Learning Similarity Metrics for Event Identification in Social Media

Hila Becker, Luis Gravano
Columbia University

Mor Naaman
Rutgers University

"Event" Content in Social Media Sites

"Event" Content in Social Media Sites

"Event"= something that occurs at a certain time in a certain place [Yang et al. '99]

Popular, widely known events

Smaller events, without traditional news coverage
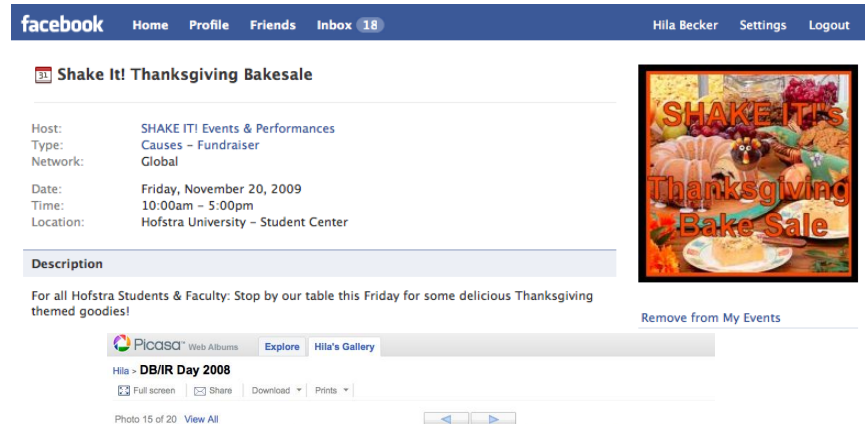
Identifying Events and Associated Social Media Documents

Identifying Events and Associated Social Media Documents

- Applications
  - Event browsing
  - Local event search
  - …

- General approach: group similar documents via clustering
  Each cluster corresponds to one event and its associated social media documents

Search Events

Popular Events

All Points West Music Festival

Friday July 31, 2009 - Sunday August 2, 2009
Liberty State Park
200 Morris Pesin Drive
Jersey City, New Jersey 07305 Get Directions

The second annual ALL POINTS WEST
MUSIC & ARTS FESTIVAL returns to Liberty State Park in Jersey City,
New Jersey Friday, July 31, Saturday, August 1 and Sunday, August 2
with Beastie Boys, Tool and Coldplay set to headline the three-day,
multi-stage music and arts festival. ALL POINTS WEST will also
feature performances from Echo & The Bunnymen, My Bloody
Valentine, Vampire Weekend, MGMT, MSTRKRFT, Gogol Bordello,
The Black Keys, Fleet Foxes, Neko Case, Yeah Yeah Yeahs, Arctic
Monkeys and more than 65 artists appearing throughout the 80-acre
historic location that has panoramic views of the Statue of Liberty, Ellis
Island and the Manhattan skyline. This year, the Queen of the Valley
stage has been changed to a Coachella-like tent creating an
environment that will showcase comedy, DJs, electronic acts and
bands.

From chris tuite   From tina gao   From surgery   From broken crow   From aguillaumee   From chris tuite

all 1,141 documents

Yankees vs Red Sox

Friday July 3, 2009 - Sunday July 5, 2009 from
7:05pm - 1:35pm
Fenway Park
4 Yawkey Way
Boston, Massachusetts Get Directions

Yankees vs Red Sox Tickets
Fenway Park in Boston, MA

Don't miss the chance to see the New York Yankees take on the Boston
Red Sox! The Yankees has the potential to win it all and will certainly be
looking to do just that. Get your Red Sox vs Yankees tickets now to see
the Bronx Bombers!
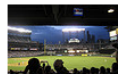
From BostonCityWa...   From stef_monk   From fnkymnky   From Lipski

all 345 documents

Recommended Events

Saturday, Aug 15

Universal Hip Hop Parade
Aug 15, 3:00pm at Marcus Garvey Boulevard a..., Brooklyn, ny

Come join the world's longest celebration of Hip Hop Culture. The 10th Annual
Universal Hip-Hop Parade steps off in the Bed-Stuy neighbourhood of Brooklyn
on Saturday August 15th at 3:00 PM from...

Monday, Nov 16

Web 2.0 Expo New York
Nov 16 at Jacob K. Javits Conventio..., New York City, ny

Web 2.0 Expo is for the builders of the next generation web: designers, developers,
entrepreneurs, marketers, business strategists, and venture capitalists, people who have
experiences to share and a...

Saturday, Jun 27

Go Long Island Community Expo
Jun 27 at The Long Island Marriott ..., Uniondale

More than a job fair. Helping people and businesses too. On Saturday June 27,
2009 Long Islanders will arrive at the Marriott in Uniondale looking for answers
and will make choices about who and...

Tuesday, Jun 23

FREE Money for Your Business
Jun 23, 3:00pm at CAMBA, New York

Come learn what Free Money is, where it can be found and how to apply for it.

Friday, Sep 11

Couture Fashion Week - New York
Sep 11, 3:00pm at The Waldorf Astoria, New York, ny

Couture Fashion Week - New York Waldorf Astoria, Park Avenue, New York, USA Couture
Fashion Week is a multiple-day upscale event presenting international couture and luxury
fashion shows. It also...

Wednesday, Sep 2

The 9th Annual New York International Salsa Congress - Dance & Music F...
Sep 2, 8:00pm at Hilton New York, New York City, ny

The New York International Salsa Congress is New York City's PREMIER Latin
Dance & Music event where the world comes to dance no matter your style
and/or age, wether you're a amateur or professional...

Saturday, Jul 4

Macy's 4th of July Fireworks
Jul 4, 9:00pm at Hudson River Park at Pier..., New York

New York, NY. April 29, 2009. Macy's 4th of July Fireworks' pyrotechnic splendor
returns to the Hudson River for the first time since 2000 to celebrate America's
birthday with the nation's largest...

Identifying Events and Associated Social Media Documents

- Applications
  - Event browsing
  - Local event search
  - …
- General approach: group similar documents via clustering
  Each cluster corresponds to one event and its associated social media documents

Event Identification: Challenges

Event Identification: Challenges

- Uneven data quality
  - Missing, short, uninformative text
  - … but revealing structured context available: tags, date/time, geo-coordinates
- Scalability
- Dynamic data stream of event information
- Number of events unknown
  - Difficult to estimate
  - Constantly changing

Clustering Social Media Documents

Clustering Social Media Documents

- Social media document representation
- Social media document similarity
- Social media document clustering framework
- Similarity metric learning for clustering
  - Ensemble-based
  - Classification-based
- Evaluation results

Social Media Document Features

Social Media Document Features

Title

DSC01325

radiohead performing

Uploaded on August 9, 2008 by **said&done**

+ said&done's photostream

Tags

- apw
- All
- Points
- West

Additional Information

- Some rights reserved
- Anyone can see this photo

- Taken in **Jersey City**, New Jersey (map)
- Taken with a Sony Ericsson K810i. More properties
- Taken on August 8, 2008
- Viewed **5** times

Social Media Document Features

Title

DSC01325

radiohead performing

Uploaded on August 9, 2008
by **said&done**

+ said&done's photostream

Tags

- apw
- All
- Points
- West

Additional Information

- Some rights reserved
- Anyone can see this photo

- Taken in **Jersey City**, New Jersey (map)
- Taken with a Sony Ericsson K810i. More properties
- Taken on August 8, 2008
- Viewed **5** times

Social Media Document Features

Title

Description

Social Media Document Features

Title

Description

DSC01325

radiohead performing

Uploaded on August 9, 2008 by **said&done**

+ said&done's photostream

Tags
- apw
- All
- Points
- West

Additional Information

ⓘ Some rights reserved

🔲 Anyone can see this photo

○ Taken in **Jersey City**, New Jersey (map)
○ Taken with a Sony Ericsson K810i. More properties
○ Taken on August 8, 2008
○ Viewed **5** times

Social Media Document Features

Title

Description

Tags

Social Media Document Features

Title

Description

Tags

DSC01325

radiohead performing

Uploaded on August 9, 2008
by **said&done**

+ said&done's photostream

Tags

🌐 apw
🌐 All
🌐 Points
🌐 West

Additional Information

ⓘ Some rights reserved
🔲 Anyone can see this photo

○ Taken in **Jersey City**, New Jersey (map)
○ Taken with a Sony Ericsson K810i. More properties
○ Taken on August 8, 2008
○ Viewed **5** times

Social Media Document Features

Title

Description

Tags

Date/Time

DSC01325

radiohead performing

Uploaded on August 9, 2008 by **said&done**

+ said&done's photostream

Tags
- apw
- All
- Points
- West

Additional Information

ⓘ Some rights reserved
Anyone can see this photo

○ Taken in **Jersey City**, New Jersey (map)
○ Taken with a Sony Ericsson K810i.
More properties

Taken on August 8, 2008
Viewed 5 times

Social Media Document Features

Title

Description

Tags

Date/Time

Social Media Document Features

Title

Description

Tags

Date/Time

Location

Social Media Document Features

Social Media Document Features

Title

Description

Tags

Date/Time

Location

All-Text

DSC01325

radiohead performing

Uploaded on August 9, 2008 by **said&done**

+ said&done's photostream

Tags

- apw
- All
- Points
- West

Additional Information

- Some rights reserved
- Anyone can see this photo

- Taken in **Jersey City**, New Jersey (map)
- Taken with a Sony Ericsson K810i. More properties.
- Taken on August 8, 2008
- Viewed 5 times

Social Media Document Similarity

Title

Description

Tags

Date/Time

Location

All-Text

Social Media Document Similarity

Title

Description

Tags

Date/Time

Location

All-Text

☐ Text: cosine similarity of tf-idf vectors

(tf-idf version?; stemming?; stop-word elimination?)

A A A   B B B

Social Media Document Similarity

Title

Description

Tags

Date/Time

Location

All-Text

☐ **Text: cosine similarity of tf-idf vectors**

(tf-idf version?; stemming?; stop-word elimination?)

☐ **Time: proximity in minutes**

time

Social Media Document Similarity

Title

Description

Tags

Date/Time

Location

All-Text

☐ **Text: cosine similarity of tf-idf vectors**
(tf-idf version?; stemming?; stop-word elimination?)

A A A   B B B
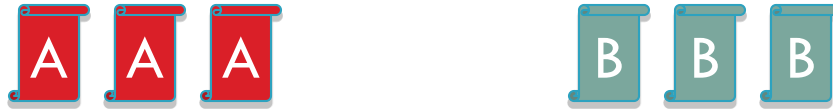
☐ **Time: proximity in minutes**

time
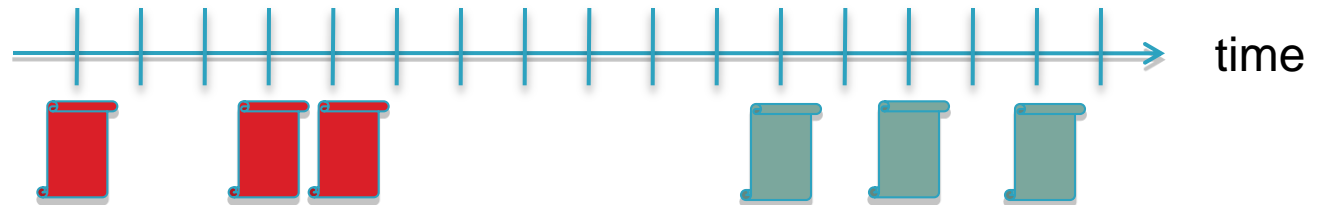
☐ **Location: geo-coordinate proximity**
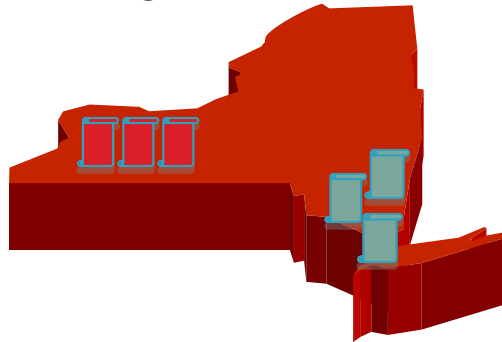
General Clustering Framework

Social media
documents

Document feature
representation

Event clusters

General Clustering Framework

Social media documents

Document feature representation

Event clusters

General Clustering Framework
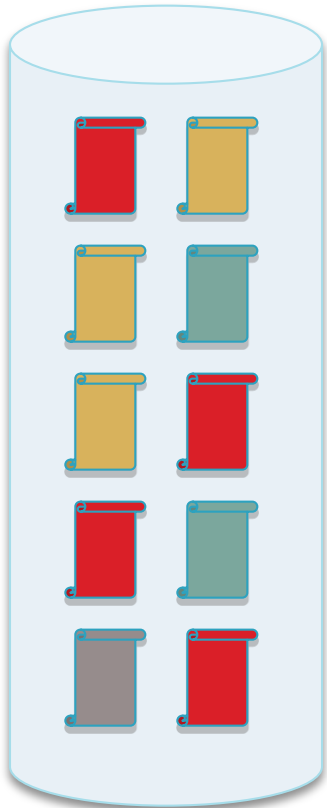
Social media documents

Document feature representation
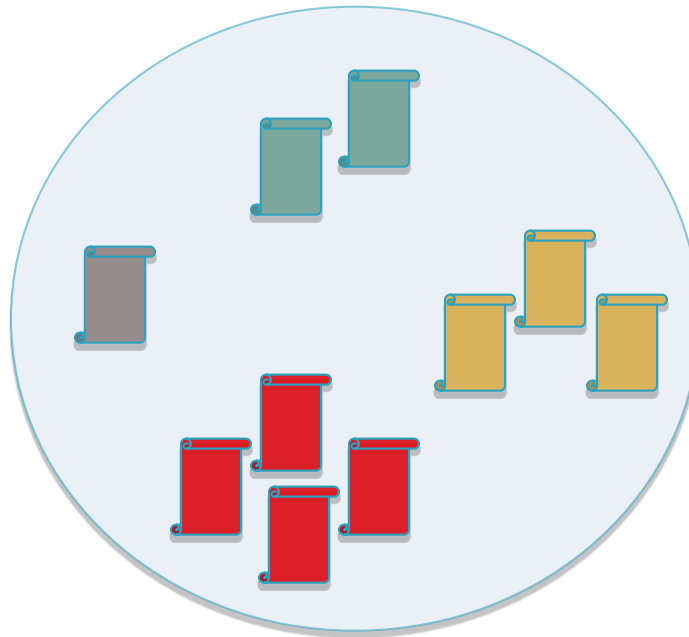
Event clusters

General Clustering Framework

Social media documents

Document feature representation

Event clusters

General Clustering Framework

Social media documents

Document feature representation

Event clusters

General Clustering Framework

Social media documents
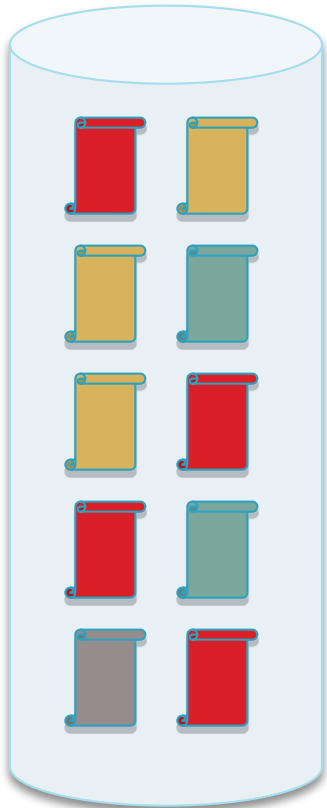
Document feature representation
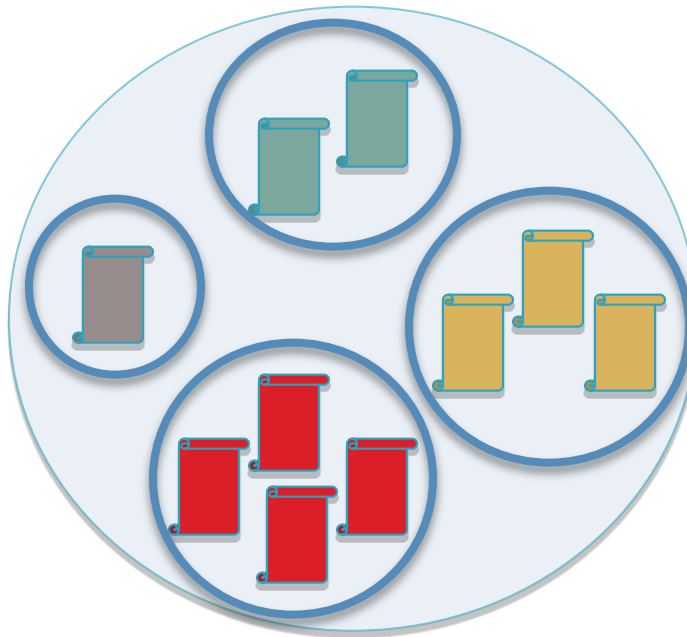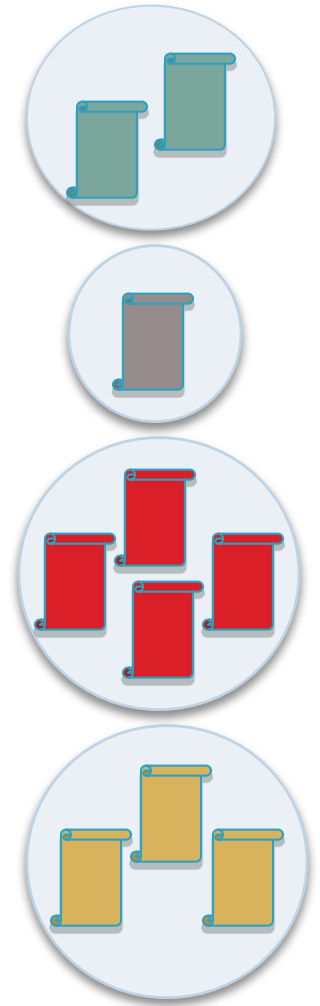
Event clusters

Clustering Algorithm

Clustering Algorithm

- Many alternatives possible! [Berkhin 2002]
- Single-pass incremental clustering algorithm
  - Scalable, online solution
  - Used effectively for event identification in textual news
  - Does not require a priori knowledge of number of clusters
  - Parameters:
    - Similarity Function σ
    - Threshold μ

Cluster Representation and Parameter Tuning

Cluster Representation and Parameter Tuning

☐ Centroid cluster representation

   ☐ Average tf-idf scores

   ☐ Average time

   ☐ Geographic mid-point

☐ Parameter tuning in supervised training phase

  Clustering quality metrics to optimize:

     ■ Normalized Mutual Information (NMI) [Amigó et al. 2008]

     ■ B-Cubed [Strehl et al. 2002]

Clustering Quality Metrics

- Characteristics of clusters:

  - Homogeneity

  - Completeness

Clustering Quality Metrics

- Characteristics of clusters:
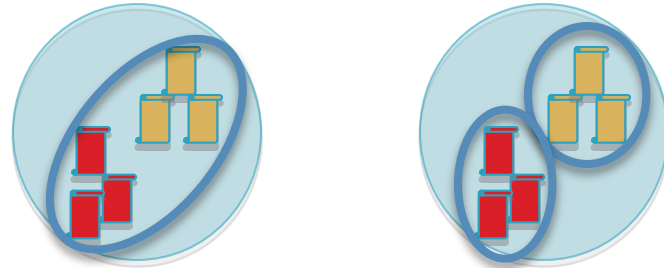
  - Homogeneity

  - Completeness

Clustering Quality Metrics

□ Characteristics of clusters:
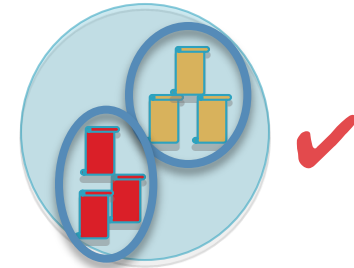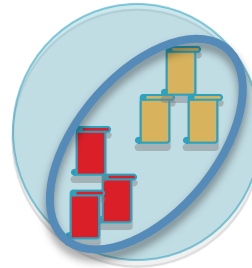
  □ Homogeneity

  □ Completeness

Clustering Quality Metrics

□ Characteristics of clusters:
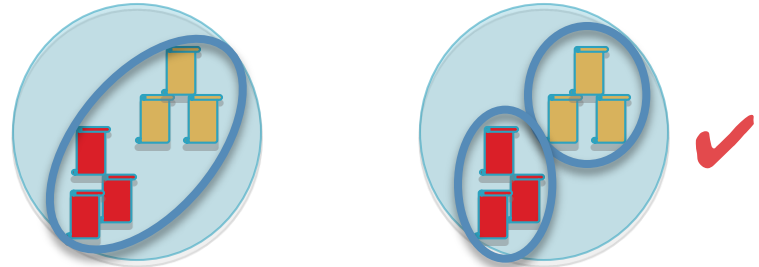
  ■ Homogeneity
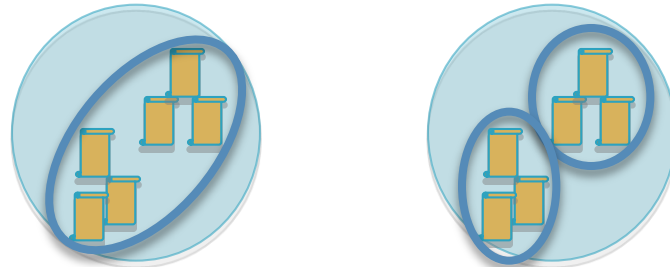
  ■ Completeness

Clustering Quality Metrics

☐ Characteristics of clusters:

☐ Homogeneity
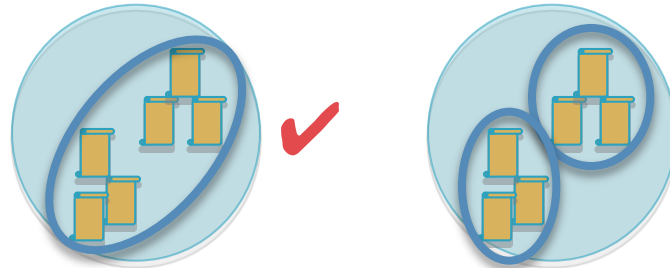
☐ Completeness

Clustering Quality Metrics

- Characteristics of clusters:

  - Homogeneity
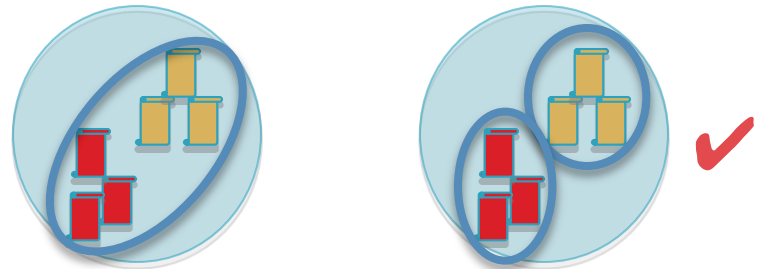
  - Completeness
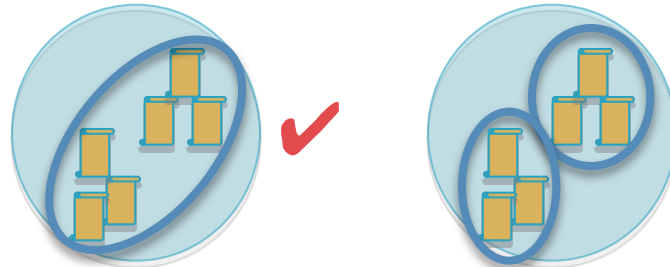
- Captured by both NMI and B-Cubed
- Optimize both metrics using a single (Pareto optimal) objective function: NMI+B-Cubed

Learning a Similarity Metric for Clustering

Learning a Similarity Metric for Clustering

- Ensemble-based similarity
  - Training a cluster ensemble
  - Computing a similarity score by:
    - Combining individual partitions
    - Combining individual similarities
- Classification-based similarity
  - Training data sampling strategies
  - Modeling strategies

Overview of a Cluster Ensemble Algorithm
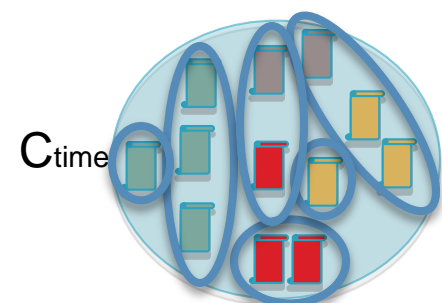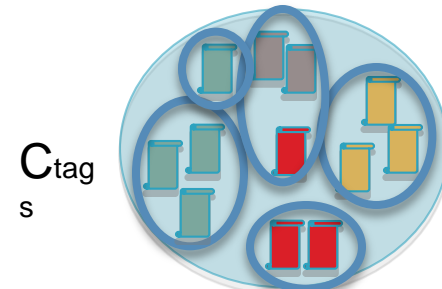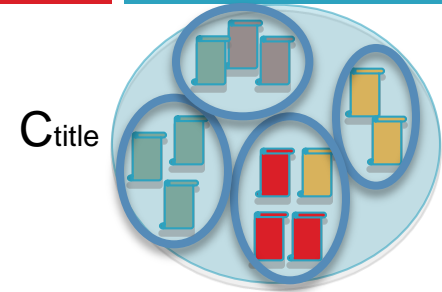
Overview of a Cluster Ensemble Algorithm

C_{title}

C_{tags}

C_{time}

Overview of a Cluster Ensemble Algorithm

C_{title}

W_{title}

Consensus Function: combine ensemble similarities

C_{tags}

W_{tags}

$f(C,W)$

C_{time}

W_{time}

Learned in a training step

Overview of a Cluster Ensemble Algorithm

Overview of a Cluster Ensemble Algorithm

C_{title}

W_{title}

C_{tags}

W_{tags}

C_{time}

W_{time}

$f(C,W)$

Overview of a Cluster Ensemble Algorithm

Overview of a Cluster Ensemble Algorithm

Overview of a Cluster Ensemble Algorithm

C_{title}

W_{title}

C_{tags}

W_{tags}

C_{time}

W_{time}

$f(C,W)$
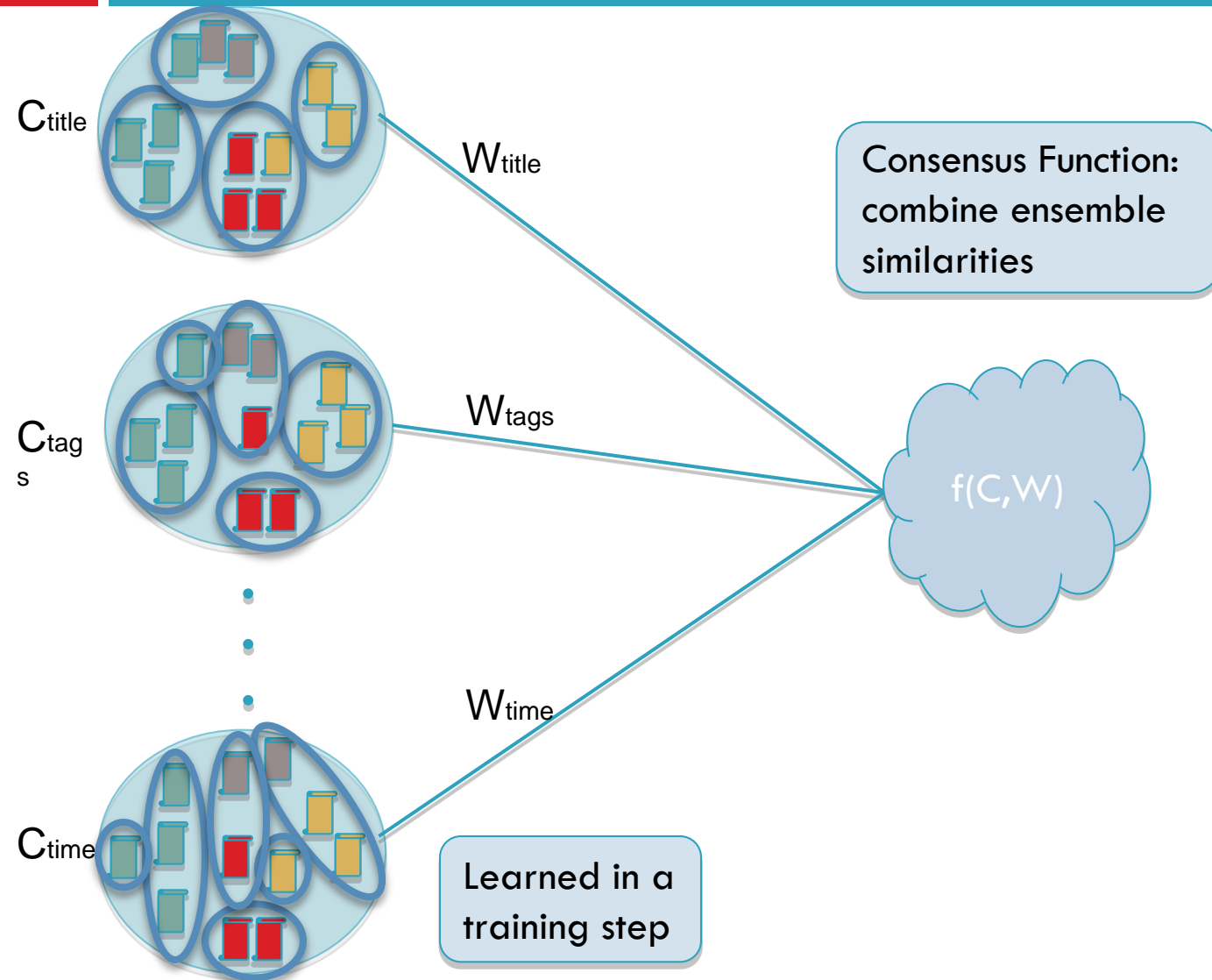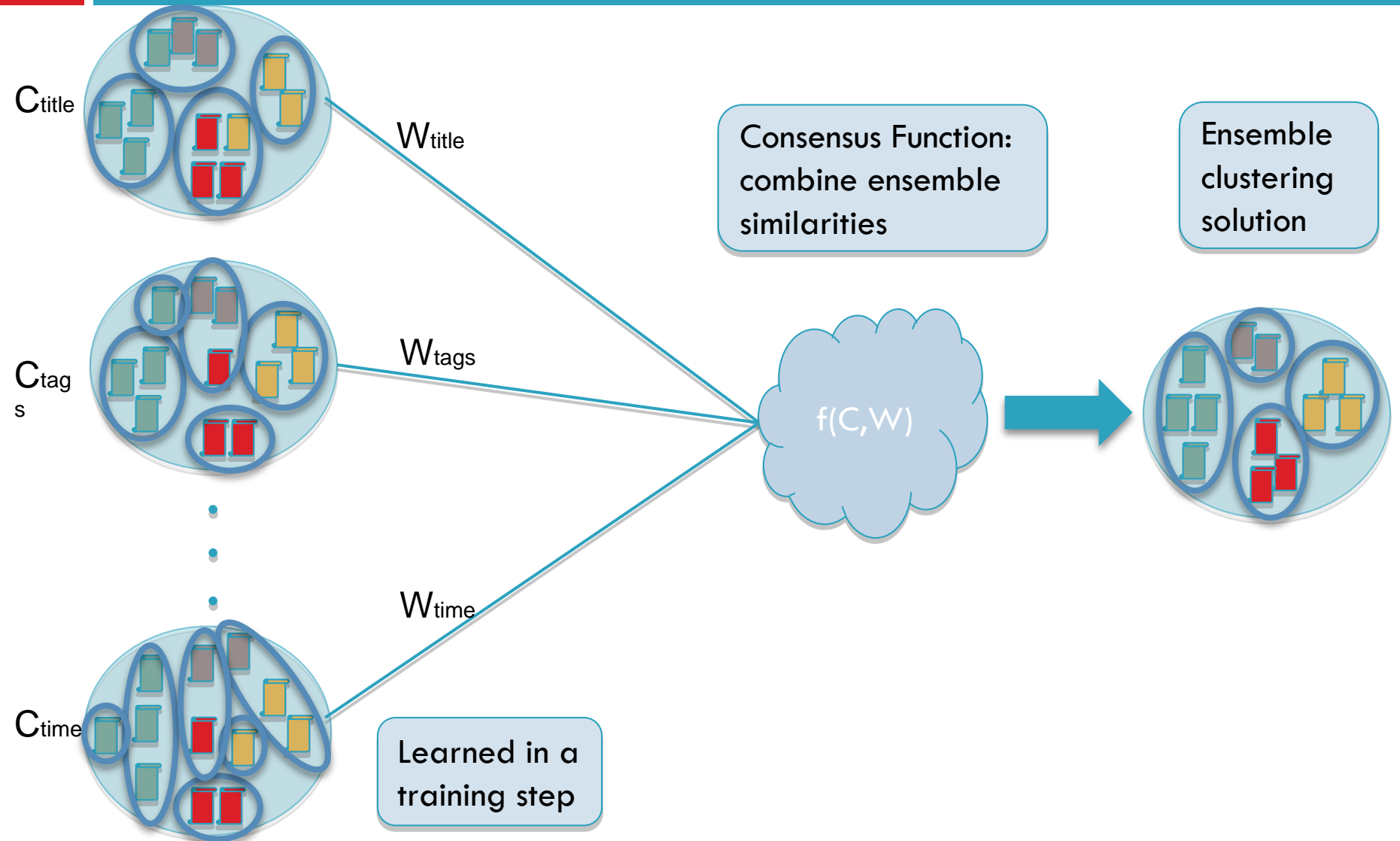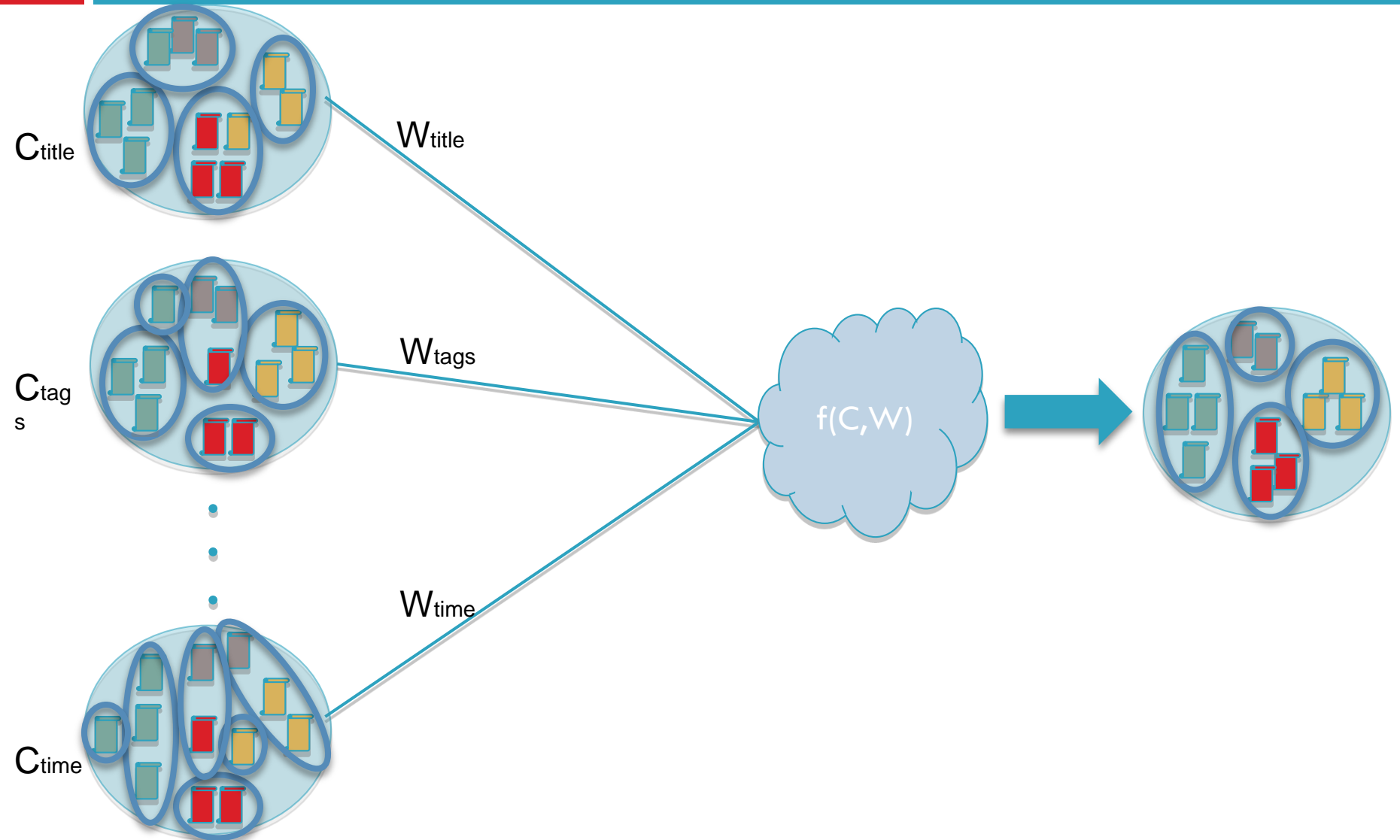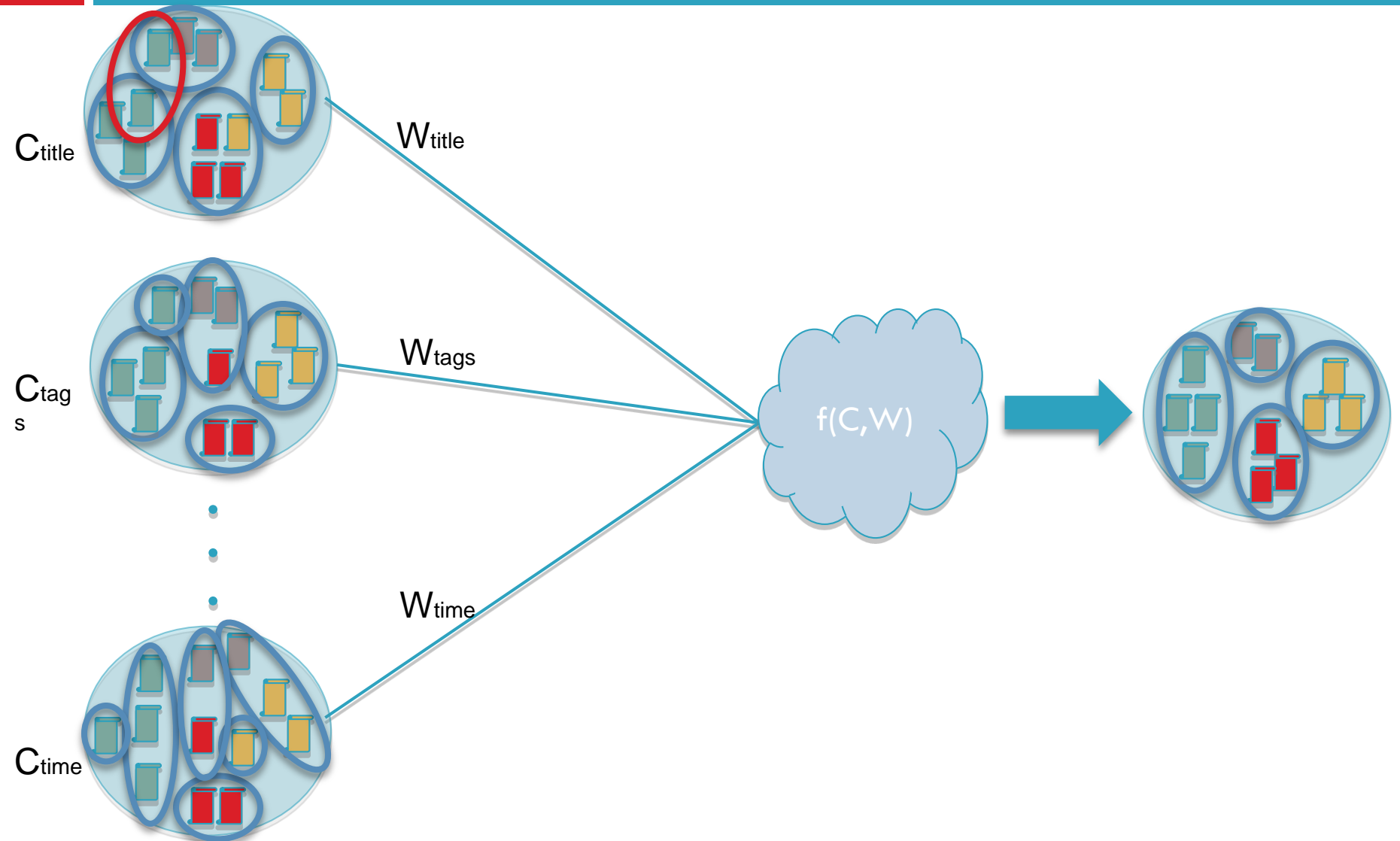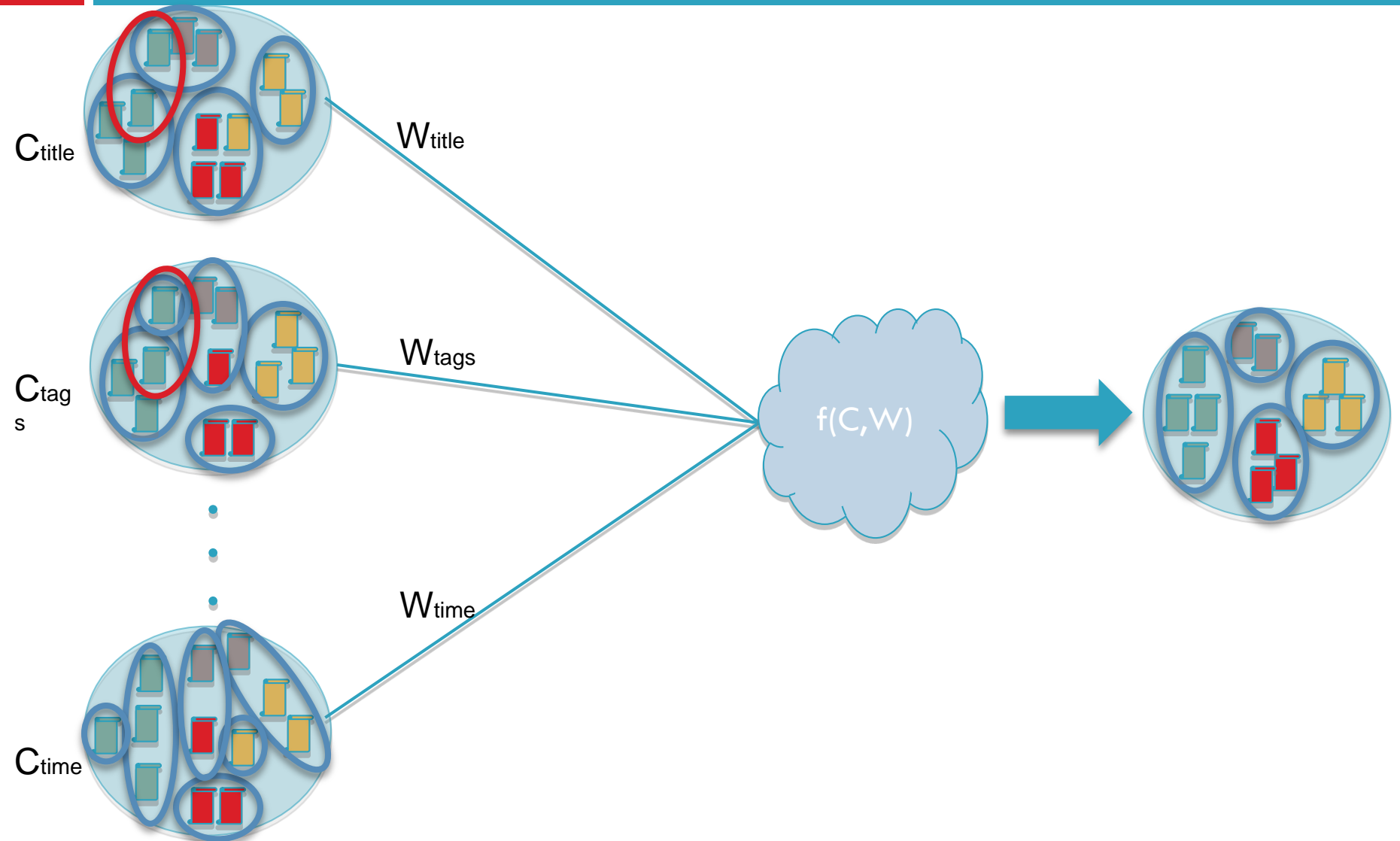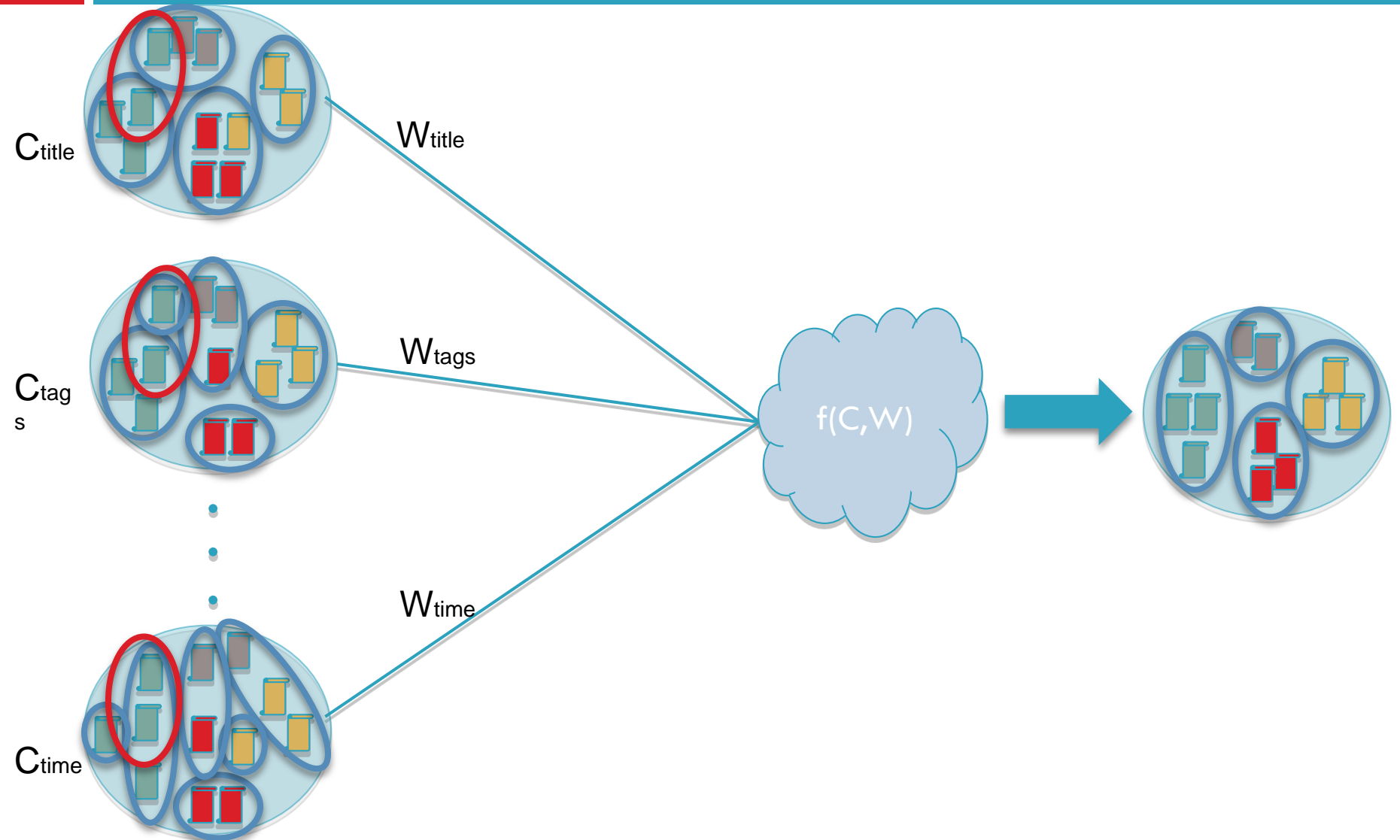
Overview of a Cluster Ensemble Algorithm
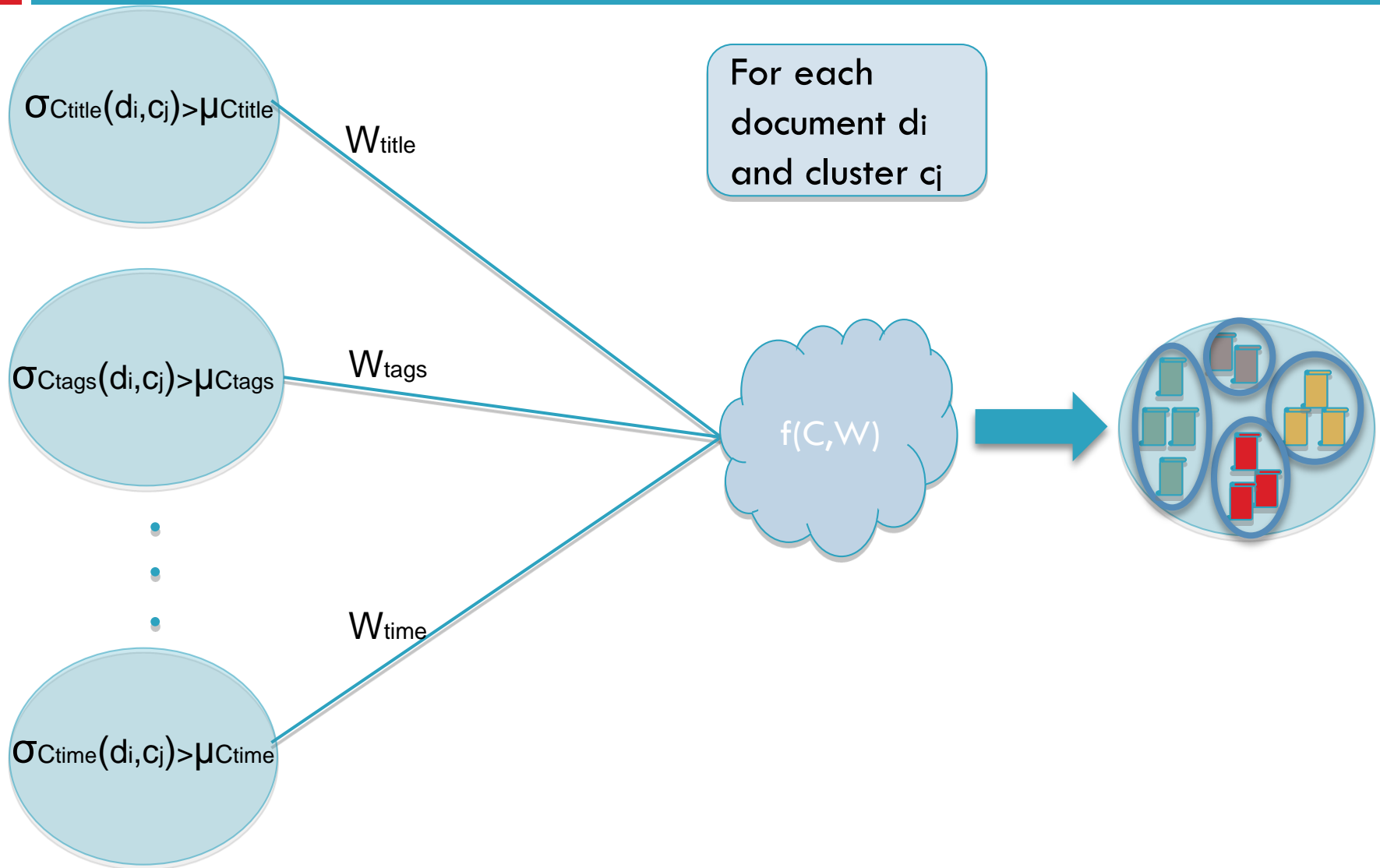
Learning a Similarity Metric for Clustering

- Ensemble-based similarity
  - Training a cluster ensemble
  - Computing a similarity score by:
    - Combining individual partitions
    - Combining individual similarities
- Classification-based similarity
  - Training data sampling strategies
  - Modeling strategies

Classification-based Similarity Metrics

Classification-based Similarity Metrics

- Classify pairs of documents as similar/dissimilar
- Feature vector
  - Pairwise similarity scores
  - One feature per similarity metric (e.g., time-proximity, location-proximity, …)
- Modeling strategies
  - Document pairs
  - Document-centroid pairs

Training Classification-based Similarity

Training Classification-based Similarity

- Challenge: most document pairs do not correspond to the same event
  - Skewed label distribution
  - Small, highly homogeneous clusters
- Sampling strategies
  - Random
    - Select a document at random
    - Randomly create one positive and one negative example
  - Time-based
    - Create examples for the first NxN documents
    - Resample such that the label distribution is balanced

Experiments: Alternative Similarity Metrics

Experiments: Alternative Similarity Metrics

- Ensemble-based techniques
  - Combining individual partitions (ENS-PART)
  - Combining individual similarities (ENS-SIM)
- Classification-based techniques
  - Modeling: document-document vs. document-centroid pairs
  - Sampling: time-based vs. random
  - Logistic Regression (CLASS-LR), Support Vector Machines (CLASS-SVM)
- Baselines
  - Title, Description, Tags, All-Text, Time-Proximity, Location-Proximity

Experimental Setup

Experimental Setup

- Datasets:
  - Upcoming
    - >270K Flickr photos
    - Event labels from the "upcoming" event database (upcoming:event=12345)
    - Split into 3 parts for training/validation/testing
  - LastFM
    - >594K Flickr photos
    - Event labels from last.fm music catalog (lastfm:event=6789)
    - Used as an additional test set

DETAILS & PHOTOS | 🗺 **MAPS**

AC DC Montreal

Saturday August 8, 2009 at 10:00pm

Olympic Stadium

4545 Pierre-De Coubertin Avenue
Montreal, Québec Get Directions

AC DC Tickets Montreal
Sat, Aug 8, 2009 10:00 PM at Olympic Stadium in Montreal, Canada

Legendary band, AC/DC, still know how to give superb performance after more than 30 years on the music scene. With Brian Johnson on lead vocals, AC/DC light up the stage with hits like "You Shook Me All Night," "Hells Bells," "Back in Black," and countless more! Be sure not to miss them when they play live at the Olympic Stadium in Montreal, Canada on August 8, 2009! Buy your AC DC tickets today before it's too late!

Click here to find best deals on AC DC tickets – Montreal now!

AC/DC is considered one of the greatest rock and roll bands in history. And now, after almost 8 years, the band is hitting the road to support their new album, Black Ice. As expected, AC/DC tickets are selling out fast! Everyone wants to see this legendary band perform live in concert! With their high energy and a reputation for putting on an exciting concert, AC/DC tickets will not last long! Act now, buy your AC/DC tickets - Montreal and see them live on stage!

Website: http://www.zimbio.com/Tickets+Online/articles/O8HwbFRwnEJ/AC...

Category: Music

▦ Add to calendar   ✉ Invite Friends   🖨 Print

Added by ▪ **dinoj54** on July 28, 2009 ⚑ **Report a problem**

Event Photos

✚ Add Photos   See all 9 photos

Been there, done that?

| I Attended | or | Interesting |

2 People saved this event. Show me ▾

Related Groups   Tags

Join a group   ✚ Add a tag?

Experimental Setup

- Datasets:
  - Upcoming
    - \>270K Flickr photos
    - Event labels from the "upcoming" event database (upcoming:event=12345)
    - Split into 3 parts for training/validation/testing
  - LastFM
    - \>594K Flickr photos
    - Event labels from last.fm music catalog (lastfm:event=6789)
    - Used as an additional test set

Event

Attendance

Videos

Images

Similar

Reviews

AUG 8 **AC/DC** PAST EVENT

The Answer

www.setlist.fm/setlist/acdc/2009/olym...

✉ Share this event   📧 Export event

34 attending | 💬 19 shouts | Added by felip_ars

Saturday 8 August 2009

Stade Olympique
4545 Rue Pierre-de-Coubertin
Montreal,qc, Montreal QC H1V 3N7
Canada
Show on Map

🚩 Flag for review

▶ Play AC/DC Radio

AC/DC

Stade Olympique
Saturday 8 August 2009

Upload poster

Did you go?

Yes ⬍   Update Calendar

34 Attended   6 Maybe

Funkafonik   felip_ars   macinbest   xizor81

mustys   thewalrusnp   Codc   -RaGe-

Soul_krusher

See more ◯

Photos

See more on Flickr (8) ◯

Tracks from the Lineup

▶ The Answer – Never Too Late full track   3:35
▶ The Answer – On And On full track   3:37
▶ The Answer – Under the Sky full track   4:06

who goes to more movies?
source: nationmaster.com

▼ 🇹🇷 Turkey   Germany ▼

86,301
people per day

408,202
people per day

save on movie rentals ❯   more people go with Visa.   **VISA**

Reviews

Setlist AC/DC at Olympic Stadium, Montreal, Canada
by setlistfm | 10 Aug 2009 | 💬 1 comment

see more ◯

📝 Review this event

Related Events

OCT 3 AC/DC
MGM Grand Hotel, Las Vegas, United States
💬 Add shout | 4 attendees

DEC 6 The Answer
Portsmouth Pyramids, Portsmouth, United Kingdom
💬 Add shout | 5 attendees

see more ◯

Experimental Setup

- Datasets:
  - Upcoming
    - >270K Flickr photos
    - Event labels from the "upcoming" event database (upcoming:event=12345)
    - Split into 3 parts for training/validation/testing
  - LastFM
    - >594K Flickr photos
    - Event labels from last.fm music catalog (lastfm:event=6789)
    - Used as an additional test set

Clustering Accuracy over Upcoming Test Set

Algorithm	NMI	B-Cubed
All-Text	0.9240	0.7697
Tags	0.9229	0.7676
ENS-PART	0.9296	0.7819
ENS-SIM	0.9322	0.7861
CLASS-SVM	0.9425	0.8095
CLASS-LR	**0.9444**	**0.8155**

- All similarity learning techniques outperform the baselines
- Classification-based techniques perform better than ensemble-based techniques

Clustering Accuracy over Upcoming Test Set

Algorithm	NMI	B-Cubed
All-Text	0.9240	0.7697
Tags	0.9229	0.7676
ENS-PART	0.9296	0.7819
ENS-SIM	0.9322	0.7861
CLASS-SVM	0.9425	0.8095
CLASS-LR	**0.9444**	**0.8155**

- All similarity learning techniques outperform the baselines
- Classification-based techniques perform better than ensemble-based techniques

Clustering Accuracy over Upcoming Test Set

Algorithm	NMI	B-Cubed
All-Text	0.9240	0.7697
Tags	0.9229	0.7676
ENS-PART	0.9296	0.7819
ENS-SIM	0.9322	0.7861
CLASS-SVM	0.9425	0.8095
CLASS-LR	**0.9444**	**0.8155**

☐ All similarity learning techniques outperform the baselines

☐ Classification-based techniques perform better than ensemble-based techniques

Clustering Accuracy over Upcoming Test Set

Algorithm	NMI	B-Cubed
All-Text	0.9240	0.7697
Tags	0.9229	0.7676
ENS-PART	0.9296	0.7819
ENS-SIM	0.9322	0.7861
CLASS-SVM	0.9425	0.8095
CLASS-LR	**0.9444**	**0.8155**

- All similarity learning techniques outperform the baselines
- Classification-based techniques perform better than ensemble-based techniques
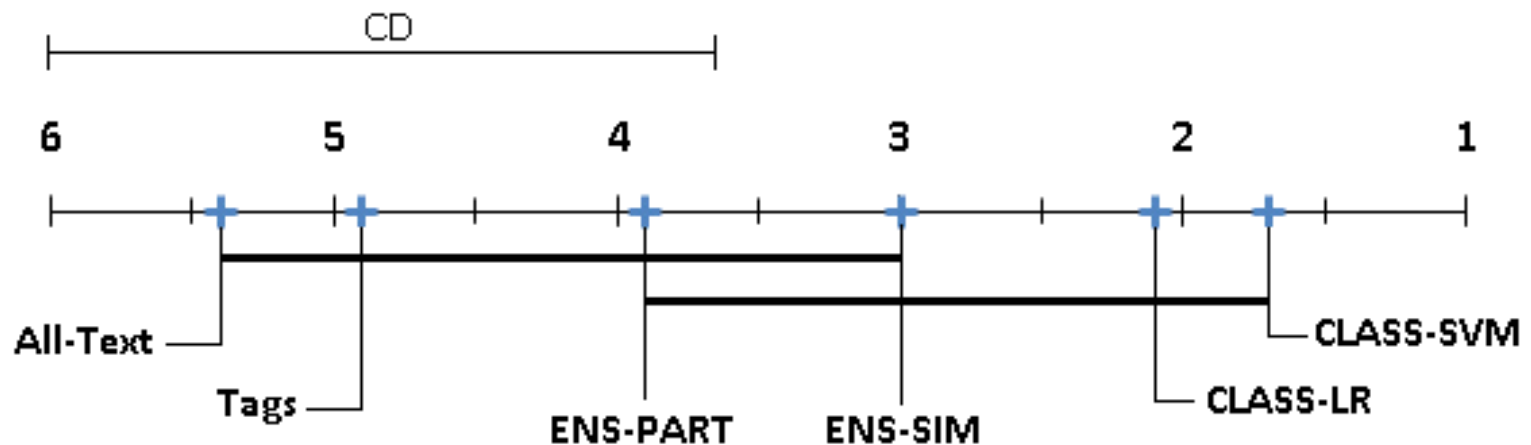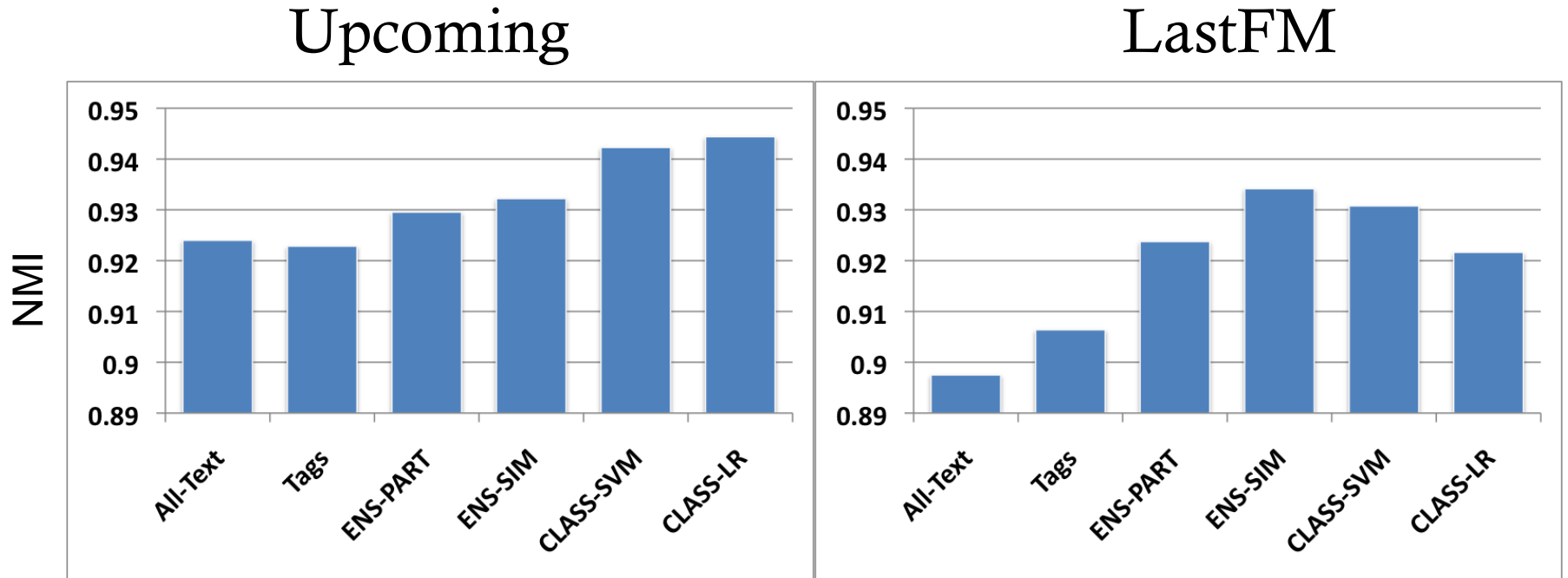
Statistical Significance Analysis

- Clustering results for 10 partitions of Upcoming test set
- Significant using Friedman test, $p < 0.05$
- Post-hoc analysis:

NMI: Clustering Accuracy over Both Test Sets

Upcoming

LastFM

- Similarity learning models trained on Upcoming data show similar trends when tested on LastFM data

Conclusions

Conclusions

- Structured context features of social media documents
  - Effective complementary cues for social media document similarity
  - Tags, Time-Proximity among highest weighted features
- Domain-appropriate similarity metrics
  - Weighted combination yields high quality clustering results
  - Significantly outperform text-only techniques
- Similarity learning models generalize to unseen data sets

Current and Future Work

- Improving clustering accuracy with social media "links" [SSM '10 poster]
- Capturing event content across sites (YouTube, Flickr, Twitter)
- Designing event search strategies

Thank You!