL
Language
Encoding
Charset
Host : TH-ORCHID @ TCL
Start crawl date : 2005-11-18
Host : JP-KIKU @ NUT
Start crawl date : 2005-11-21
Crawler Status
Language Distribution
of 347,875 pages from : all sites
Language %
English 37.3
Hebrew 30.4
Thai 15.3
Turkish 6.1
Norwegian 5.1
Italian 2.2
Russian 2.1
Malay 1.4
Done
スタート
2 Firefox
7:40 AM

เสมือนกระจกเงาสำหรับสำรวจความรู้ในสังคม

ให้รู้ถึง แหล่งที่มา ภาษา และเนื้อหา

WLE-OCA

www.tcllab.org/weblang