Hashtag-centric Immersive Search on Social Media

Yuqi Gao, Jitao Sang, Tongwei Ren, Changsheng Xu

State Key Laboratory for Novel Software Technology, Nanjing University

National Lab of Pattern Recognition, Institute of Automation, CAS

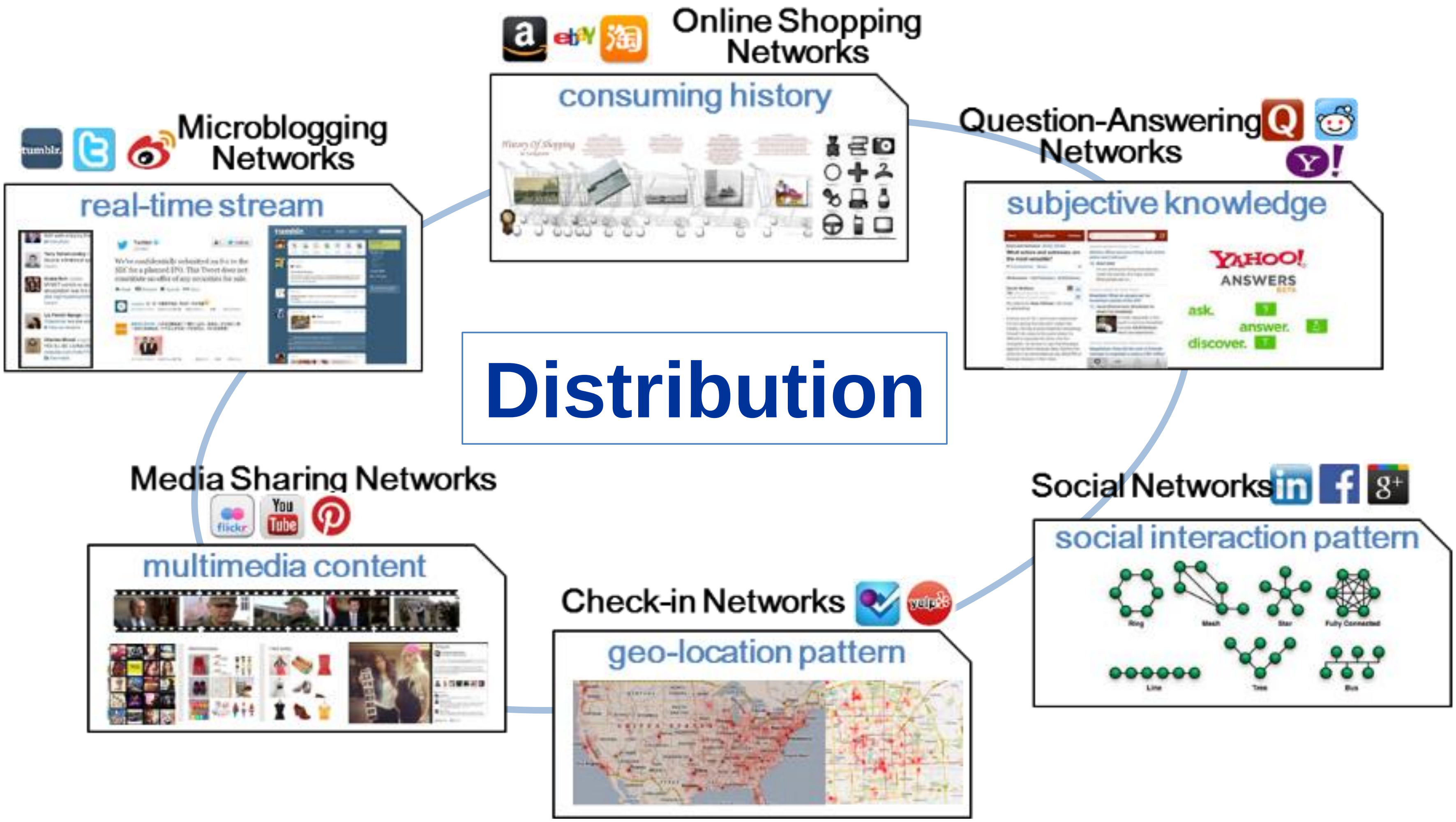
Outline

1. Motivation
2. Data Analysis
3. Solution
4. Experiment
5. Discussion

Outline

1 **Motivation**

Information: Multi-modality → Multi-source

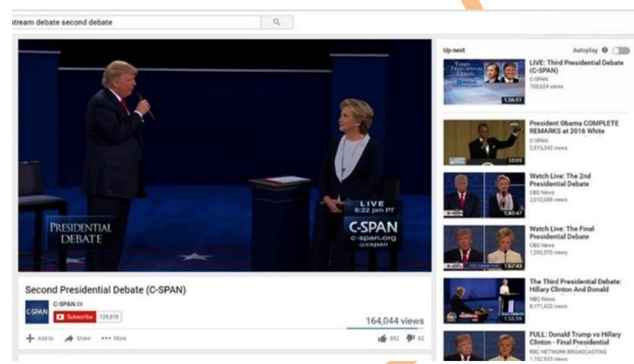


Information: Multi-modality ⟶ Multi-source

Access

ELECTION 2016

twitter



YouTube

Instagram

Immersive Information Access

Portal era → **Search era** → **Social media**

Portal era	Search era	Social media
All-web Passive index	**All-web** Active query	**Single OSN?** (Online Social Network)

Immersive Information Access

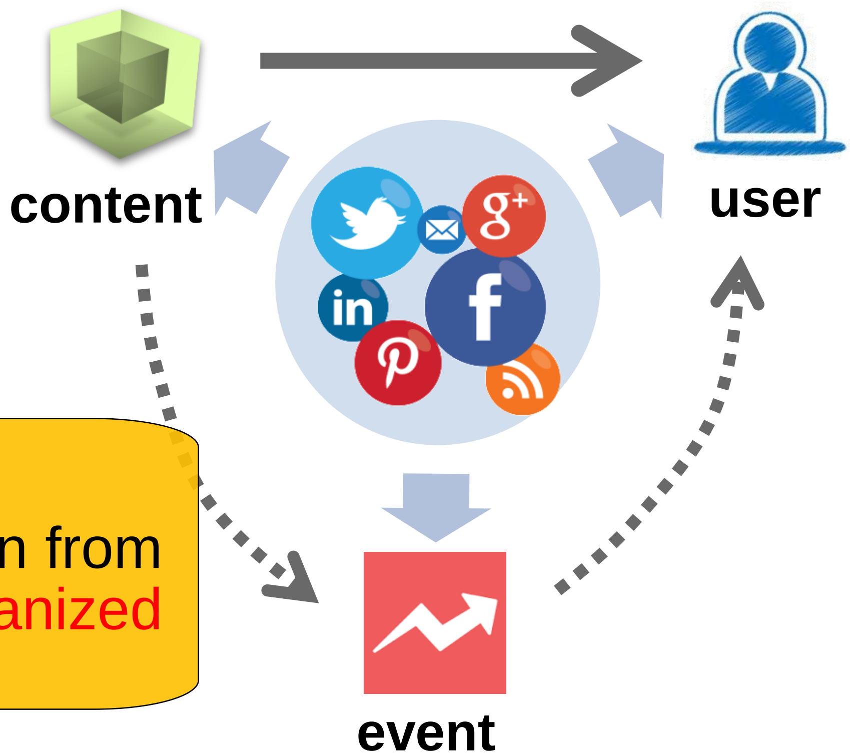
Portal era
All-web
Passive index

Search era
All-web
Active query

Social media
Cross-OSN
Immersive Service

YAHOO!

Google

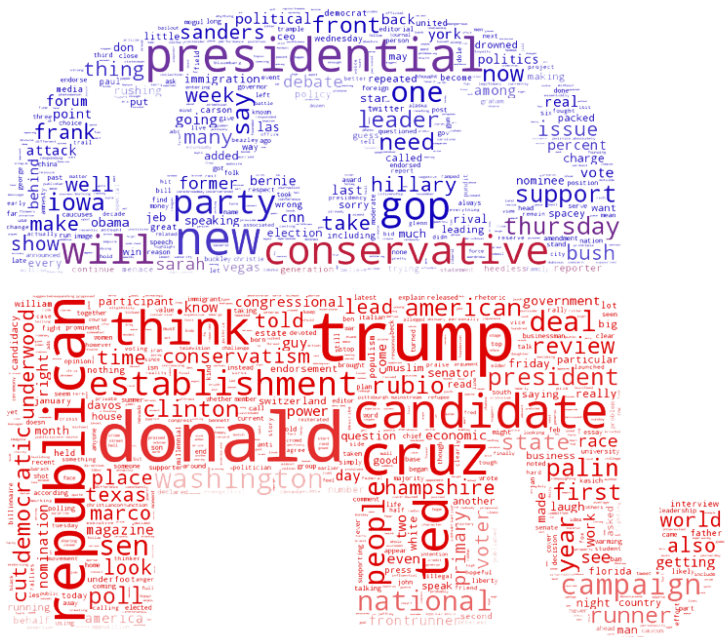


Cross-OSN search:
Given a query, the related information from different OSNs are aggregated, organized and demonstrated as search results.

Challenge: Relevance & Organization

1 **Relevance**
Noisy and biased results

2 **Organization**
Different modalities + search options



Election 2016

twitter

flickr

YouTube

relevance **interest** **#view**

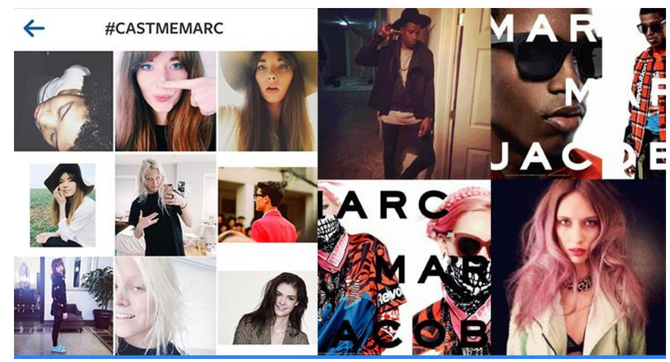
Hashtag: UGC Annotation + Management

1 Relevance
Noisy and biased results

2 Organization
Different search options + modalities

User-annotated

Management



naturally **cross-OSN**

Motivation:
We exploit hashtag as bridge for cross-OSN information integration and demonstration.

Outline

2 **Data Analysis**

Data Collection

Google Trends → 347 Queries → twitter / flickr / YouTube

Examples:

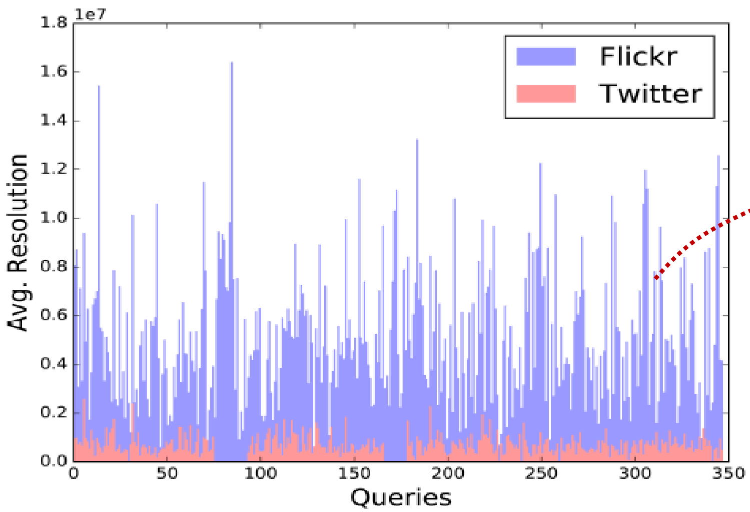
Election 2016	MLB
Overwatch	Arsenal
Halloween	Pokemon
Nobel Prize	Presidential debate

Question :

- What's the advantage of integrating different OSN search results? → **Single-OSN search comparison**
- How people use hashtag across different OSNs? → **Cross-OSN hashtag usage analysis**

Single-OSN Search Comparison: Information Richness

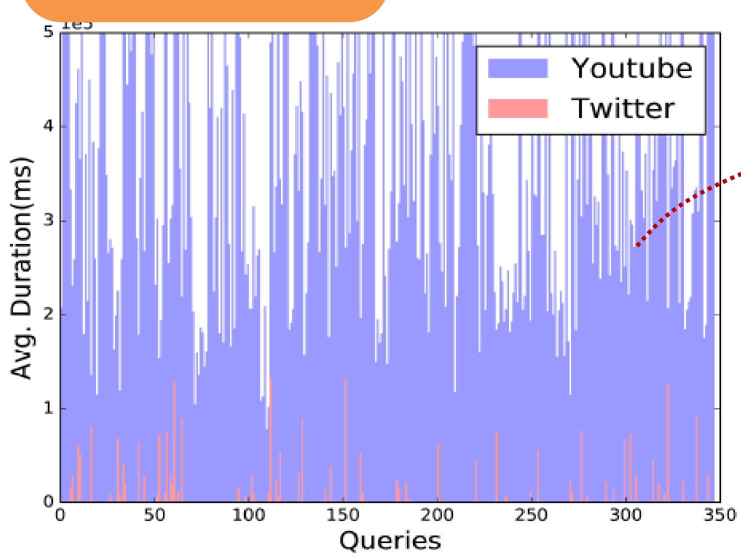
1 **Image Resolution**



480*360

900*853

2 **Video Duration**

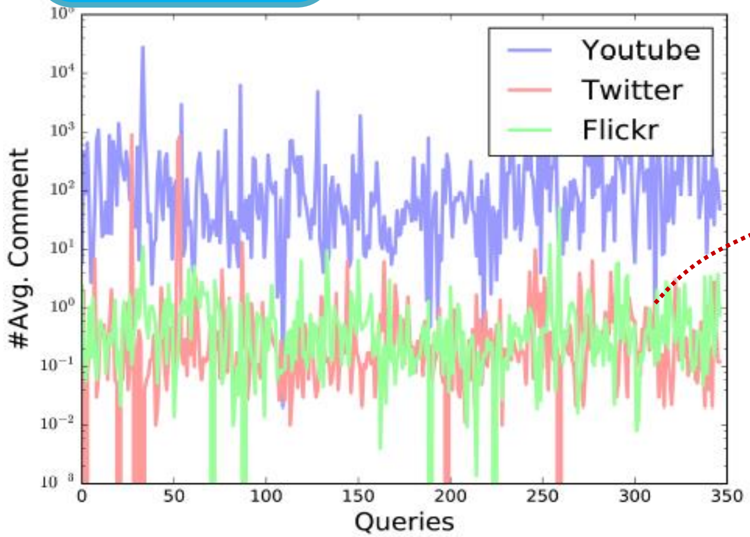


31s

21m47s

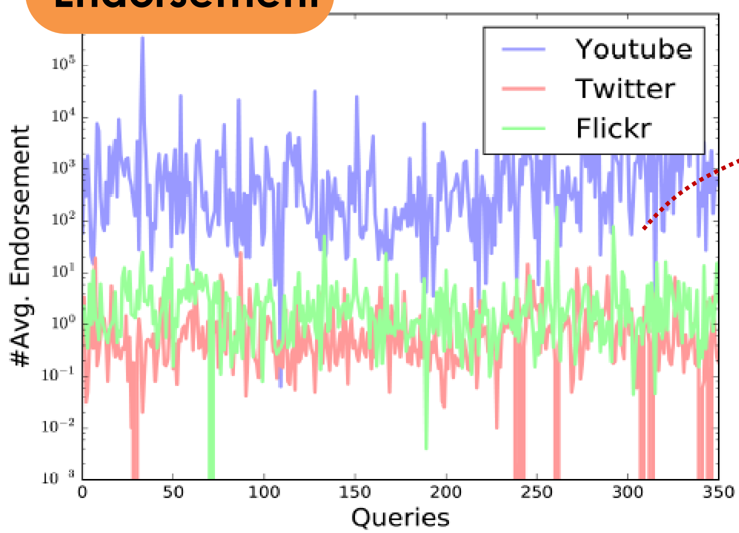
Single-OSN Search Comparison: User interaction

1 **Comments**



2 retweet **34 comment**

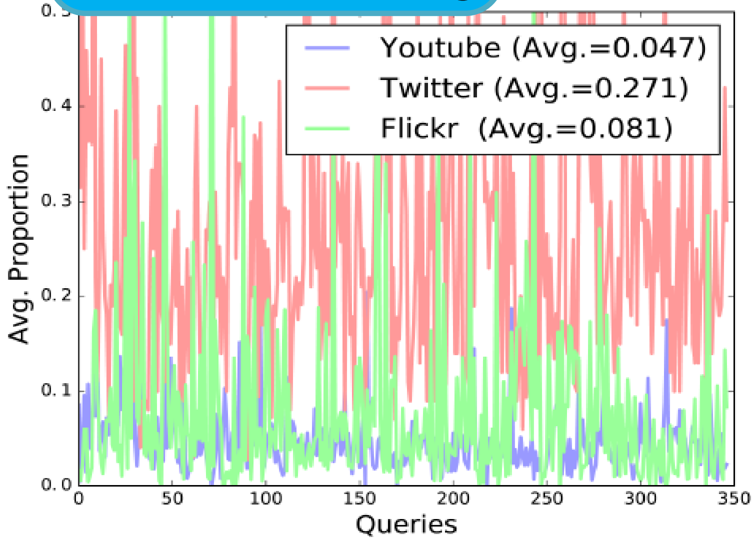
2 **Endorsement**



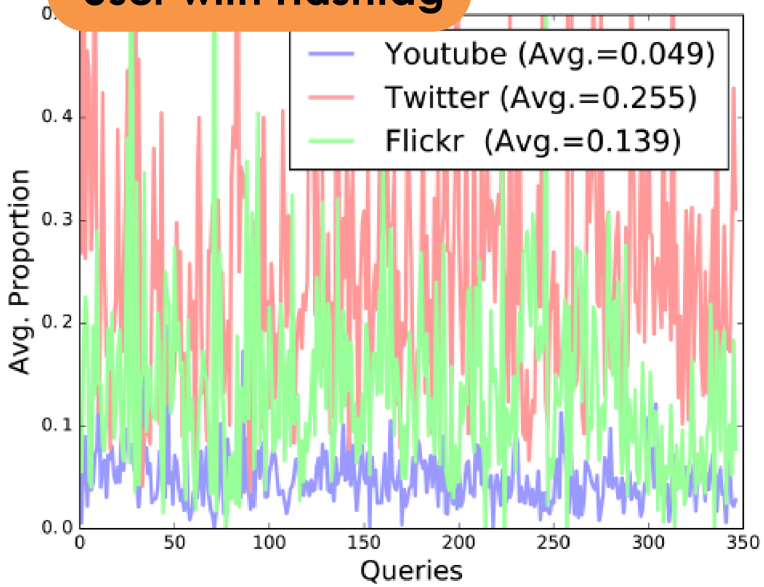
20 favorites **243 likes**

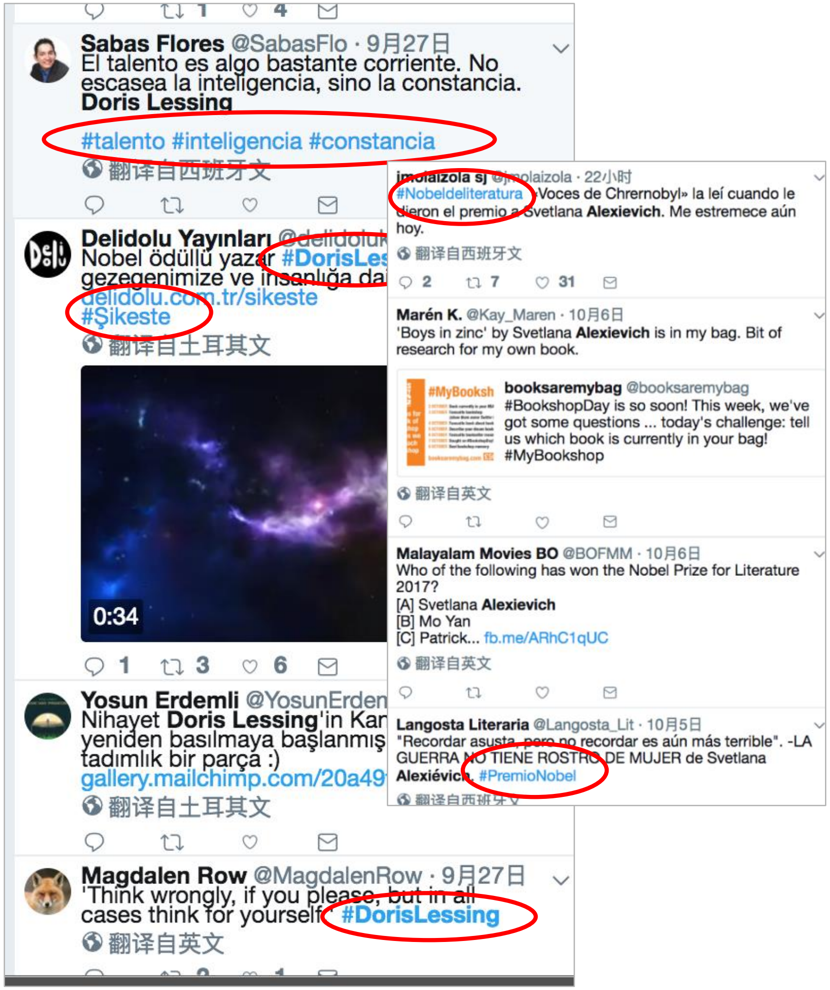
Cross-OSN Hashtag Usage Analysis: Popularity

1 **Result with Hashtag**



2 **User with Hashtag**





Cross-OSN Hashtag Usage Analysis: Diversity (1)

#Unique Hashtag (per query)

YouTube	Twitter	Flickr
17.77	28.42	27.61

Hashtag from search result of "Arsenal" (Partially)

Twitter	Flickr	YouTube
#EPL	#AlwaysTimeForCakevia	#afc
#RealMadrid	#alexanderhleb	#LIVE
#CristianoRonaldo	#wilshere	#Ludogorets
#AFC	#manchest	#PES2017
#Arsenal	#London	#MarqueeMatchups
#soccer	#MUFC	#EPL
#COYG		#COYG
#FootballNews		#SFC

Cross-OSN Hashtag Usage Analysis: Diversity (2)

NFr score of cross-OSN lists μ_1, μ_2

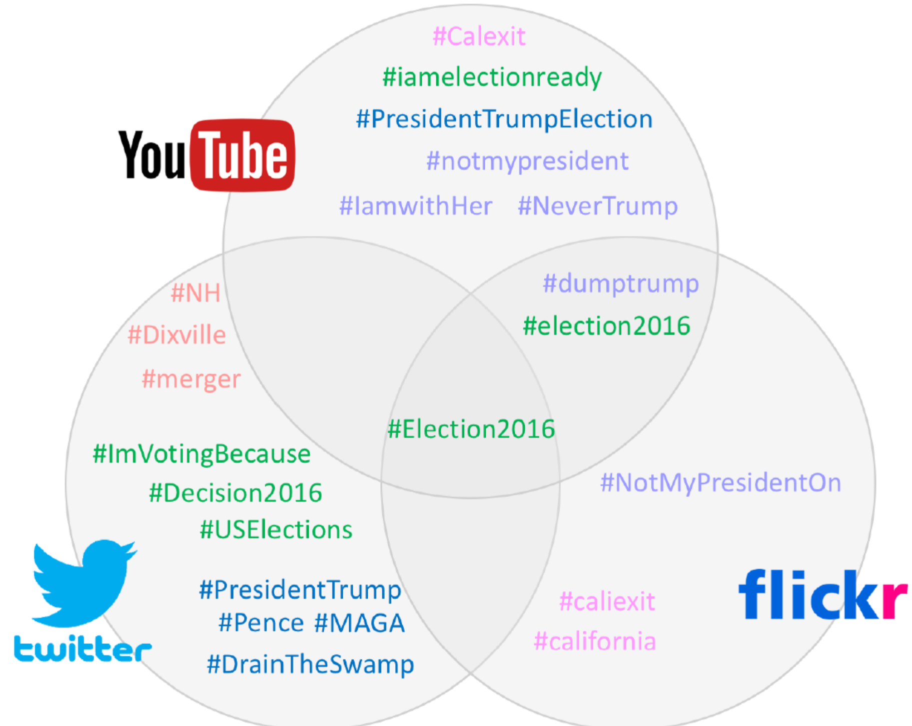
YouTube&Flickr	Twitter&YouTube	Flickr&YouTube
0.1006	0.0857	0.0375

$$NFr(\mu_1, \mu_2) = 1 - \frac{Fr^{|S|}(\mu_1, \mu_2)}{max \quad Fr^{|S|}}$$

$$Fr^{|S|}(\mu_1, \mu_2) = \sum_{i=1}^{|S|} |\mu_1(i) - \mu_2(i)|$$

$Fr^{|S|}$ equals $1/2|S|^2$ when $|S|$ is even

$Fr^{|S|}$ equals $1/2(|S| + 1)(|S| - 1)$ when $|S|$ is odd



Hashtags from search result of "Election 2016" from different OSNs

Data Observations

Single-OSN search comparison
- Information Richness
- User Interaction

Necessity
- Guarantee a better multi-modal search experience
- Enable more advanced features like social interaction

Cross-OSN hashtag usage analysis
- Popularity Analysis
- Diversity Analysis

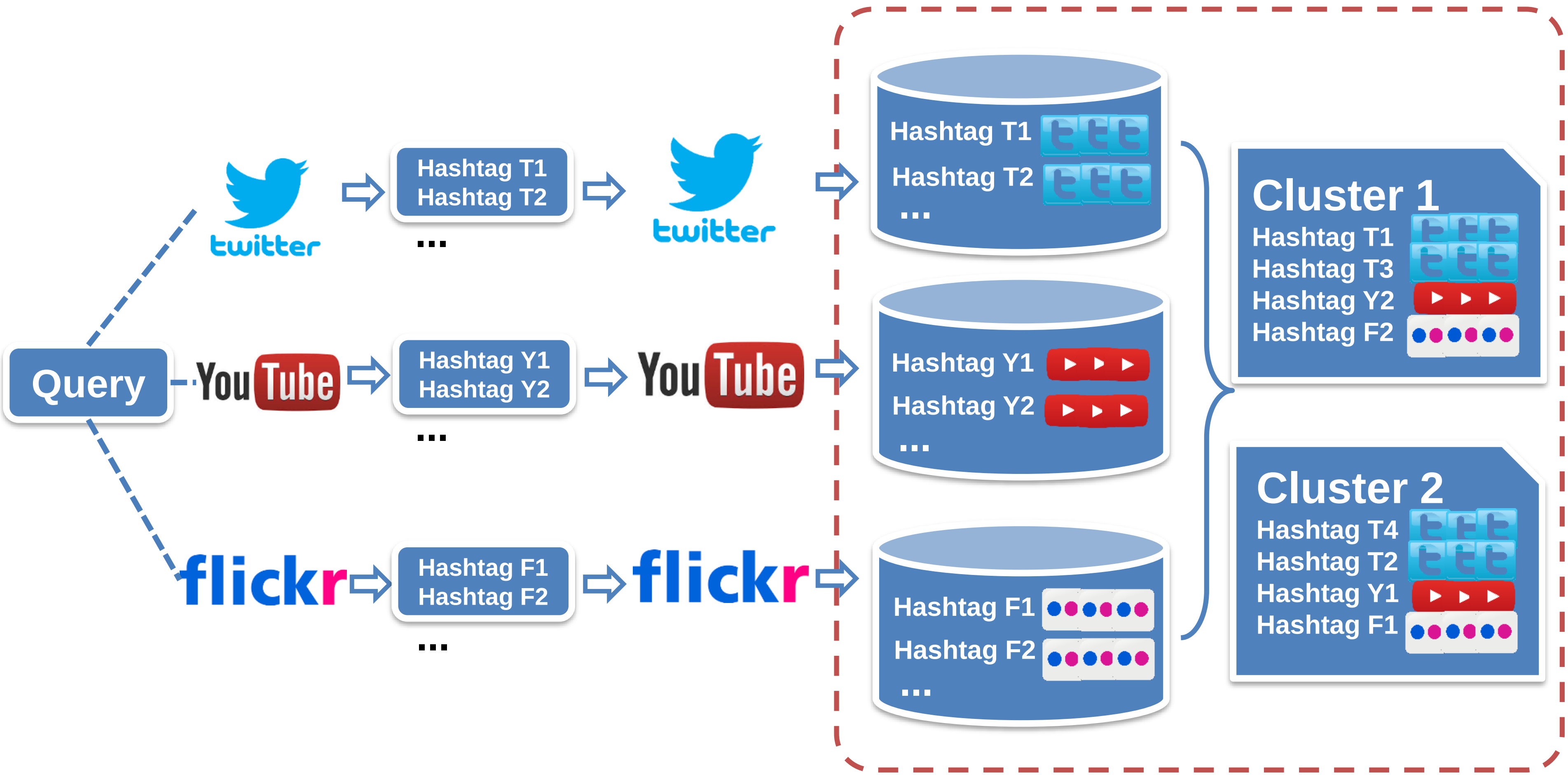
Feasibility
- Hashtag is widely used across different OSNs.

Inspiration to Solution
- **Multiple hashtag** → fine-grained semantic exploration;
- **OSN-distinctive hashtag** →higher topic level for integration.

Outline

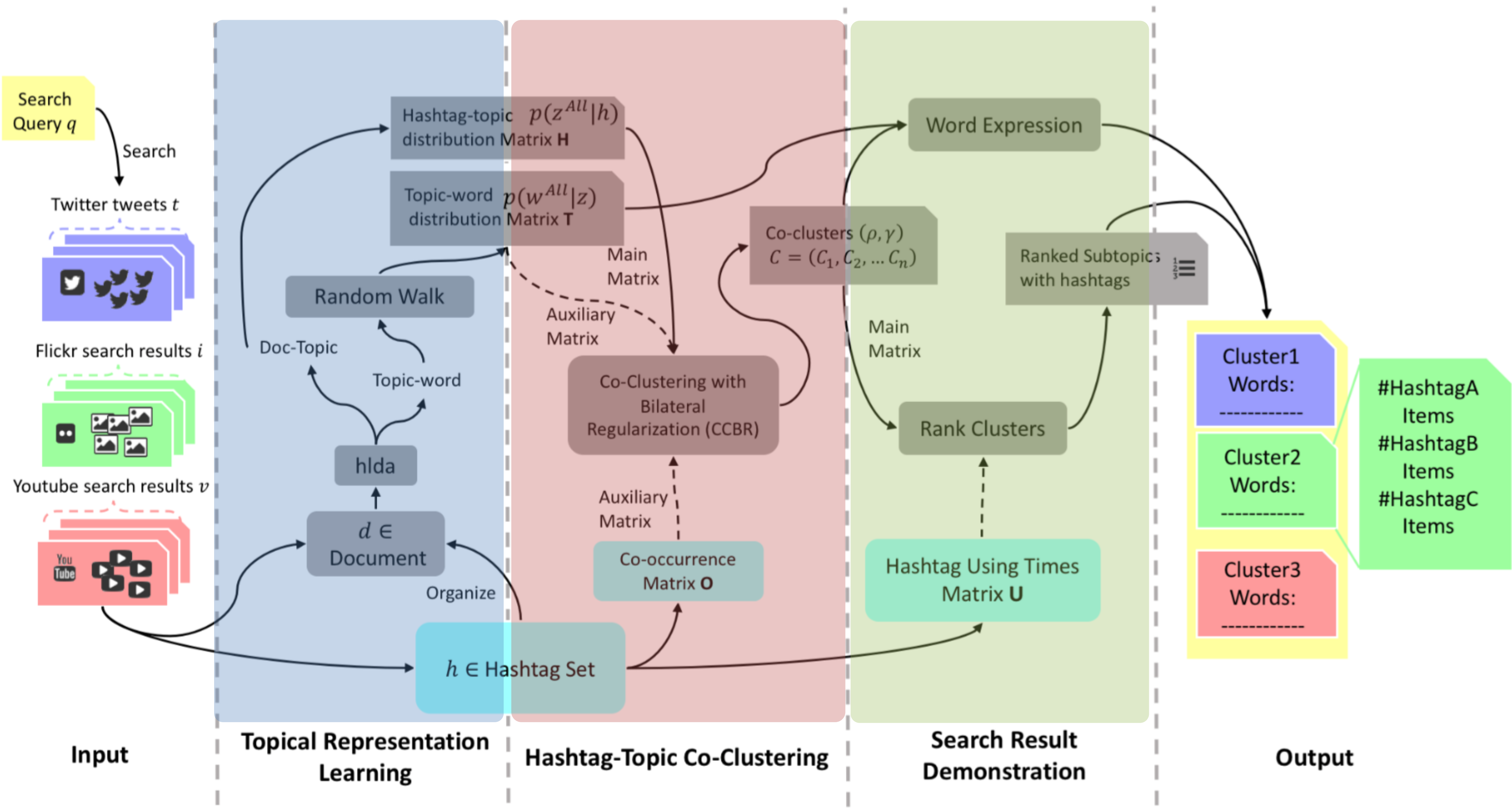
3 **Solution**

Data Flowchart

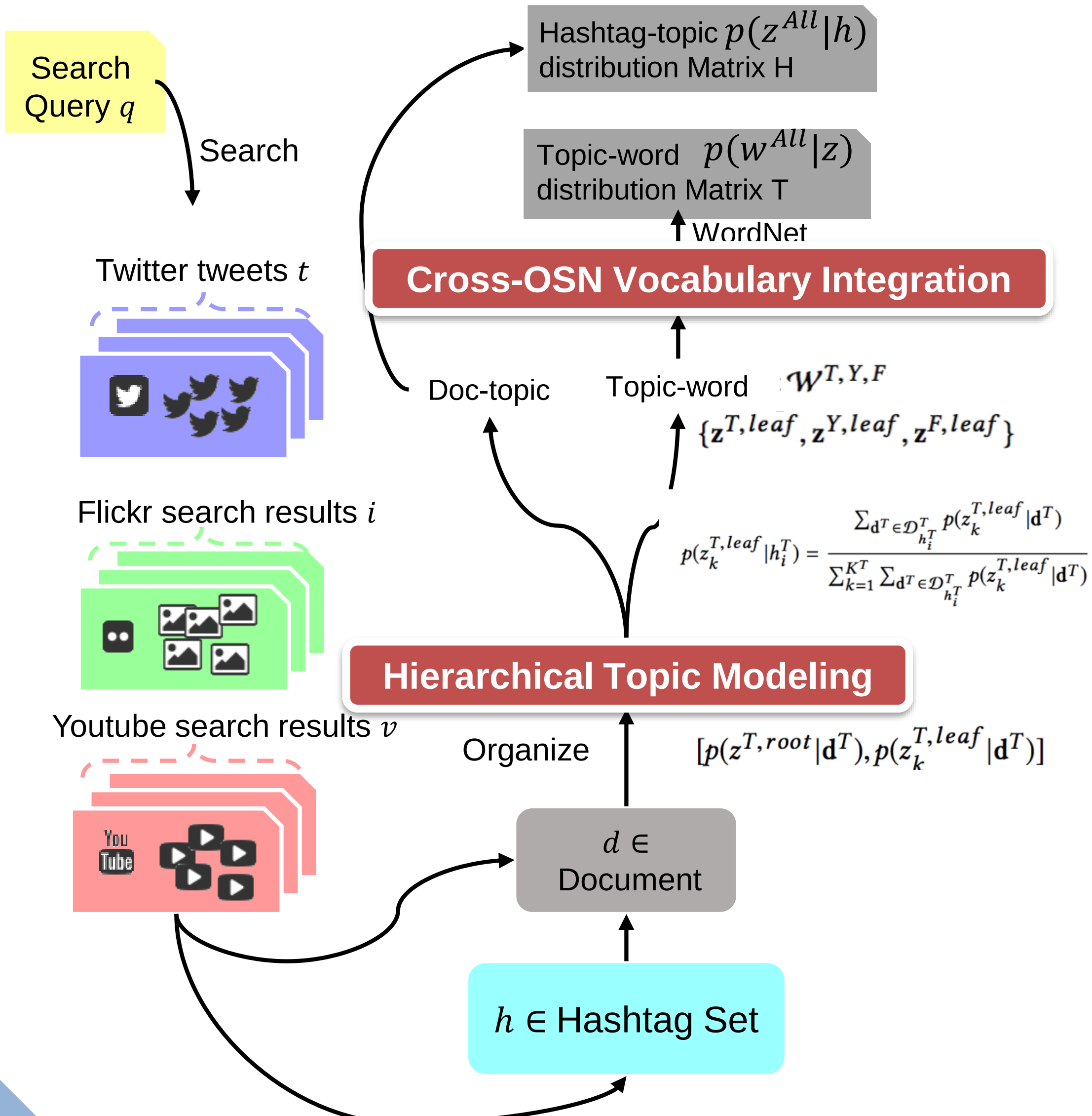


Hashtag Clustering

Solution Framework



Stage 1: Topical Representation Learning



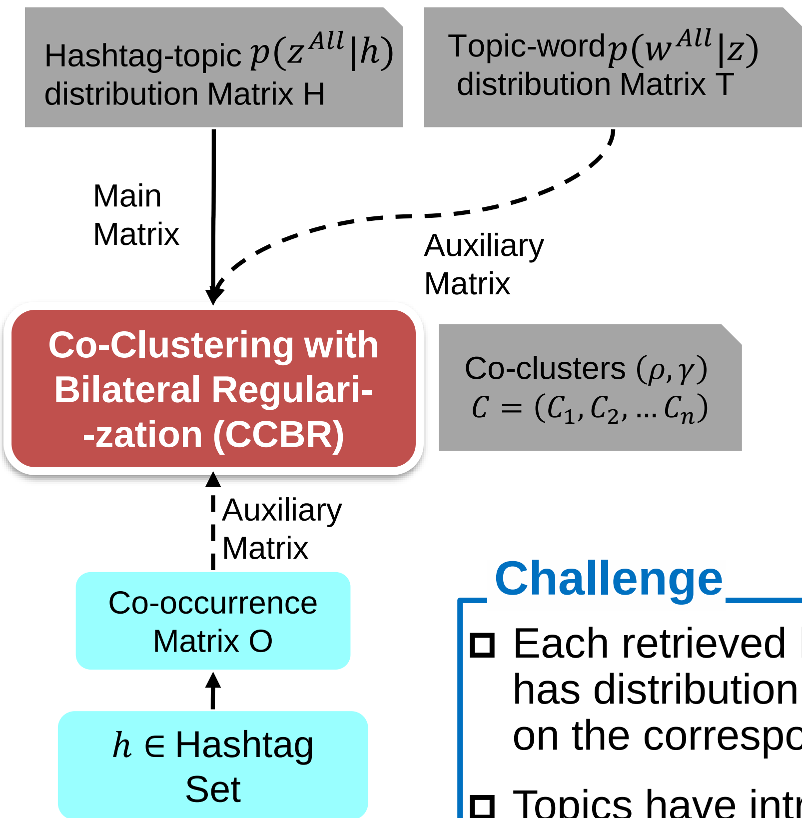
Challenge

- How to explore the fine-grained semantics under the same query topic?
- How to connect the biased vocabulary sets from different OSNs?

Solution

- **Hierarchical Topic Modeling**: utilize hLDA to represent each hashtag over the **subtopics** of the corresponding OSN $\{z^{T,leaf}, z^{Y,leaf}, z^{F,leaf}\}$
- **Cross-OSN Vocabulary Integration**: conduct random walk to propagate the subtopic relevance and obtain a **cross-OSN topic space** z^{All}

Stage 2: Hashtag-Topic Co-Clustering



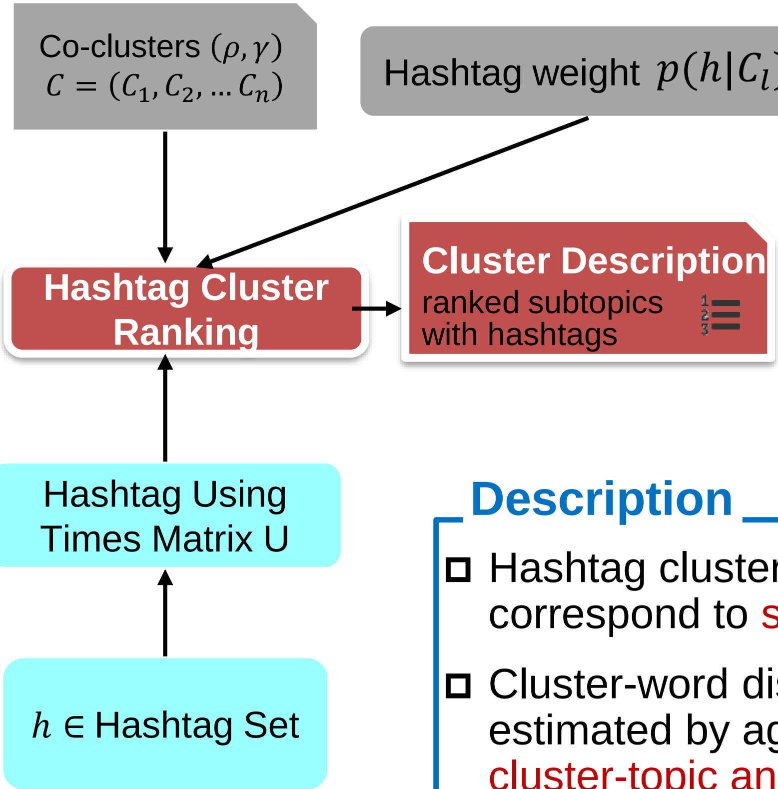
Challenge

- Each retrieved hashtag only has distribution over the topics on the corresponding OSN.
- Topics have intra-relation both within OSN and cross OSNs.

Solution

- **Hashtag-topic Co-Clustering:** cluster cross-OSN hashtags and topics simultaneously;
- **Co-Clustering with Bilateral Regularization (CCBR)**
  ① Regularizing **topic semantic relation** for column/topic clustering: topic-hashtag involvement + topic-word distribution;
  ② Regularizing **hashtag co-occurrence** for row/hashtag clustering: hashtags that co-occurring in the same item have high probability to contribute to the same subtopic.

Stage 3: Search Result Demonstration



Description

- Hashtag cluster is expected to correspond to subtopics.
- Cluster-word distribution is estimated by aggregating the cluster-topic and topic-word distribution.

$$p(w|C_l) = \sum_{z^t \in z^{all}} p(z^t|C_l) \cdot p(w|z^t)$$

Organization

- **Cluster - Hashtag - Item**
- Items within unique hashtag are organized **chronologically.**
- Hashtags within unique cluster are organized via **cluster-hashtag weight** $p(h|C_l)$.
- Cluster Ranking:
  ① Popularity constrain: cluster with more hashtag-annotated items in the search result set should be ranked higher;
  ② Smooth constrain: clusters with similar semantic relation deserve close ranks.

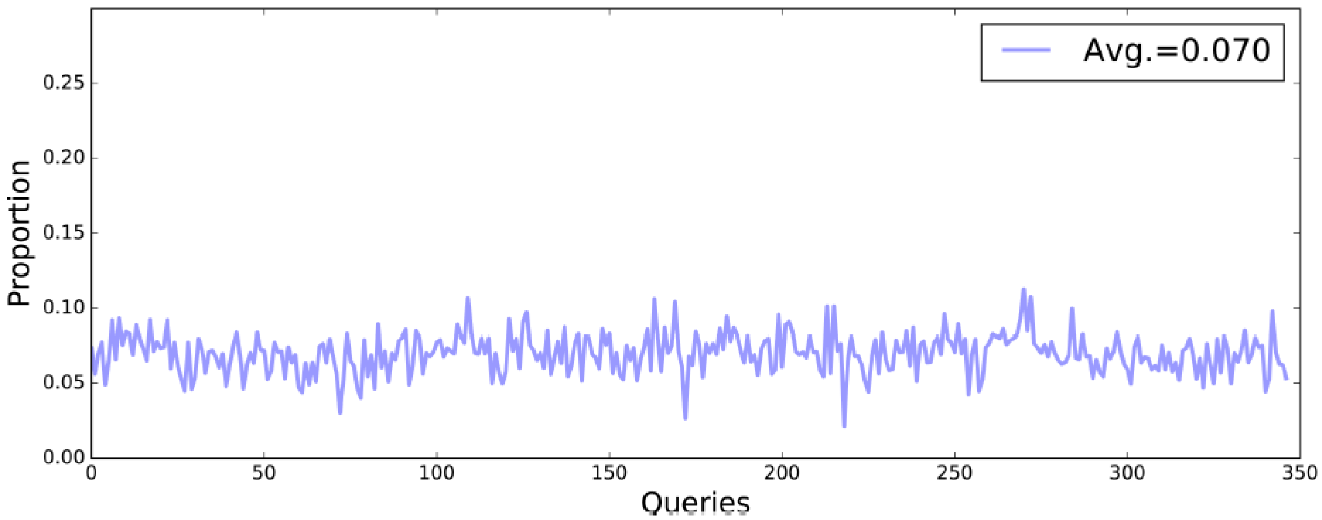
$$\kappa_{ij} = exp\left(-\frac{\sum_{z^t \in z^{all}} (p(z^t|C_i) - p(z^t|C_j))^2}{2\sigma^2}\right)$$

Outline

4 **Experiments**

Vocabulary overlap proportion.

It is shown only about **7%** vocabulary is shared between the three OSNs, which validates the necessity for random walk-based vocabulary integration.



Results of Topical Representation Learning

Derived topics for query "Election 2016".

- The discovered topics have a wide coverage.
- Random walk connects between different vocabulary spaces and enhances the topic representation with cross-OSN words.

Platform	Topic
Twitter	ranks,states,worldpolitics,meet ...
	published,newsletter, history,online,...
Flickr	million,trump,election,votes,riches...
	nation,people,language,americanelection...
Youtube	trump,donald,live,rally,presidenttrump,...
	country,children, feel,hold, citizens,...

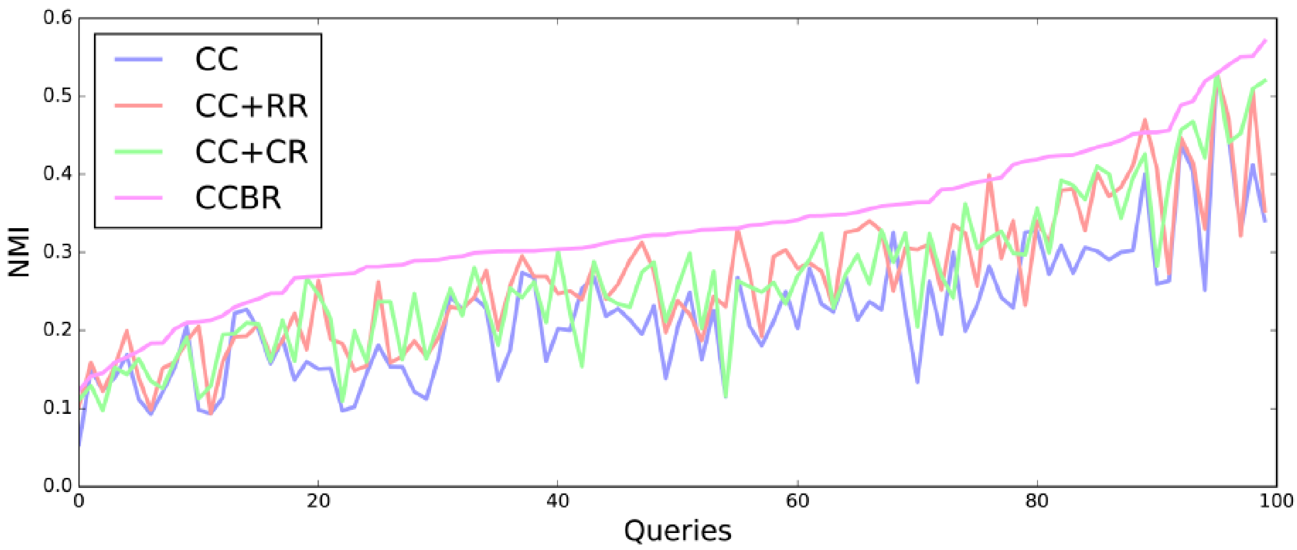
Twitter

Flickr

YouTube

Results of Hashtag Clustering

- **CC**: original Bregman Co-Clustering
- **CC+RR:** Co-Clustering with hashtag co-occuRrence Regularization
- **CC+CR**: Co-Clustering with intra-topic Correlation Regularization
- **CCBR**: Co-Clustering with Bilateral Regularization

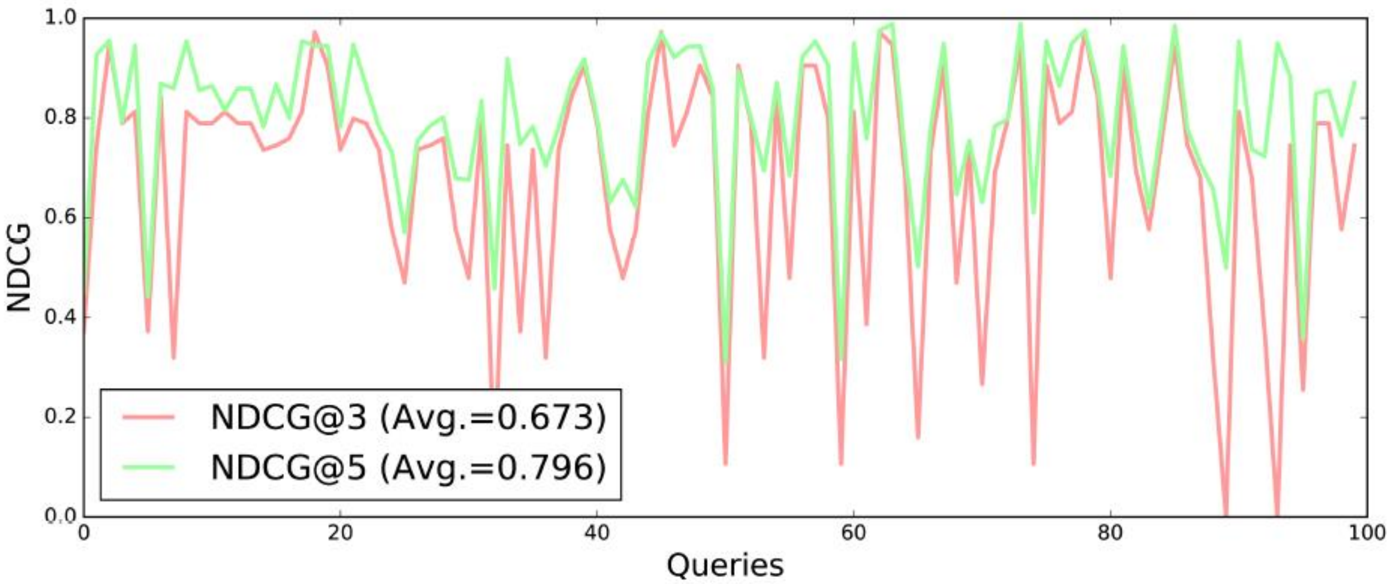


Negative Pearson correlations indicate that queries with larger #hashtag have a lower NMI

Method	CC	CC+CR	CC+RR	CCBR
NMI & #hashtag	-0.689	-0.633	-0.643	-0.526

Pearson correlation coefficient

Search Result Demonstration

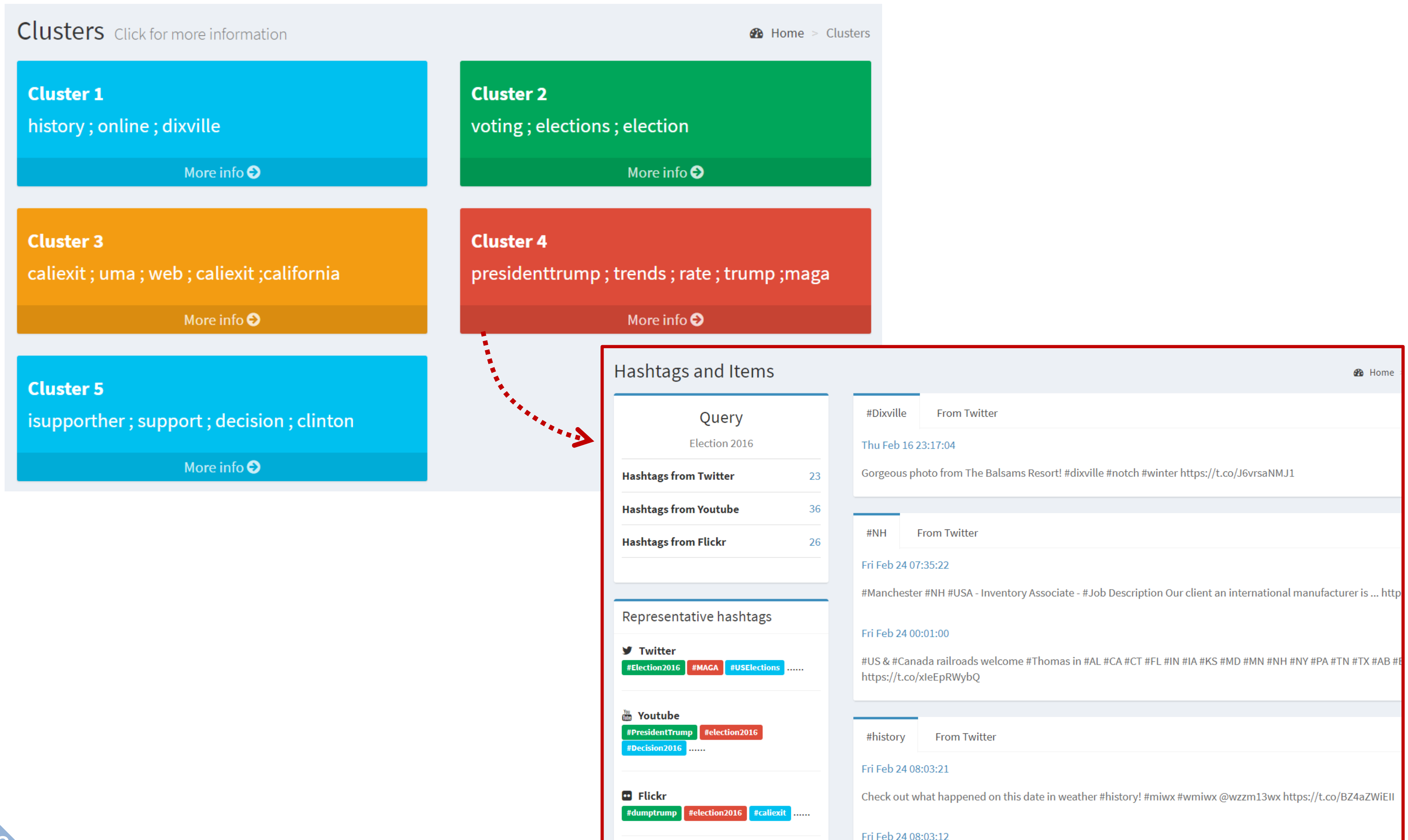


NDCG of hashtag cluster rank for different queries

- NDCG@5 = 79.6%: the cluster rank solution is reasonable and practical in real application.
  (most queries have a ground-truth of 5-9 clusters).
- When only examining the rank-1 cluster, the proposed rank method achieves performance with average NDCG@1=37%.

Demonstration: Cluster-Hashtag-Item Hierarchy

https://hashtagasbridge.github.io/Hashtag



Contribution and Limitation

Contribution:

- Discussed and analyzed the Cross-OSN hashtag usage;
- Positioned the problem of Cross-OSN immersive search, and introduced a preliminary hashtag-centric solution.

Limitation:

- **Time cost**: the first stage of cross-OSN topical representation learning prevents a practical solution.
- **Narrow focus on event queries**: remains unknown whether can apply to general queries.
- **Insufficient utilization of contextual data**, e.g., time (enable topic evolution), hyperlink (for better clustering).
- **Lack of exploiting representative OSN features**, e.g., Twitter list, Flickr group, YouTube channel, authoritative Users.

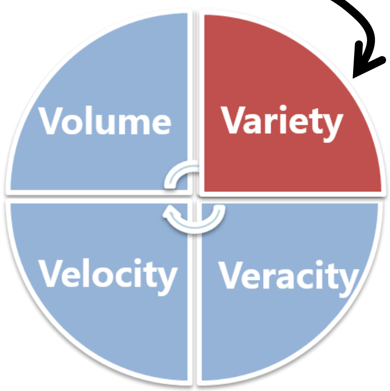
Outline

5

Discussion:
Cross-Modal ⇨ Cross-OSN

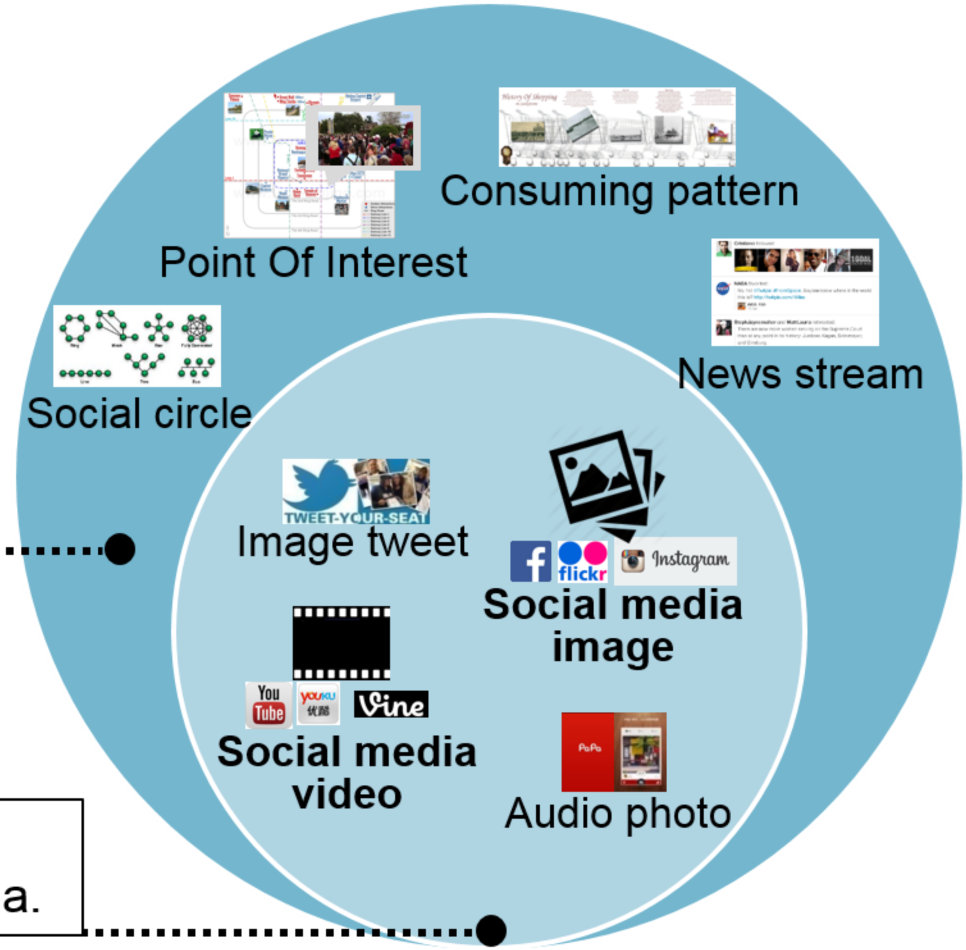
Social Multimedia: Special ➡ General

Multimedia:
Media of different **types** and **sources**



General Social Multimedia:
all types of individual and aggre-
-gated data on social media
(beyond modality, beyond object).

Special Social Multimedia:
multi-modal objects on social media.



General Social Multimedia Analysis

Web Multimedia

- **WEB1.0**
- Data is professionally edited.
- Core problem is media understanding.

Social Multimedia

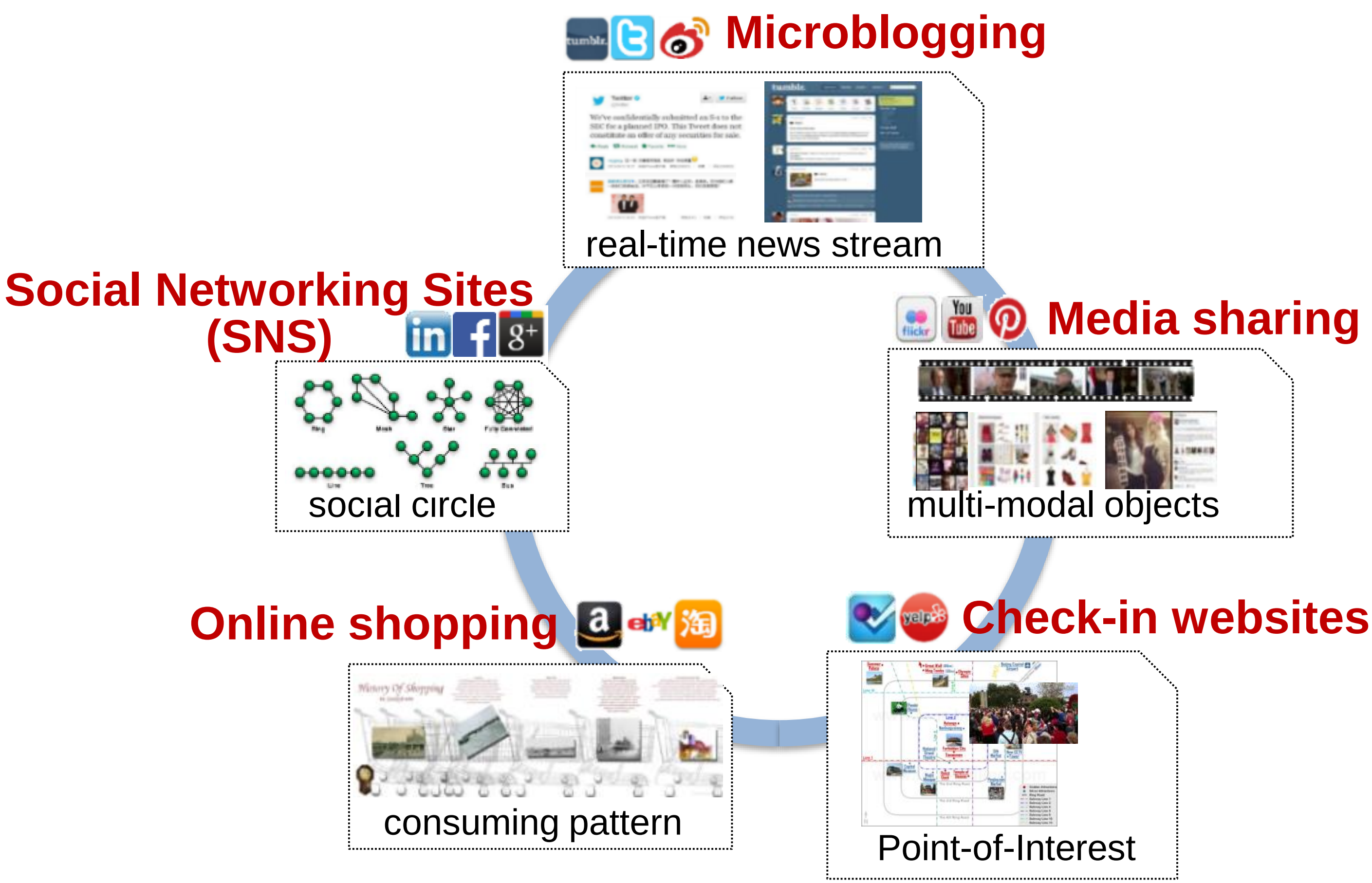
- Interaction is key.
- User contributes to data generation.
- User modeling is one basic problem.

General Social Multimedia

- **Media type**: beyond modalities.
- **Granularity**: beyond objects.
- **Association**: beyond semantic.

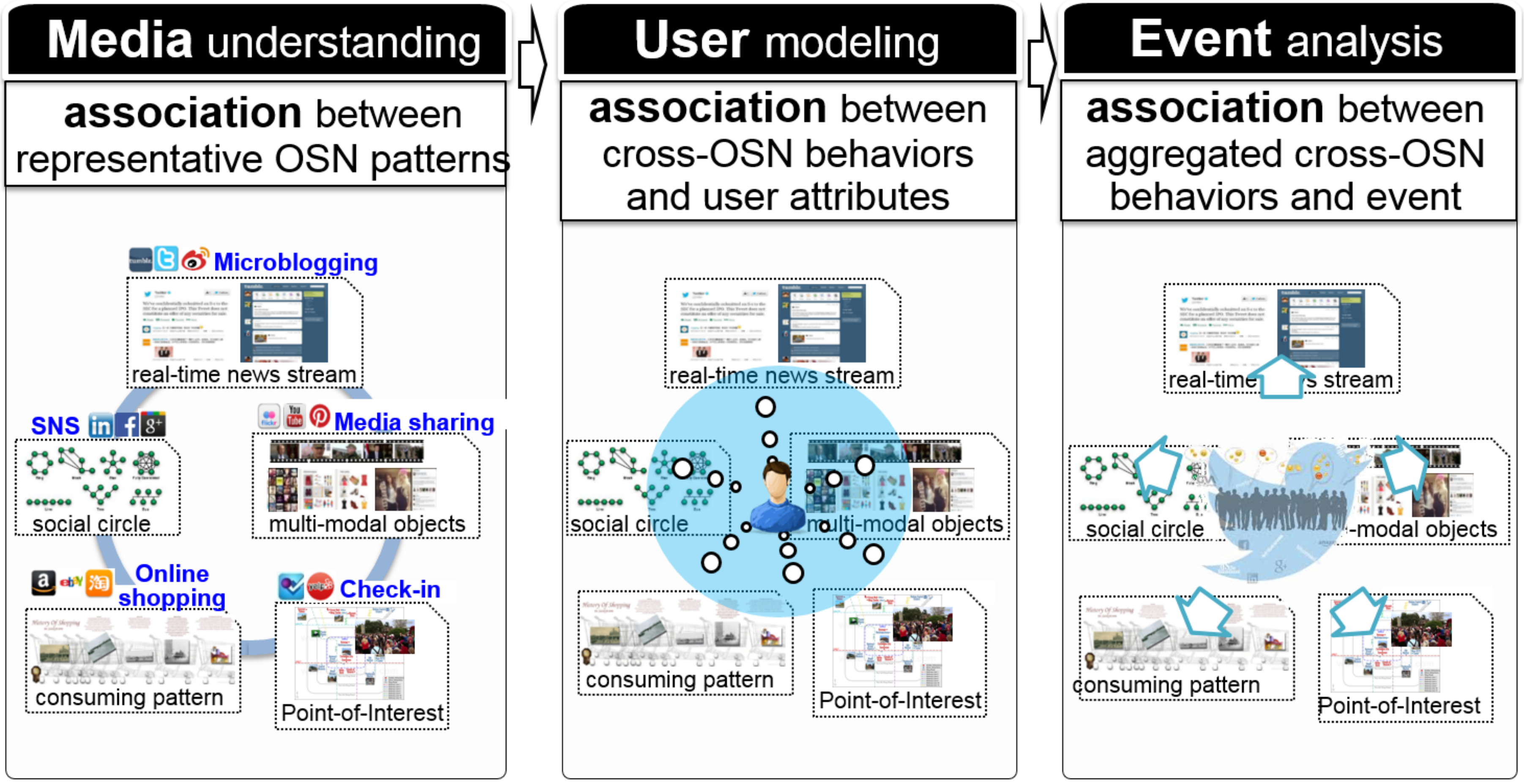
Cross-OSN: an Instantiation

Cross-OSN (Online Social Networks)



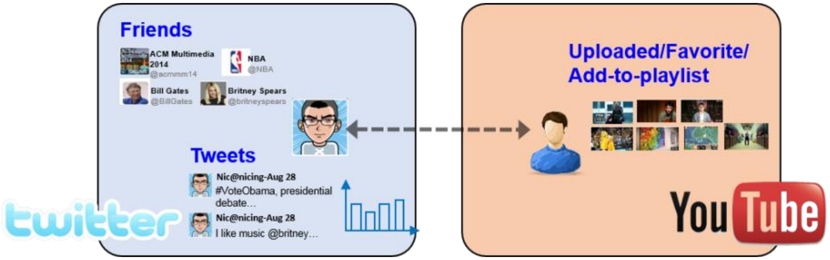
General Social Multimedia distributes among OSNs.
Cross-OSN provides both dataset & application scenarios.

Cross-OSN Data Mining

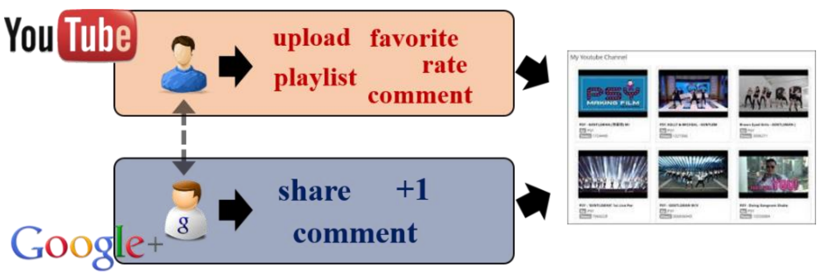


Jitao Sang, Changsheng Xu, Ramesh Jain.
Social Multimedia Mining: from Special to General. *ISM 2016*, Invited Paper.

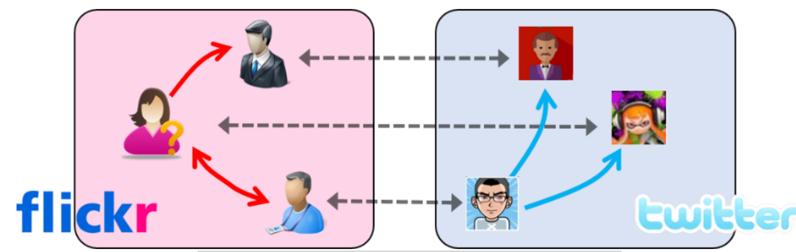
Cross-OSN Dataset (User-centric)



CrossOSN-U: Hetero



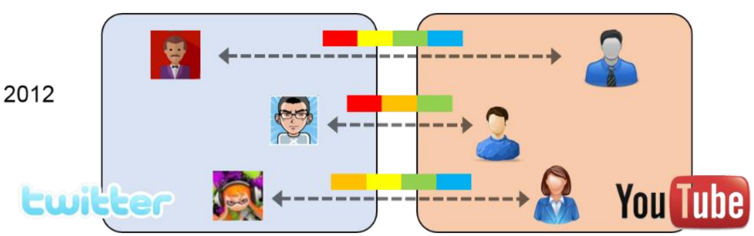
CrossOSN-U: Homo



CrossOSN-U: SN

Event List

- T1: US presidential election 2012
- ...
- T13: Iphone5 release
- ...



CrossOSN-U: Event

Thank you!

Questions?