

Emerging Topic Detection for Organizations from Microblogs

Yan Chen*, Hadi Amiri+, Zhoujun Li* and Tat-Seng Chua+

*State Key Laboratory of Software Development Environment, Beihang University, Beijing, China

+School of Computing, National University of Singapore, Singapore

The 36th Annual ACM SIGIR Conference.
Dublin, Ireland. 28th July-1st August, 2013.

Outline

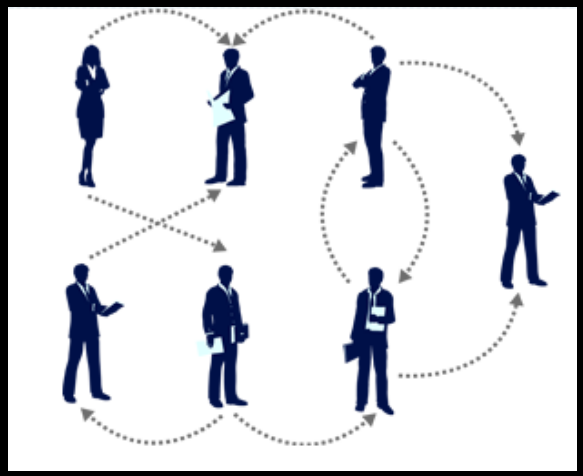
- Background
- Organization-related Data Selection
- Hot Emerging Topic Detection
- Experiments and Analysis
- Conclusion and Future Work

Outline

- Background
- Organization-related Data Selection
- Hot Emerging Topic Detection
- Experiments and Analysis
- Conclusion and Future Work

Background

- Microblog Services
  - Interaction
  - Feature
    Real time
  - Users
    Individuals
    Organizations
    eg: banks, universities, government organizations, and so on.



Background

Optus

@Optus

We're here to hear from you. If you've got questions or just something you'd like to tell us, we're online Monday-Friday from 9am to 8pm & Saturday 9am-5pm AEST

Australia http://www.optus.com.au

Follow

77,825 TWEETS

2,966 FOLLOWING

23,687 FOLLOWERS

Telstra

@Telstra

We're here to provide customer support and answer any Telstra questions you might have whenever it works for you - 24 hours a day, 7 days a week!

Australia http://www.telstra.com.au

Follow

84,767 TWEETS

3,742 FOLLOWING

39,489 FOLLOWERS

Vodafone Australia

@Vodafone_AU

Follow us for all the latest network news, product and service announcements, and special promotions. For help and support, please tweet @vodafoneau_help.

Australia http://www.vodafone.com.au

Follow

9,357 TWEETS

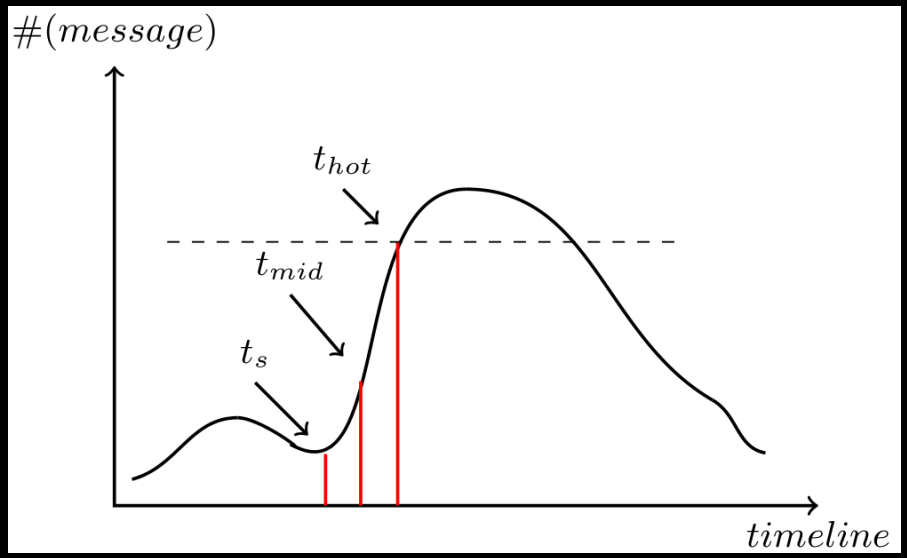
19,115 FOLLOWING

23,825 FOLLOWERS

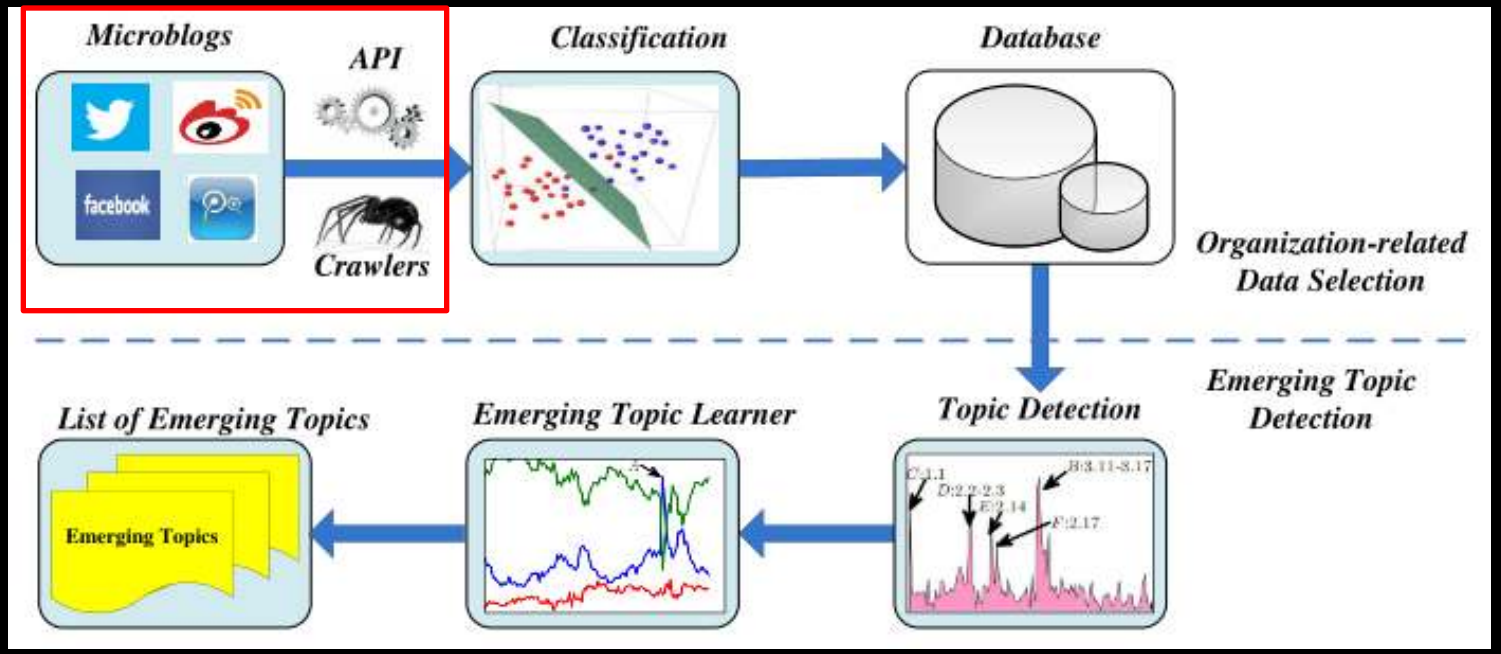
Motivation

- Organizations expect to:
  - Track the evolution of any identified relevant topics.
  - Be informed of any new emerging topics.

- Hot Emerging Topic
  - Novel
  - Hot and viral in the near future



Overview of framework



- Stages:
  - Data crawlers
  - Classification
  - Live topic detection
  - Live hot emerging topic detection

Focus and Contributions

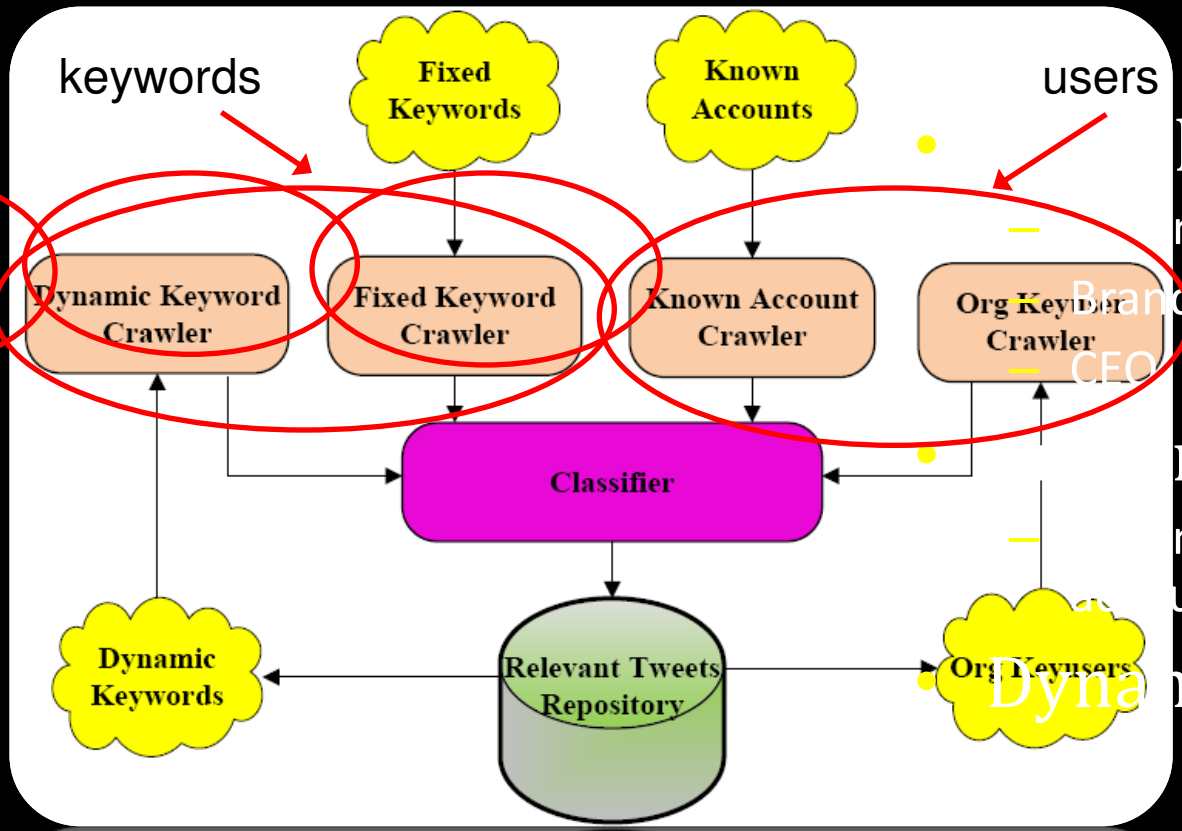
- A multi-source crawling strategy

- Techniques for hot emerging topic detection

Outline

- Background
- Organization-related Data Selection
- Hot Emerging Topic Detection
- Experiments and Analysis
- Conclusion and Future Work

Organization-related Data Selection



keywords

nization Name

Brands

CEO

n Accounts

nization Official

unts

Dyna mic Keywords

Dynamic Keywords Generation

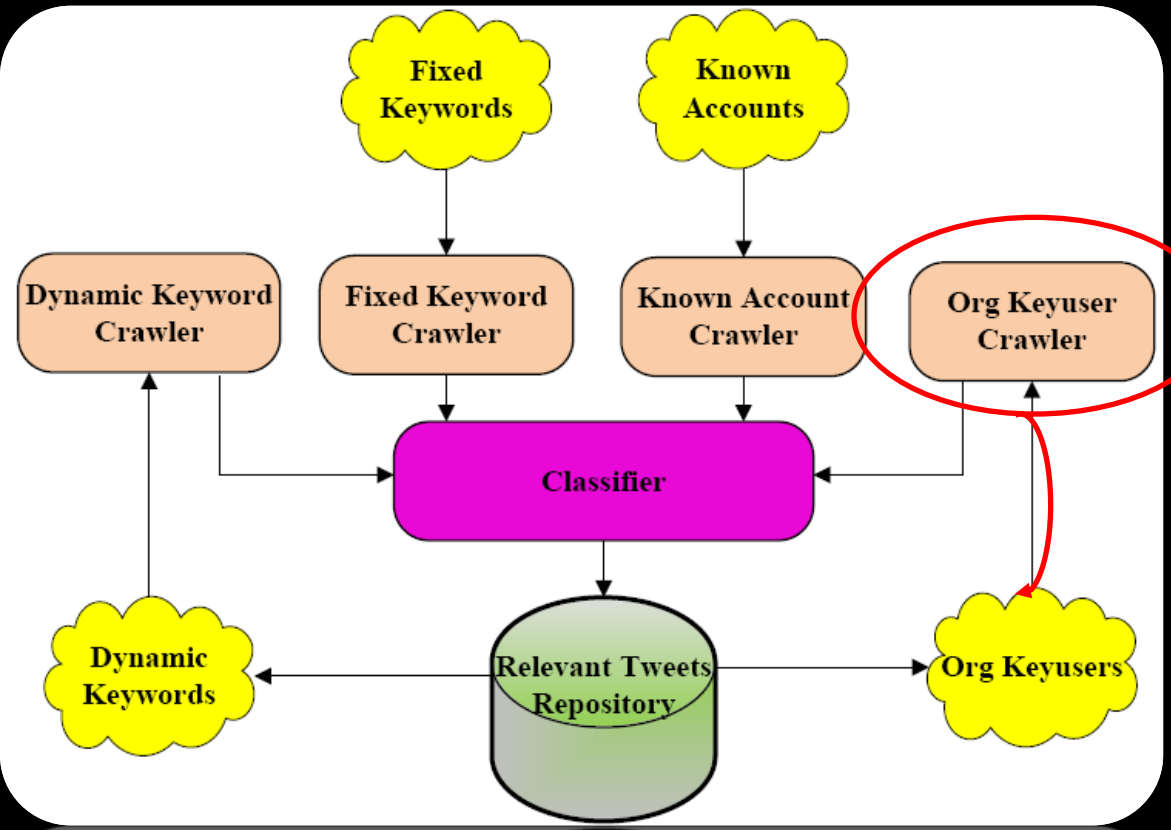
- Definition:
  - Newly introduced representative terms.

- Methods:
  - Foreground [t-T]
  - Background
    [t-2T, t-T],
    [t-T] of previous day
    [t-T] of one week ago
  - Chi-square distribution

$$\chi_i^2 = \begin{cases} \frac{(f_i - b_i)^2}{b_i} + \frac{[(100-f_i)-(100-b_i)]^2}{100} & \text{if } f_i > b_i; \\ 1 & \text{if } otherwise. \end{cases}$$

  - Rank top N as dynamic keywords

Organization-related Data Selection



- Fixed keywords
  - Organization Name
  - Brands
  - CEO
- Known Accounts
  - Organization official accounts
- Dynamic Keywords
- Org Keyusers

Graph-based Org Keyusers Generation

- Organization user relationship graph
  - *Nodes*: known accounts, all users posted at least one organization relevant tweets, their friends and followers;
  - *Edges*: social relationship between nodes.

- Method
  - A time interval T (e.g.: 24 hours)
  - A subset of users U - post at least one relevant tweets in [t − T, t]
  - Incorporating the activity degree (tweeting times in current time interval) of user into graph by a Pagerank similar algorithm.

$$auth(u_i) = \alpha \sum_{u_j \in follower(u_i)} \frac{auth(u_j)}{|following(u_j)|} + (1-\alpha)\frac{|Tw_{\Delta t}^{u_i}|}{|Tw_{\Delta t}|},$$

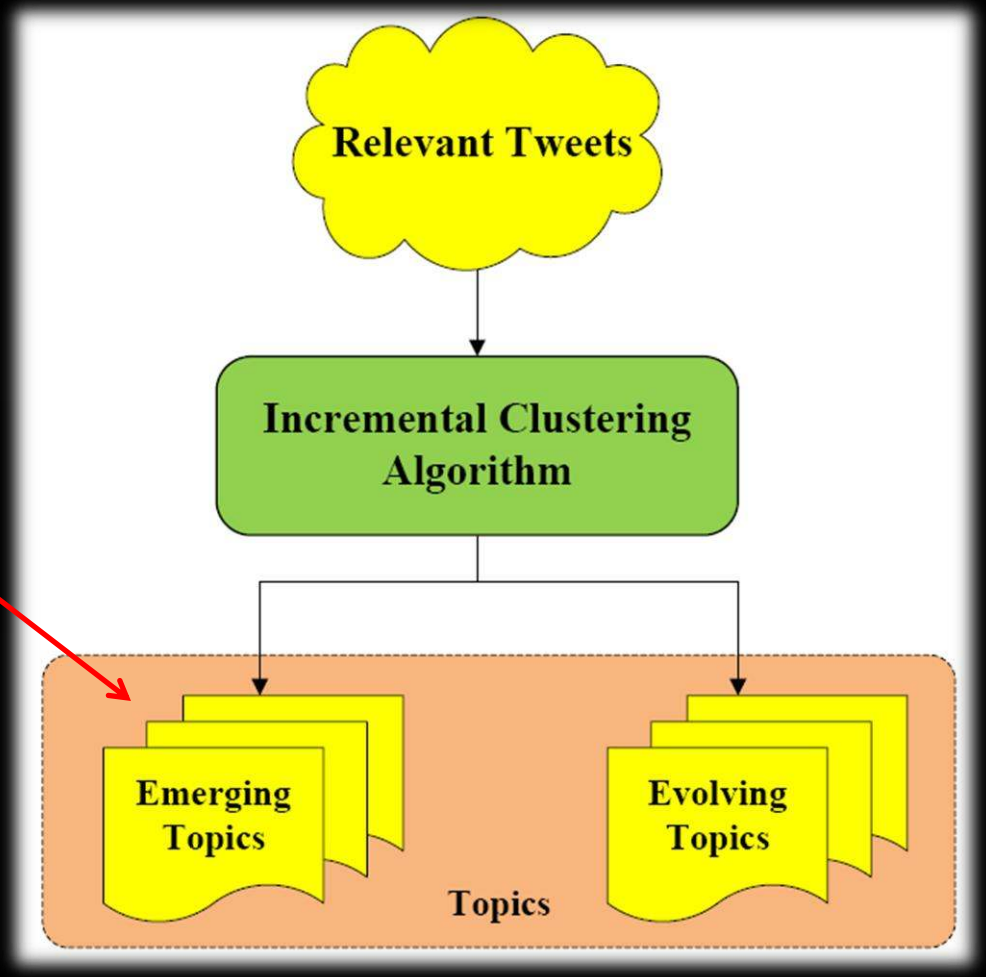
  - Top N from U as key users

Outline

- Background and Motivation
- Related Work
- Organization-related Data Selection
- Hot Emerging Topic Detection
- Experiments and Analysis
- Conclusion and Future Work

Topic Detection

- A single-pass incremental clustering algorithm



Features for Hot Emerging Topic Detection

- Frequency Rate based features:
  - Increasing rate of users number
  - Increasing rate of tweets number
  - Increasing rate of retweets number
- Influence based features:

Topical User Authority

- Observations
  - Posted many tweets about topic *tp*;
  - Posted more tweets retweeted by other users in U_{tp};
  - More followers in U_{tp}.

$$auth_{tp}(u_i) = \beta \frac{r_{u_i}}{\sum r_{u_j}} + \varphi \frac{f_{u_i}+1}{\sum f_{u_j}} + \omega \frac{q_{u_i}+1}{\sum q_{u_j}},$$

  - r_{ui} is the total number of relevant tweets posted by u_i;
  - f_{ui} is the total number of u_i's followers who exist in U_{tp} ;
  - q_{ui} is the total number of u_i's relevant tweets retweeted by others;
  - weighting parameters

Topical Tweet Influence

- Observations
  - Be retweeted by a higher number of times;
  - Posted by a topic authority user;
  - Have the potential to influence more users.

$$auth_{tp}(tw_i) = \log(1 + auth_{tp}(u_{tw_i})) + \sum_{u \in U_{rtw_i}} \log(1 + auth_{tp}(u)),$$

- Term score
  - By tweets that appeared in;

$$Weight_{tp}(w_i) = \frac{\sum_{\forall tw_j \in Tw_{tp} \wedge w_i \in tw_j} auth_{tp}(tw_j)}{\sum_{\forall w \in W_{tp}} \sum_{\forall tw \in Tw_{tp} \wedge w \in tw} auth_{tp}(tw)}$$

Features for Hot Emerging Topic Detection

- Frequency Rate based features:
  - Increasing rate of users number
  - Increasing rate of tweets number
  - Increasing rate of retweets number
- Influence based features:
  - The overlap of Org key users and Topic key users
  - The overlap of Org keywords and Topic keywords
  - The Influence of the tweets' accumulated score

Hot Emerging Topic Detection

- Two factors
  - Insufficient training data
  - Imbalance of positive and negative data
- Semi-supervised classifiers
  - Co-training Classifier
  - Semi-Ensemble Classifier

Semi-supervised Classifiers

- Co-training Classifier
  - Features divided into two views

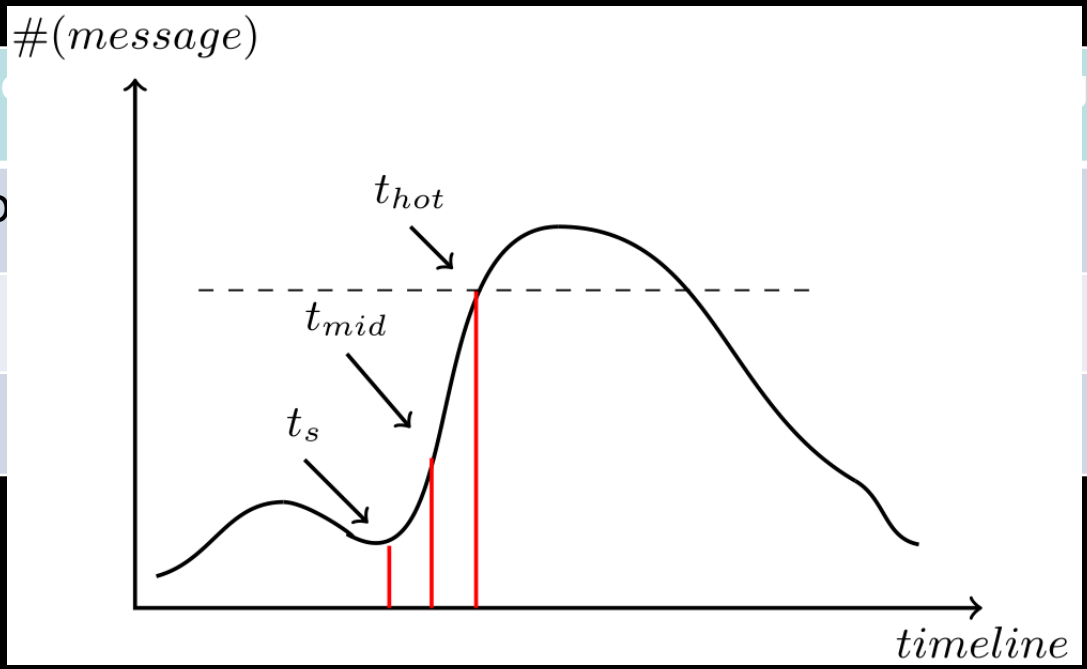
- Semi-Ensemble Classifier
  - Voting based

Outline

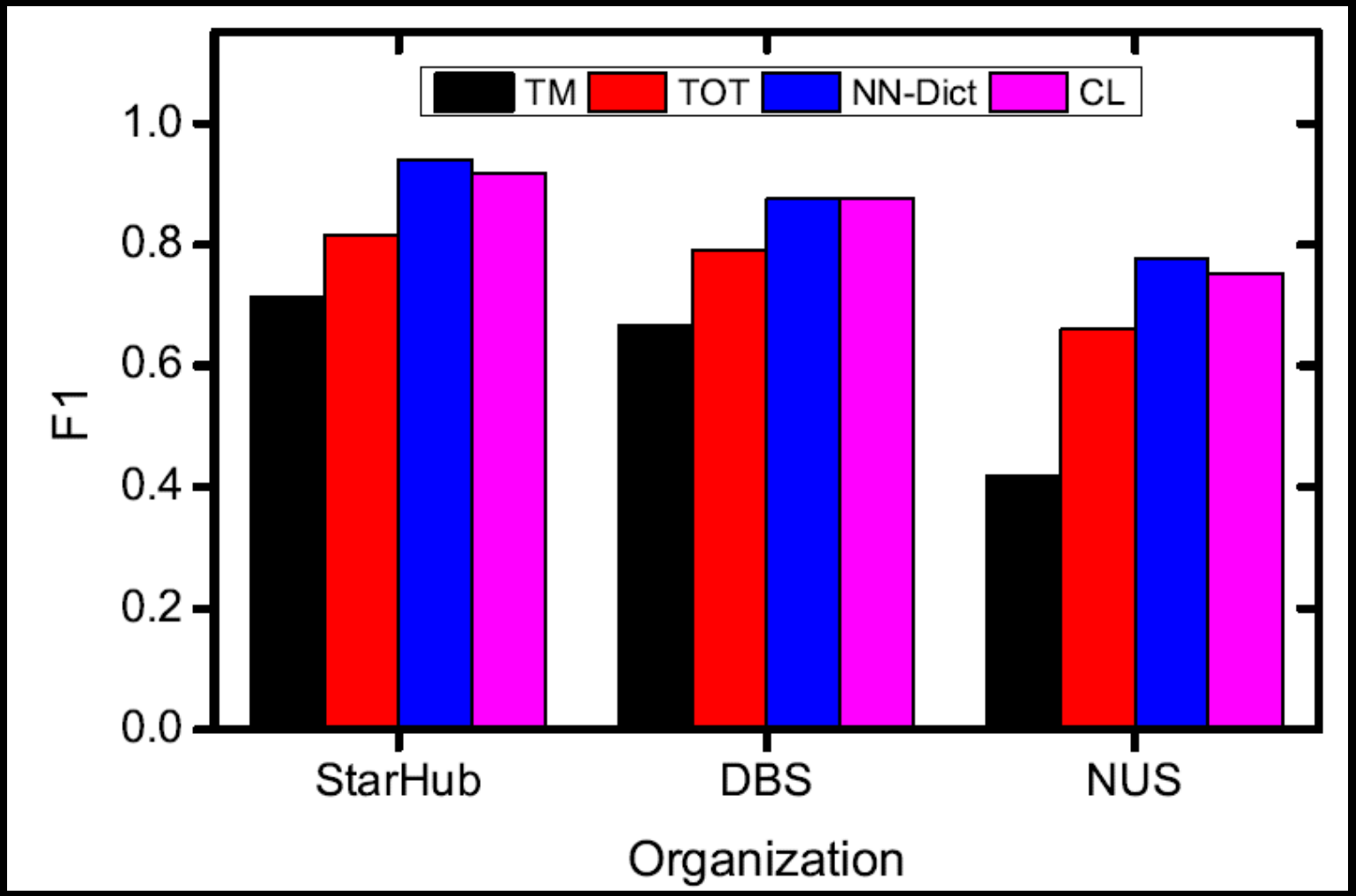
- Background and Motivation
- Organization-related Data Selection
- Hot Emerging Topic Detection
- Experiments and Analysis
- Conclusion and Future Work

Datasets

Organization	Time Duration	# Tweets	#Users	#Emerging Topic
StarHub	10 Oct - 9 Nov, 2012	51,708	15,792	24
DBS	15 Oct - 14 Nov, 2012	130,791	44,454	17
NUS	14 - 27 Oct, 2012	142,091	36,973	5

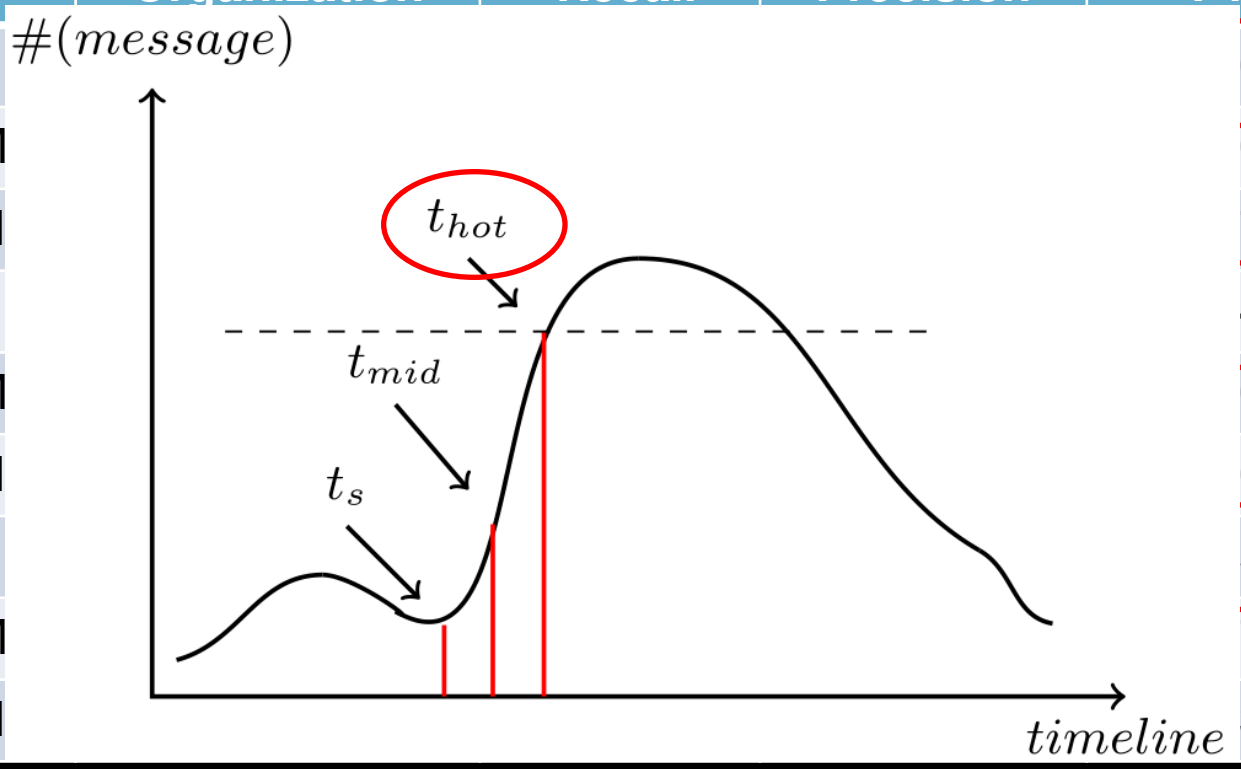


Performance of Topic Detection



Performance of Hot Emerging Topic Detection

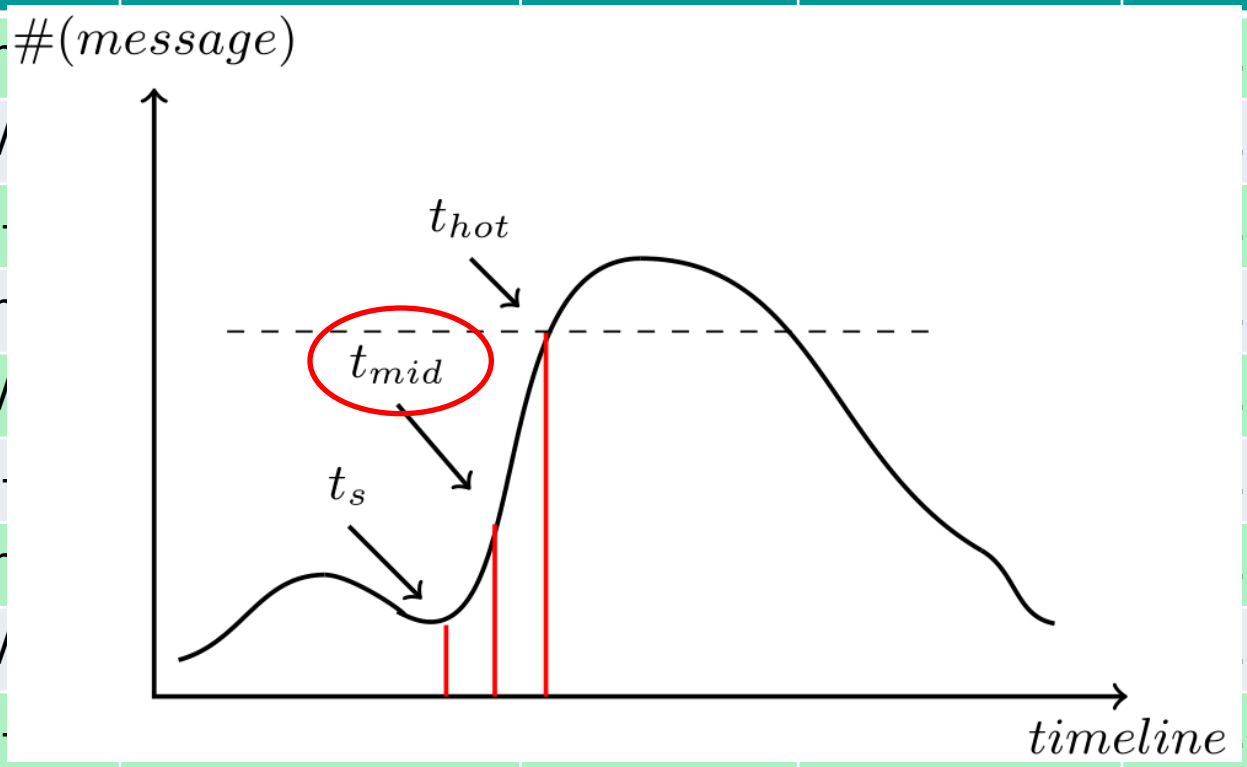
Methods	Organization	Recall	Precision	F1
CL+En				0
CL+TSVM				0
CL+Semi-N				7
CL+En				4
CL+TSVM				0
CL+Semi-N				0
CL+En				5
CL+TSVM				7
CL+Semi-N				3



$T_L = t_{hot}$

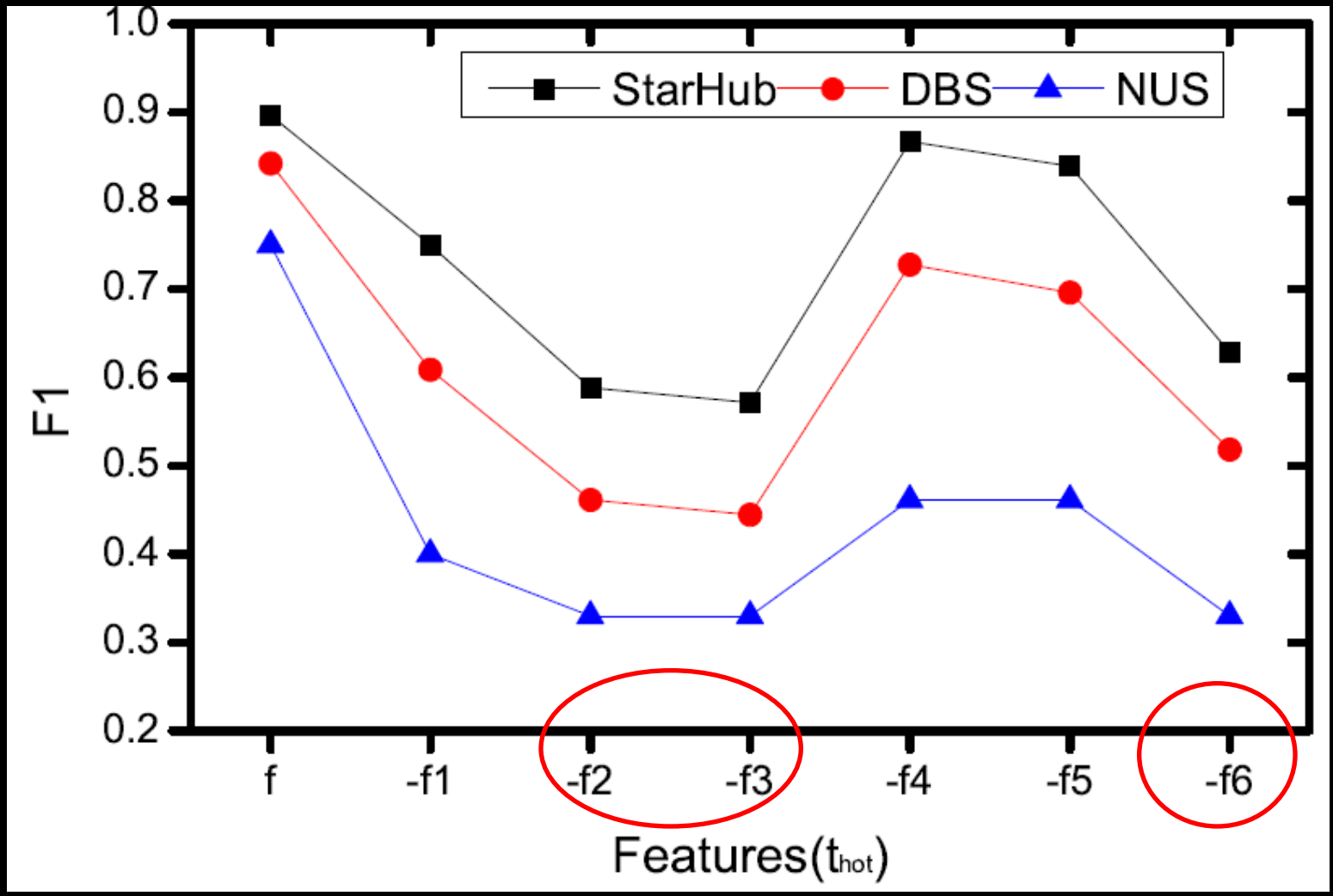
Performance of Hot Emerging Topic Detection

Methods	Organization	Recall	Precision	F1
CL+Er				77
CL+TSV				71
CL+Semi-				69
CL+Er				78
CL+TSV				74
CL+Semi-				70
CL+Er				57
CL+TSV				50
CL+Semi-				50

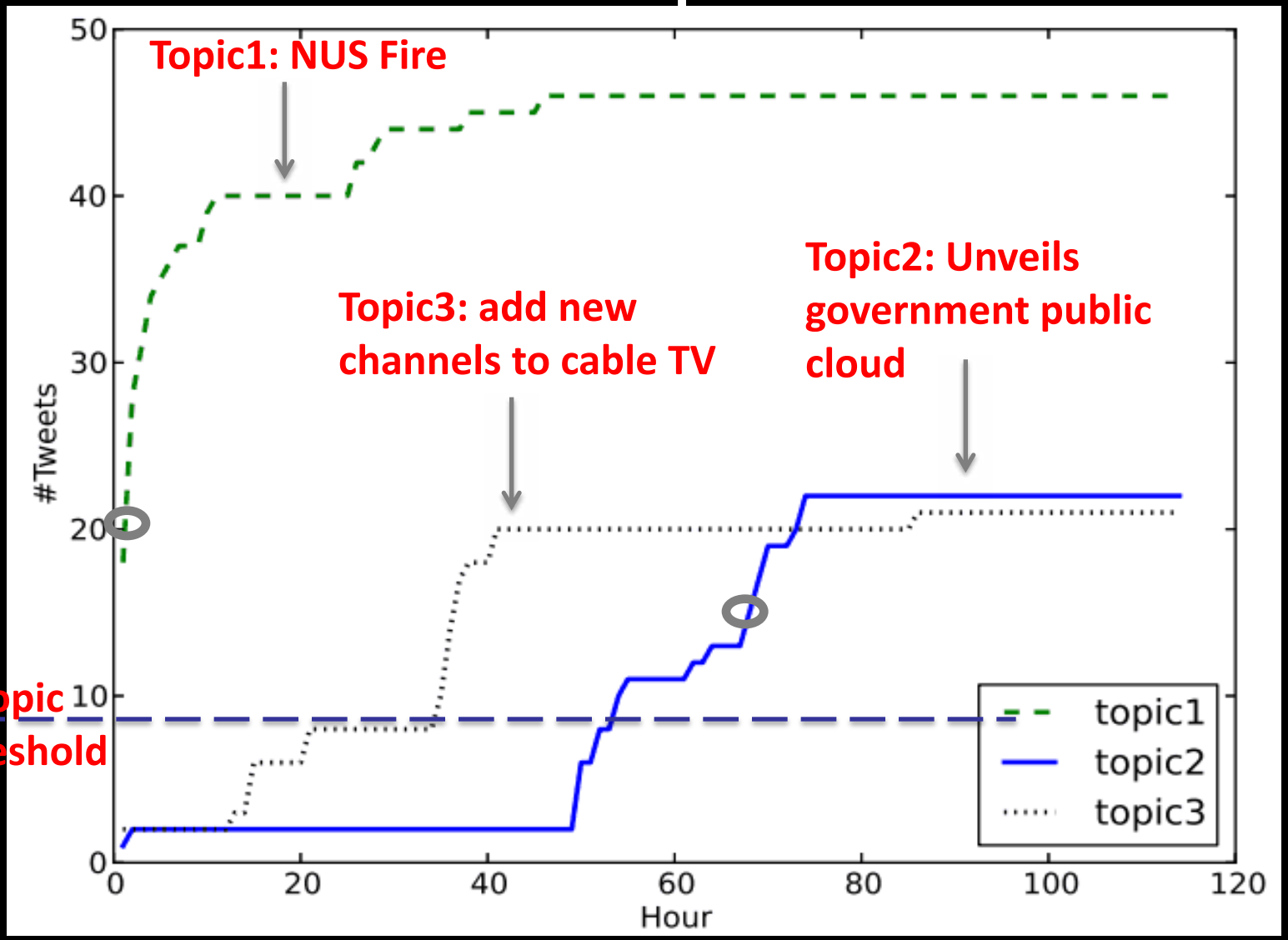


$T_L = t_{mid}$

Emerging Feature Analysis



Example


Topic1: NUS Fire
Topic3: add new channels to cable TV
Topic2: Unveils government public cloud
Topic Threshold
#Tweets
Hour
topic1
topic2
topic3
0
10
20
30
40
50
0
20
40
60
80
100
120

Outline

- Background and Motivation
- Organization-related Data Selection
- Emerging Topic Detection
- Experiments and Analysis
- Conclusion and Future Work

Conclusion

- Introduced four sources of crawling the organization data from multiple perspectives.
- Extracted non text emerging features to discover hot emerging topics.
- Developed semi-supervised learners to facilitate timely identification of hot emerging topics for organizations.
- Detected close to 90% of hot topics with a precision of over 70%. This is an encouraging results for hot emerging topic detection.

Future work

- Extend framework to general entities (e.g. People, Location, Events)

- Topic summary for end users.

Thank you!

Q&A