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” = 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
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.

**Popular Events**

**All Points West Music Festival**
Friday July 31, 2009 - Sunday August 2, 2009
Liberty State Park, 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.

**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!

**Recommended Events**

**Saturday, Aug 15**
Universal Hip Hop Parade
Aug 15, 3:00pm at Marcus Garvey Boulevard, Brooklyn, NY
Come join the world's longest celebration of Hip Hop Culture. The 10th Annual Universal Hip Hop Parade steps off in the Red-Stay 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 Javits Convention Center, 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, NY
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 whose 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 multi-day upscale event presenting international couture and luxury fashion shows. It is...

**Wednesday, Sep 2**
The 9th Annual New York International Salsa Congress - Dance & Music Festival
Sep 2, 9:00pm 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, whether you're an amateur or professional...

**Saturday, Jul 4**
Macy's 4th of July Fireworks
Jul 4, 9:00pm at Hudson River Park Pier 21, New York, NY
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

DSC01325

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

radiohead performing
Social Media Document Features

DSC01325

radiohead performing

Additional Information
- Some rights reserved
- Anyone can see this photo
- Taken in Jersey City, New Jersey (map)
- Taken with a Sony Ericsson K810i
- Taken on August 8, 2008
- Viewed 5 times
Social Media Document Features

DSC01325

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

radiohead performing
Social Media Document Features

Title:

DSC01325

Description:

radiohead performing
Social Media Document Features

DSC01325

radiohead performing
Social Media Document Features

**Title**
DSC01325

**Description**
radiohead performing

**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.
- Taken on August 8, 2008
- Viewed 5 times
Social Media Document Features

Title

Description

Tags

DSC01325

radiohead performing

Additional Information
- Some rights reserved
- Anyone can see this photo
- Taken in Jersey City, New Jersey (map)
- Taken with a Sony Ericsson K810i
- Taken on August 8, 2008
- Viewed 5 times
Social Media Document Features

Title: DSC01325

Description: radiohead performing

Date/Time: Taken on August 8, 2008

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

Uploaded on August 9, 2008 by said&done
Social Media Document Features

- Title
- Description
- Tags
- Date/Time
- Location

**DSC01325**

Uploaded on August 9, 2008 by said&done

- 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 9, 2008
- Viewed 5 times

radiohead performing
Social Media Document Features

Title: DSC01325

Description: radiohead performing

Date/Time: August 9, 2008

Location: Jersey City, New Jersey

Tags: apw, All, Points, West

Additional Information:
- Some rights reserved
- Anyone can see this photo
- Taken in Jersey City, New Jersey
- Taken with a Sony Ericsson K810i
- More properties
- Taken on August 8, 2008
- Viewed 5 times
Social Media Document Features

- **Title**: DSC01325
- **Description**: radiohead performing
- **Date/Time**: Taken on August 8, 2008
- **Location**: Taken in Jersey City, New Jersey
- **Tags**: apw, All, Points, West
- **Additional Information**: Some rights reserved, Anyone can see this photo
Social Media Document Similarity
Social Media Document Similarity

- **Text:** cosine similarity of tf-idf vectors
  (tf-idf version?; stemming?; stop-word elimination?)
Social Media Document Similarity

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

- **Time:** proximity in minutes
Social Media Document Similarity

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

- **Time**: proximity in minutes

- **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
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 $\sigma$
- Threshold $\mu$
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
Overview of a Cluster Ensemble Algorithm

Consensus Function: combine ensemble similarities

Learned in a training step
Overview of a Cluster Ensemble Algorithm

Consensus Function: combine ensemble similarities

Ensemble clustering solution

Learned in a training step
Overview of a Cluster Ensemble Algorithm

\[ f(C, W) \]
Overview of a Cluster Ensemble Algorithm
Overview of a Cluster Ensemble Algorithm

\[ f(C, W) \]
Overview of a Cluster Ensemble Algorithm
Overview of a Cluster Ensemble Algorithm

For each document $d_i$ and cluster $c_j$:

- $\sigma_{C_{\text{title}}}(d_i, c_j) > \mu_{C_{\text{title}}}$
- $\sigma_{C_{\text{tags}}}(d_i, c_j) > \mu_{C_{\text{tags}}}$
- $\sigma_{C_{\text{time}}}(d_i, c_j) > \mu_{C_{\text{time}}}$

$W_{\text{title}}$, $W_{\text{tags}}$, $W_{\text{time}}$
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
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
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
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
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

- All similarity learning techniques outperform the baselines
- Classification-based techniques perform better than ensemble-based techniques
Clustering Accuracy over Upcoming Test Set

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>NMI</th>
<th>B-Cubed</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Text</td>
<td>0.9240</td>
<td>0.7697</td>
</tr>
<tr>
<td>Tags</td>
<td>0.9229</td>
<td>0.7676</td>
</tr>
<tr>
<td>ENS-PART</td>
<td>0.9296</td>
<td>0.7819</td>
</tr>
<tr>
<td>ENS-SIM</td>
<td>0.9322</td>
<td>0.7861</td>
</tr>
<tr>
<td>CLASS-SVM</td>
<td>0.9425</td>
<td>0.8095</td>
</tr>
<tr>
<td>CLASS-LR</td>
<td>0.9444</td>
<td>0.8155</td>
</tr>
</tbody>
</table>
Clustering Accuracy over Upcoming Test Set

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>NMI</th>
<th>B-Cubed</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Text</td>
<td>0.9240</td>
<td>0.7697</td>
</tr>
<tr>
<td>Tags</td>
<td>0.9229</td>
<td>0.7676</td>
</tr>
<tr>
<td>ENS-PART</td>
<td>0.9296</td>
<td>0.7819</td>
</tr>
<tr>
<td>ENS-SIM</td>
<td>0.9322</td>
<td>0.7861</td>
</tr>
<tr>
<td>CLASS-SVM</td>
<td>0.9425</td>
<td>0.8095</td>
</tr>
<tr>
<td>CLASS-LR</td>
<td>0.9444</td>
<td>0.8155</td>
</tr>
</tbody>
</table>

- All similarity learning techniques outperform the baselines
- Classification-based techniques perform better than ensemble-based techniques
Clustering Accuracy over Upcoming Test Set

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>NMI</th>
<th>B-Cubed</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Text</td>
<td>0.9240</td>
<td>0.7697</td>
</tr>
<tr>
<td>Tags</td>
<td>0.9229</td>
<td>0.7676</td>
</tr>
<tr>
<td>ENS-PART</td>
<td>0.9296</td>
<td>0.7819</td>
</tr>
<tr>
<td>ENS-SIM</td>
<td>0.9322</td>
<td>0.7861</td>
</tr>
<tr>
<td>CLASS-SVM</td>
<td>0.9425</td>
<td>0.8095</td>
</tr>
<tr>
<td>CLASS-LR</td>
<td><strong>0.9444</strong></td>
<td><strong>0.8155</strong></td>
</tr>
</tbody>
</table>

- 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:
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!