

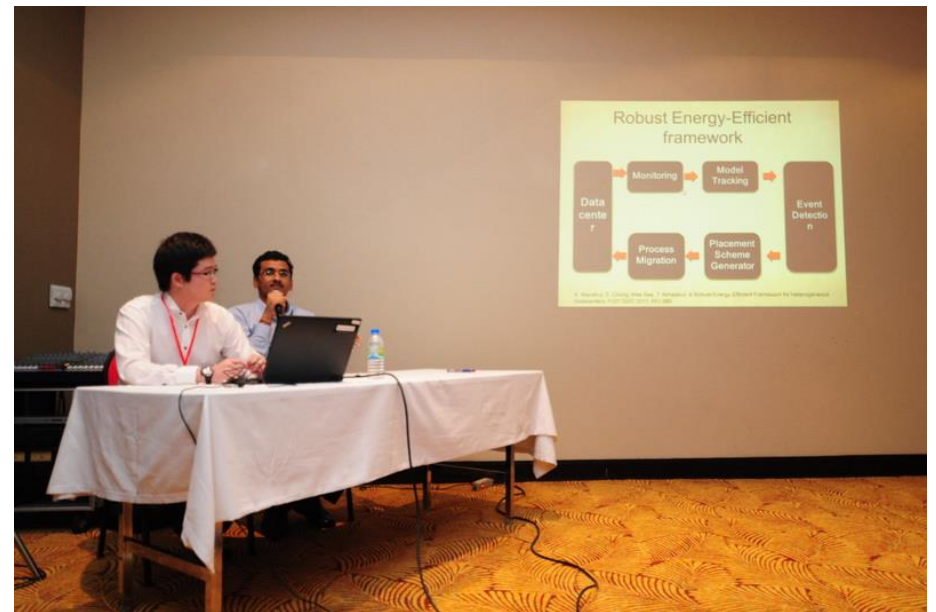
## Overview of ECTI-CON 2012

- ECTI-CON 2012 is the ninth annual international conference.
- Organized by Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) Association, Thailand.
- May 16-18, 2012 Hua-Hin Thailand.
- Promote technological progress of thailand.
- The program of ECTI-CON 2012 will consist of 7 area.

# Overview of ECTI-CON 2012

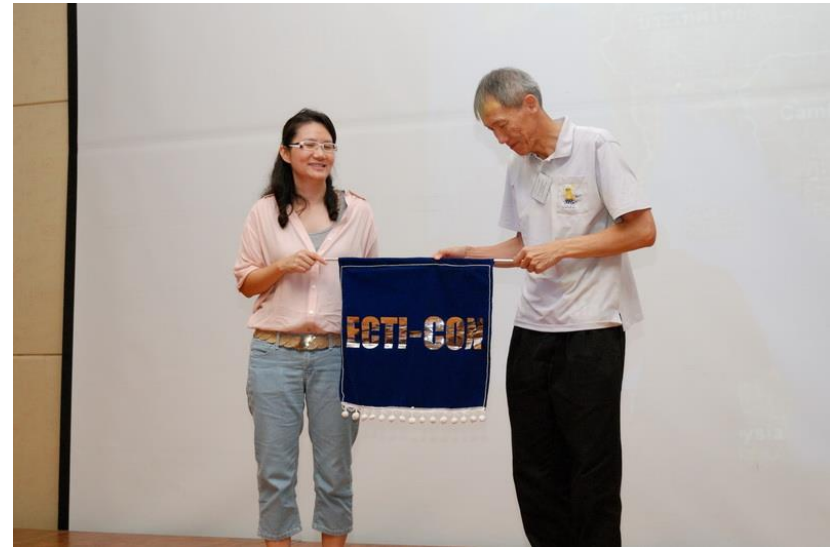
- The program of ECTI-CON 2012 will consist of 7 area.
  - Device, Circuits and Systems
  - Computers and Information Technology
  - Communication Systems
  - Controls
  - Electrical Power Systems
  - Signal Processing
  - Other Related Fields

# The atmosphere in ECTI-CON 2012





# The atmosphere in ECTI-CON 2012



# **Search Result Clustering for Thai Twitter Based on Suffix Tree Clustering**

**Santipong Thaiprayoon, Alisa Kongthon,  
Pornpimon Palingoon and Choochart Haruechaiyasak**

Speech and Audio Technology Laboratory (SPT)  
National Electronics and Computer Technology Center (NECTEC)  
Thailand Science Park

# Presentation Outline

- Background and motivation
- Our proposed approach
  - Suffix Tree Clustering with Label Merging (STC-LM)
- Resource and Experiment
  - Corpus
  - Experiment
  - Evaluation results
- Conclusion
- Future works

# Background and Motivation

- Today **Twitter** has become a popular online medium for posting and sharing news and events.
- There are many Twitter posts containing **similar messages** on Twitter.
- There are some existing Twitter search engines such as TwitterSearch, TwitterScan and Topsy.
- Users have difficulties in finding relevant results from **long list of search results**.

# Problem of Twitter search engines

Results for น้ำท่วม

Query for: flooding

Top people · View all

ช่วยเหลือน้ำท่วม @thaiiflood  
ศูนย์ข้อมูลช่วยเหลือผู้ประสบกับน้ำท่วม

Follow

Tweets Top / All

1 new Tweet

satinee @satinee  
คาด 2-3 สัปดาห์ รัฐเคสียร์ชัดเจน น้ำท่วม 1.2 แสนล. :  
[thaireform.in.th/reform-the-new...](http://thaireform.in.th/reform-the-new...) via @thaireform

Expand

Thaireform @thaireform  
คาด 2-3 สัปดาห์ รัฐเคสียร์ชัดเจน น้ำท่วม 1.2 แสนล. :  
[thaireform.in.th/news-highlight...](http://thaireform.in.th/news-highlight...) via @thaireform

Expand

Nonflood @Nonflood  
RT @traffy RT @DailyNewsTwit: ชุดทดสอบไฟรั่ว ฝีมือเยาวชนวิทย์จุฬาฯ  
แรงบันดาลใจจากเหตุน้ำท่วม [bit.ly/JrHNWS](http://bit.ly/JrHNWS) #nonflood #thaiiflood

Expand Reply Retweet Favorite

Traffy.in.th @traffy  
RT @DailyNewsTwit: ชุดทดสอบไฟรั่ว ฝีมือเยาวชนวิทย์จุฬาฯ แรงบันดาลใจ  
จากเหตุน้ำท่วม [bit.ly/JrHNWS](http://bit.ly/JrHNWS)

Expand

Similar messages  
on Twitter

Similar messages  
on Twitter



# What is Search Result Clustering (SRC)

- Clustering small collections of documents, search results or document abstract into thematic categories.

The screenshot shows a Google search for 'thailand'. The search bar at the top has 'thailand' entered. Below the search bar, there are tabs for 'web', 'news', 'images', 'maps', 'blogs', 'wikipedia', 'jobs', and 'more'. The 'web' tab is selected, and the search results are displayed. On the left side, there is a sidebar with various filters like 'Everything', 'Images', 'Maps', 'Videos', 'News', 'Blogs', 'Books', and 'More'. The main search results area shows a list of results, including 'Thailand - Wikipedia, the free encyclopedia', 'Thailand officially the Kingdom of Ratcha Anachak Thai', 'Prostitution in Thailand - Ying', 'Tourism Authority of Thailand', 'Thailand Travel Information', '2 Apr 2012 - Thailand tourism history, culture, transport and w', 'News for thailand', '2012 Young Artist', 'Bangkok Post - 2', 'Chon Buri teenage to become the 201', 'Bangkok Post showing the ...', 'Thailand may help boost agric', 'Gulf Daily News - 3 hours ago', 'STOCKS NEWS THAILAND-Siam Cement up; Siam Makro down on MSCI ind', and 'Reuters Africa - 1 hour ago'. A red box highlights a section titled 'All Results (133)' which lists various clusters of results: 'Thailand Travel (12)', 'Exporter (23)', 'Maps (14)', 'Southeast Asia (11)', 'Market (11)', 'Bangkok, Thailand (13)', 'Photographs (10)', 'Business, Bangkok Post, Thailand Knight Ridder (8)', 'Plant (6)', and 'Project, Planning (6)'. Below this list, there are links for 'more' and 'all clouds', and a search bar labeled 'find in clouds:' with a 'Find' button. On the right side, there is a section titled 'Top 130 results of at least 389,000,000 retrieved for the query thailand (c) thailand'. This section includes a definition of 'thailand' as a noun: 'Thailand, Kingdom of Thailand, Siam -- (a country of southe Malay peninsula; "Thailand is the official name of the former Siam")'. Below this, there is a section titled 'Thailand—Profile' which provides a brief overview of Thailand, including its population (61.8 million), location (in the heart of the Andaman Sea and Gulf of Thailand in the south), and known names (Siam until 1939). Further down, there is a section titled 'Vacation In Thailand' which encourages planning a trip to Thailand and provides a link to the official tourism site. Below this, there is a section titled 'Bangkok hotels' which provides information about finding hotels, attractions, and more, and a link to bangkok.sawadee.com. Finally, there is a section titled 'learn photography' which provides information about learning photography on travel and a link to www.thaiphoto.net.

Google

thailand

web news images maps blogs wikipedia jobs more »

thailand

Search

advanced preferences

About 227,000,000 results (0.37 s)

Search

Everything

Images

Maps

Videos

News

Blogs

Books

More

Khlong Luang, Pathum Thani

Change location

The web

Pages from Thailand

Any time

Past hour

Past 24 hours

Past 2 days

Past week

Past month

Past year

Custom range...

Thailand - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Thailand

Thailand officially the Kingdom of Ratcha Anachak Thai; IPA: [râ:t

Prostitution in Thailand - Ying

Tourism Authority of Thailand

www.tourismthailand.org/ - Ca

The official site of Tourism Auth

information, Travel guide, maps

Thailand Travel Information

www.lonelyplanet.com/thailand

2 Apr 2012 - Thailand tourism

history, culture, transport and w

in ...

News for thailand

2012 Young Artist

Bangkok Post - 2

Chon Buri teenage

to become the 201

Bangkok Post showing the ...

Thailand may help boost agric

Gulf Daily News - 3 hours ago

STOCKS NEWS THAILAND-Siam Cement up; Siam Makro down on MSCI ind

Reuters Africa - 1 hour ago

clouds sources sites time

All Results (133)

Thailand Travel (12)

Exporter (23)

Maps (14)

Southeast Asia (11)

Market (11)

Bangkok, Thailand (13)

Photographs (10)

Business, Bangkok Post, Thailand Knight Ridder (8)

Plant (6)

Project, Planning (6)

more | all clouds

find in clouds:

Find

Top 130 results of at least 389,000,000 retrieved for the query thailand (c)

thailand

noun - Thailand, Kingdom of Thailand, Siam -- (a country of southe Malay peninsula; "Thailand is the official name of the former Siam")

Thailand—Profile

THAILAND—PROFILE(2001 pop. 61.8 million). Thailand lies in the heart of the Andaman Sea and Gulf of Thailand in the south. Known as Siam until 1939, it has a rich culture.GeographyWith an area

Vacation In Thailand

Plan Your Trip To Thailand At The Official Tourism Site of Thailand!

www.tourismthailand.org/US

Bangkok hotels

Find Bangkok hotels, attractions and more. Get reviews and discounts.

bangkok.sawadee.com

learn photography

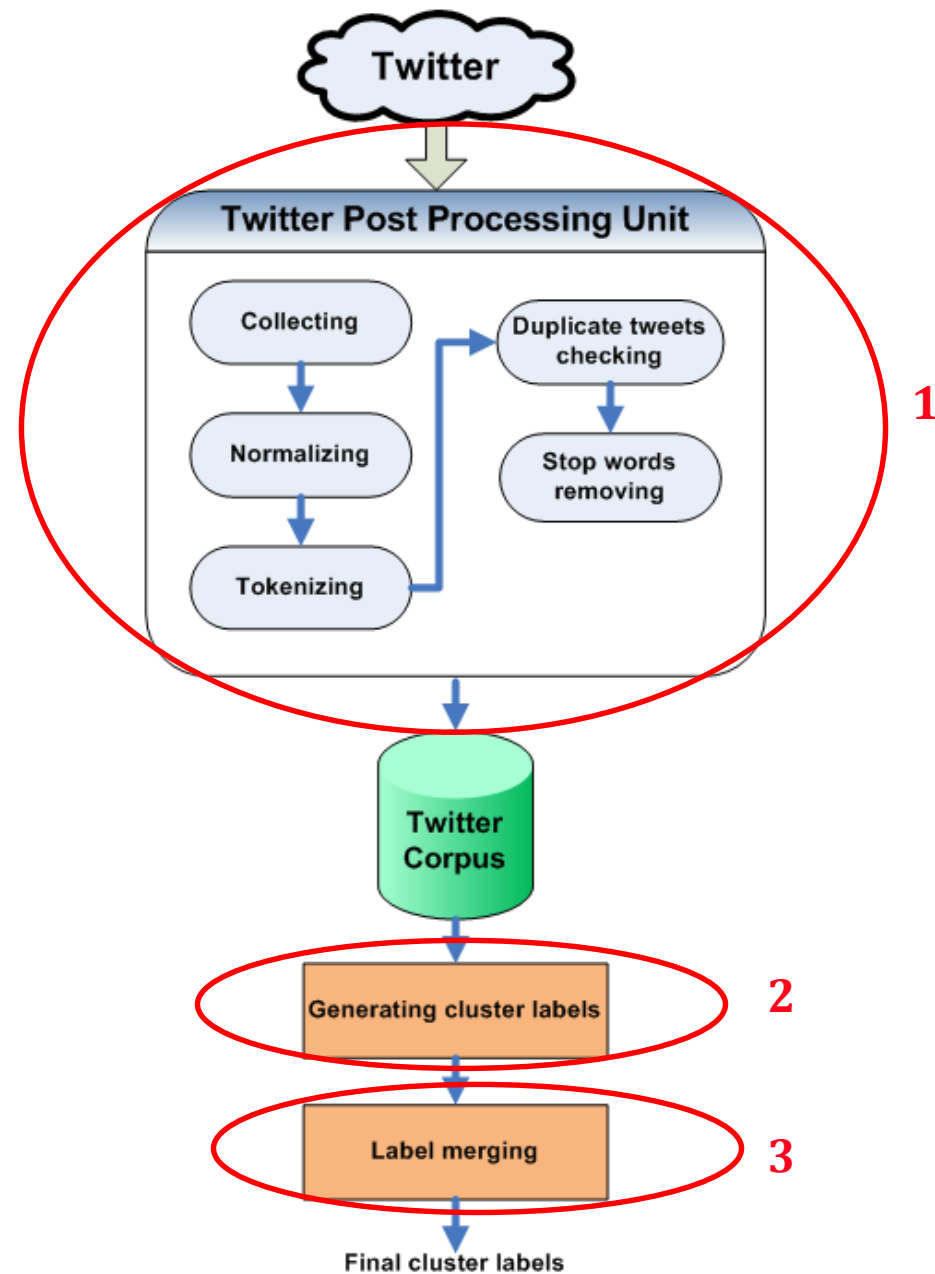
learn photography on your travel with professional photographer

www.thaiphoto.net

## Background and Motivation (cont'd)

- Many previous works use Suffix Tree Clustering (STC) algorithm for clustering search results.
- We apply the STC algorithm which is a part of Carrot2 framework.
- However, the main drawbacks of Carrot2 are.
  - Some of the returned cluster labels are unmeaningful.
  - It is unable to create a hierarchical structure of cluster labels.

# Our Proposed Approach



# Twitter Post Processing Unit

- Collecting: using the twitter4j library
- Normalizing: strip the HTML tags, number and special symbols.
- Tokenizing: segment a given text into word tokens
- Duplicate tweet checking: filter out the near-duplicate tweets by using Jaccard coefficient.
- Stopword: remove the Thai and English stopword

## Generating Cluster labels

- This component will return the top-N tweets results from the Twitter corpus.
- The tweet results are cluster labels generated.



# Label Merging

- Each cluster label is converted from abbreviated form to full word form.
- We automatically tag the POS of each word and remove the cluster labels which do not contain any noun.

Original Carrot2 Framework (Baseline)	STC-LM Approach
<div> <div>Merge</div> <div>Create Two-level</div> </div> <div> <div>สู้น้ำท่วม (fight against flooding)</div> <div>น้ำท่วมจั้ง (retained flood)</div> <div><del>ช่วย (help) — Verb</del></div> <div>จ่ายเงินช่วยน้ำท่วม (donated money for flooding)</div> <div>ผู้ว่าฯ กทม. (bangkok governor)</div> <div>น้ำท่วม (flooding)</div> <div>ช่วยน้ำท่วม (help flooding)</div> <div>ยิ่งลักษณ์ (Yingluck)</div> <div><del>ฟื้นฟู (flood relief) — Verb</del></div> </div>	<div> <div>- สู้น้ำท่วม (fight against flooding)</div> <div>— น้ำท่วมจั้ง (retained flood)</div> <div>— จ่ายเงินช่วยน้ำท่วม ท่วม (donated money for flooding)</div> <div>— ช่วยน้ำท่วม (help flooding)</div> <div>- ผู้ว่าฯ กรุงเทพมหานคร (bangkok governor)</div> <div>- ยิ่งลักษณ์ (Yingluck)</div> </div>

# Thai Tweet Corpus

- We collected tweets from many **Twitter profile pages**.
- 160,000 tweets.
- We selected popular keywords under 12 topics such as flooding, survival kits and oil prices.
- Each tweet is cleaned and tokenized by using the Twitter post processing unit.

- We apply the carrot2 framework to perform all the experiments.
- We used the default setting of carrot2 framework for generating cluster labels.
- We limited the maximum final cluster labels to 20
- We compare cluster label results between our approach and original Carrot2 by using precision, recall and F1 measure.

# Performance Evaluation Metric

$$P = \frac{\text{number of returned cluster labels that are relevant}}{\text{number of returned cluster labels}}$$

$$R = \frac{\text{number of returned cluster labels that are relevant}}{\text{number of relevant number of cluster labels}}$$

$$F = 2 \times \frac{P \times R}{P + R}$$

# Evaluation Example

Human	Carrot2 (Baseline)	STC-LM
<u>ระดับ ดัน กัน น้ำ</u>	อพยพ	<u>ระดับ ดัน กัน น้ำ</u>
<u>ดัน กัน น้ำ ปึก</u>	<u>ระดับ ดัน กัน น้ำ</u>	<u>ดัน กัน น้ำ ปึก</u>
<u>กระสอบทราย</u>	<u>ดัน กัน น้ำ ปึก</u>	<u>แนว ดัน กัน น้ำ</u>
<u>แนว ดัน กัน น้ำ</u>	<u>กระสอบทราย</u>	<u>ดัน กัน น้ำ คลอง</u>
<u>คลอง มหา สวัสดิ์</u>	นนทบุรี	<u>เหนือ ดัน กัน น้ำ</u>
<u>ดัน กัน น้ำ คลอง</u>	<u>แนว ดัน กัน น้ำ</u>	<u>เรือ ดัน กัน น้ำ</u>
<u>เหนือ ดัน กัน น้ำ</u>	คลอง	<u>ดัน กัน น้ำ ริม</u>
<u>น้ำ กระสอบทราย</u>	<u>คลอง มหา สวัสดิ์</u>	<u>น้ำ กระสอบทราย</u>
<u>คลอง ประปา</u>	<u>ดัน กัน น้ำ คลอง</u>	<u>กระสอบทราย ดัน</u>
<u>พัง ดัน กัน น้ำ</u>	<u>เหนือ ดัน กัน น้ำ</u>	<u>คลอง มหา สวัสดิ์</u>
<u>กระสอบทราย ดัน</u>		<u>คลอง ประปา</u>
<u>เรือ ดัน กัน น้ำ</u>		<u>เจ้า พระยา</u>
<u>ดัน กัน น้ำ ริม</u>		นนทบุรี จังหวัด
$P = 7/10 = 0.7$ $R = 7/16 = 0.43$ $F1 = 2 * (0.7 * 0.43) / (0.7 + 0.43) = 0.53$		
$P = 12/13 = 0.92$ $R = 12/16 = 0.75$ $F1 = 2 * (0.92 * 0.75) / (0.92 + 0.75) = 0.82$		
Precision	0.7	0.92
Recall	0.43	0.75
F-Measure	0.53	0.82



# Cluster Label Evaluation Results

Topics	Baseline			STC-LM		
	P	R	F1	P	R	F1
พ่อปลาบ (plaboo's father)	0.900	0.642	0.750	1.000	0.785	0.880
น้ำท่วม (flood)	0.700	0.538	0.608	0.615	0.615	0.615
งูแมนบ้า (mambas)	0.900	0.642	0.750	0.900	0.642	0.750
คปท. (flood center)	0.700	0.583	0.636	0.666	0.666	0.666
กระสอบทราย (sandbag)	0.900	0.692	0.782	0.909	0.769	0.833
ถุงยังชีพ (survival kits)	0.600	0.461	0.521	0.777	0.538	0.636
บิ๊กแบ็ก (big bag)	0.800	0.571	0.666	1.000	0.714	0.833
คันกันน้ำ (dike)	0.700	0.437	0.538	0.857	0.750	0.800
ซี7 ตบป้องหู (C7 assault)	0.600	0.352	0.444	0.687	0.647	0.666
ราคาน้ำมัน (oil prices)	0.900	0.642	0.750	1.000	0.642	0.782
ปรับ ครม. (cabinet shuffle)	0.600	0.461	0.521	0.777	0.538	0.636
บริหารจัดการน้ำ (water management)	0.500	0.294	0.370	0.428	0.352	0.387
<b>Average</b>	<b>0.733</b>	<b>0.527</b>	<b>0.612</b>	<b>0.802</b>	<b>0.639</b>	<b>0.707</b>

# Conclusion

- We present a new approach called Suffix Tree Clustering with Label Merging (STC-LM).
- STC-LM help generating more informative cluster labels.
- The main task of the STC-LM is to merge and create two-level label structure.
- 70% F1 measurement (improved 9%)

## Future Works

- Considering timelines for cluster labels.
- Using incorporate probabilistic and machine learning into our label merging algorithm

**Thank you for your attention**

**Questions / Comments ?**