Geo-Temporal Mining and Searching of Events from Web-based Image Collections

Massimiliano Ruocco

Department of Computer and Information Science, Norwegian University of Science and Technology

May 22, 2014
Outline

1. Introduction and Motivation
2. Research Questions
3. Part I - Mining and Detection of Events in Media Sharing Applications
4. Part II - Exploring Geo-Temporal Distribution of Tags in Social Media
5. Conclusion and Future Work
# Outline

1. **Introduction and Motivation**

2. **Research Questions**

3. **Part I - Mining and Detection of Events in Media Sharing Applications**

4. **Part II - Exploring Geo-Temporal Distribution of Tags in Social Media**

5. **Conclusion and Future Work**
Introduction

Events as unit of daily life of a person: wedding, concert, conference.

Event documented by different media, more than one person.

Event-centric knowledge: lastfm¹, Lanyrd², Eventful³

¹http://www.last.fm/
²http://www.lanyrd.com/
³http://www.eventful.com/
Introduction

Events as **unit** of **daily life** of a **person**: *wedding, concert, conference*

Introduction

- Events as unit of daily life of a person: wedding, concert, conference
- Event documented by different media, more than one person

1 http://www.last.fm/
2 http://www.lanyrd.com/
3 http://www.eventful.com/
Introduction

- Events as **unit** of **daily life** of a **person**: *wedding, concert, conference*
- **Event** documented by **different media**, **more than one** person
- **Event-centric** knowledge: *lastfm*¹, *Lanyrd*², *Eventful*³

¹http://www.last.fm/
²http://www.lanyrd.com/
³http://www.eventful.com/
Introduction

Lack of coverage
Introduction

Lack of coverage

EVENTS

PICTURES
Introduction

PICTURES

Lack of coverage

EVENTS
Flickr and LastFM

Coldplay at Emirates Stadium

With Marina & the Diamonds and Ash

Emirates Stadium
Drayton Park
London N5 1BU
United Kingdom

Show on Map
Web: www.arsenal.com

www.arsenal.com/news/news-archive/emirates-stadium-to-host-coldplay...

Arsenal Football Club has announced that world-famous rock band Coldplay will be performing two dates at Emirates Stadium next summer.

British band Coldplay, who have just released their new album ‘Mylo Xyloto’, will appear at the Gunners’ home on Friday 1st June 2012 and Saturday 2nd June 2012.

Emirates Stadium has previously played host to Bruce Springsteen in 2008 and the Summertime Ball in 2009. Arsenal Chief Executive Ivan Gazidis said: “We are absolutely delighted that Coldplay have chosen to perform at Emirates Stadium next summer. We are all very excited at the prospect of Coldplay performing to a packed Emirates Stadium.

Event added by Lylas1 | Flag for review

Photos

Recent Activity

jojojo10 went to Coldplay at Emirates Stadium. 16 days ago
silverreind went to Coldplay at Emirates Stadium. October 2013
AlexBalchin went to Coldplay at Emirates Stadium. August 2013
ralph went to Coldplay at Emirates Stadium. September 2013
Flickr and LastFM

- Flickr: 6 Billion pictures (August 2011)
- Flickr/lastfm: 560K pictures (May 2014)
Events in Media Sharing Applications - Research Tasks
Events in Media Sharing Applications - Task
Focus of the Research Work

Event-related detection and search of images over media-sharing applications
# Outline

1. Introduction and Motivation

2. Research Questions

3. Part I - Mining and Detection of Events in Media Sharing Applications

4. Part II - Exploring Geo-Temporal Distribution of Tags in Social Media

5. Conclusion and Future Work
[RQ] - Can a raw geographical and temporal information, surrounding annotated resources, improve browsing, detection and text-based searching of events?
Research Questions

[RQ] - Can a raw geographical and temporal information, surrounding annotated resources, improve browsing, detection and text-based searching of events?

[RQ1] Can raw geographical and temporal information surrounding resources improve the clustering and detection of resources geotagged and timestamped in a dataset?

[RQ2] Can a raw geographical and temporal term profile improve existing retrieval and query expansion models for the task of tag-based search of event-related resources?

[RQ3] How can raw geographical and temporal data be modelled and cooperate to improve and extend the retrieval performance of a tag-based search of event-related resources?

[RQ4] Can a geographical and temporal tag profile improve tag co-occurrence and tag similarity?

[RQ5] How can a geographical and temporal term profile be used to infer semantics to the tags?
Contributions
Contributions

[C1] Detection of Event-related Resources

[C1.1] Improving Clustering and Extraction of Event-Related Pictures by considering geographical and temporal dimension

[C1.2] Improving Effectiveness of Detection and Grouping of Event-related pictures by using external knowledge sources
Contributions

- **[C1]** Detection of Event-related Resources
  - **[C1.1]** Improving Clustering and Extraction of Event-Related Pictures by considering geographical and temporal dimension
  - **[C1.2]** Improving Effectiveness of Detection and Grouping of Event-related pictures by using external knowledge sources

- **[C2]** Spatiotemporal Analysis of Tag Point Pattern
  - **[C2.1]** Definition of novel features from Geographical Distribution of tags for extraction of locational tags
  - **[C2.2]** Definition of novel tag relatedness similarity measure improving retrieval effectiveness in QE framework
Contributions

[C1] Detection of Event-related Resources
  - [C1.1] *Improving Clustering and Extraction of Event-Related Pictures by considering geographical and temporal dimension*
  - [C1.2] *Improving Effectiveness of Detection and Grouping of Event-related pictures by using external knowledge sources*

[C2] Spatiotemporal Analysis of Tag Point Pattern
  - [C2.1] *Definition of novel features from Geographical Distribution of tags for extraction of locational tags*
  - [C2.2] *Definition of novel tag relatedness similarity measure improving retrieval effectiveness in QE framework*

[C3] Tag-based Search of Event-Related Pictures
  - [C3.1] *Analyzing effectiveness of textual metadata for retrieval purpose*
  - [C3.2] *Improving tag-based search of event-related images of timestamped queries*
  - [C3.3] *Improving tag-based search of event-related images of not timestamped queries*
Outline

1. Introduction and Motivation
2. Research Questions
3. Part I - Mining and Detection of Events in Media Sharing Applications
4. Part II - Exploring Geo-Temporal Distribution of Tags in Social Media
5. Conclusion and Future Work
Part I: Overview

- **Novel** Definition of **Event** for Media-Sharing Applications
Part I: Overview

- **Novel** Definition of **Event** for Media-Sharing Applications

- [C1.1] *Improving Clustering and Extraction of Event-Related Pictures by considering geographical and temporal dimension*
  - Novel algorithm for extraction and group of event-related images leveraging on temporal and geographical information
Part I: Overview

- **Novel** Definition of *Event* for Media-Sharing Applications

- **[C1.1]** *Improving Clustering and Extraction of Event-Related Pictures by considering geographical and temporal dimension*
  - Novel algorithm for extraction and group of event-related images leveraging on temporal and geographical information

- **[C1.2]** *Improving Effectiveness of Detection and Grouping of Event-related pictures by using external knowledge sources*
  - Novel algorithm for extraction and group event-related images leveraging on external knowledge
Definition of Event

*Something happening in a certain place at a certain time* [Cite Yang and Carbonell]
Definition of Event

Definition of Event in Media Sharing Apps

*Something happening in a certain place at a certain time annotated with a certain tag*
Event Mining of Social Media Pictures by Suffix Tree Structure

Definition of Event

**Definition of Event in Media Sharing Apps**

 Something happening in a certain place at a certain time annotated with a certain tag

**Definition of Event Cluster \( \mathcal{E} (1) \)**

\( \mathcal{E} : \text{event cluster} \Rightarrow \{ \forall (i, j \in \mathcal{E}), G_i = G_j \land V_i = V_j \land t_i = t_j \} \)
Definition of Event

Definition of **Event in Media Sharing Apps**

*Something happening in a certain place at a certain time* annotated with a certain tag

Definition of **Event Cluster** $\mathcal{E}$ (1)

$\mathcal{E} : event\ cluster \Rightarrow \{ \forall (I_i, I_j \in \mathcal{E}), G_i = G_j \land V_i = V_j \land t_i = t_j \}$
Definition of Event

**Definition of Event in Media Sharing Apps**

*Something happening in a certain place at a certain time annotated with a certain tag*

**Definition of Event Cluster $\mathcal{E}$ (1)**

$\mathcal{E} : event\ cluster \Rightarrow \{\forall (I_i, I_j \in \mathcal{E}), G_i = G_j \land V_i = V_j \land t_i = t_j\}$

$\mathcal{E} : event\ cluster \nRightarrow \{\forall (I_i, I_j \in \mathcal{E}), G_i = G_j \land V_i = V_j \land t_i = t_j\}$
**Definition of Event in Media Sharing Apps**

*Something happening in a certain place at a certain time annotated with a certain tag*

**Definition of Event Cluster $\mathcal{E}$ (2)**

$\mathcal{E} : \text{event cluster} \iff \{ \exists (V, G, t) : S_{VGt} = S_{Gt} \}$

*Figure: Example of event cluster*
Proposed Approach - Overview

- **Main Objective:** Extract a group of images related to an event from a collection of pictures
Main Objective: Extract of group of images related to an event from a collection of pictures

Figure: Overview of the system
Event Mining of Social Media Pictures by Suffix Tree Structure

Proposed Approach - Core

- Based on suffix tree data structure
Proposed Approach - Core

- Based on **suffix tree** data structure
- **Suffix Tree Clustering (STC)** algorithm
Event Mining of Social Media Pictures by Suffix Tree Structure

**Proposed Approach - Core**

- Based on **suffix tree** data structure
- **Suffix Tree Clustering (STC)** algorithm
- Proposed **Tag Extension**:

**Figure**: Example of annotation expansion. The original tag set $T$ is shown on the left side. After the extension with temporal-string and location-string, the set becomes $T''$. 
Proposed Approach - Detection of event clusters

- detection of event cluster according to the proposed Event Cluster Definition
Proposed Approach - Detection of event clusters

- detection of event cluster according to the proposed Event Cluster Definition

\[
\frac{|S_{VGt} \cap S_{Gt}|}{\max(S_{VGt}, S_{Gt})} \geq K,
\] (1)
Proposed Approach - Detection of event clusters

- **detection of event cluster** according to the proposed Event Cluster Definition

\[
\frac{|S_{VGt} \cap S_{Gt}|}{\max(S_{VGt}, S_{Gt})} \geq K, \quad (1)
\]

- **Semantic** meaning of the branches
Proposed Approach - Detection of event clusters

- **detection of event cluster** according to the proposed **Event Cluster Definition**

\[
\frac{|S_{VGt} \cap S_{Gt}|}{\max(S_{VGt}, S_{Gt})} \geq K, \tag{1}
\]

- **Semantic** meaning of the branches
Proposed Approach - Detection of event clusters

- **detection** of event cluster according to the proposed Event Cluster Definition
  \[
  \frac{|S_{VGt} \cap S_{Gt}|}{\max(S_{VGt}, S_{Gt})} \geq K,
  \]

- **Semantic** meaning of the branches
- **Merging** candidate event clusters
  - Semantic similarity: merging overlapping candidate event clusters.
  - Refinement step exploring spatiotemporal neighbourhood
### Dataset

**San Francisco Dataset**
- Taken around San Francisco
- $\sim 230K$ pictures
- Geotagged and timestamped

**Barcelona Dataset**
- Taken in Barcelona
- $\sim 200K$ pictures
- Geotagged and timestamped
Event Mining of Social Media Pictures by Suffix Tree Structure

Experimental Setup

- Clustering performance by IR metrics:
  - Precision at rank $k$: $P@k = \frac{|C_k \cap R_k|}{k}$
  - Tuning of $K$ in event cluster hypothesis $\frac{|S_{VGt} \cap S_{Gt}|}{\max(S_{VGt}, S_{Gt})} \geq K$
- Different granularity in time and space
Event Mining of Social Media Pictures by Suffix Tree Structure

Results and Discussions

1. soccer, race, duboc guerrero, goal, runner, athlete, mile marker, kfc height

2. bicycle, cyclocross, race, basp, ox, hunter

3. decision, gilt, onestrongonefight, marriage

4. ubergizmo, effortlessly card, tech,

5. drk, birthday, markhopkin, topofthemark

6. science, camera digital xsi

7. fun, indie mart

8. zoology, wildlife, meggie

9. www, media prom, media, prom
Results and Discussions

- Better performance for relaxed value of $K$

**Figure:** $P@50$ at different value of $K$ over different granularity in space and time (e.g., $D100$ means day as temporal granularity and 100 meters as spatial granularity)
Results and Discussions

- Better performance for relaxed value of $K$

**Figure:** $P@50$ at different value of $K$ over different granularity in space and time (e.g., $D100$ means day as temporal granularity and 100 meters as spatial granularity)

- Many candidate event clusters tend to **merge with stronger** candidate event clusters
Results and Discussions

- Better performance for relaxed value of $K$

![Bar Chart](chart.png)

**Figure:** $P@50$ at different value of $K$ over different granularity in space and time (e.g., $D100$ means day as temporal granularity and 100 meters as spatial granularity)

- Many candidate event clusters tend to **merge** with **stronger** candidate event clusters
- **Collective** events (e.g., protest, conference, parade) **pushed up** in the ranked list of clusters
Summary: Contributions vs Research Questions

- Novel algorithm for extraction of group of images related to an event from a collection of pictures
- Automatic annotation of event clusters
- Incremental Clustering algorithm
- Event clustering on multi-person set of photos
- Use of Geo-Temporal information and text annotations
- Analysis of effectiveness over different granularity in time and space
- Effectiveness compared with similar approach

[RQ1] Can raw geographical and temporal information surrounding resources improve the clustering and detection of resources geotagged and timestamped in a dataset?
NEXT QUESTION: Can external knowledge improve the detection and clustering of event-related pictures?
Introduction

MediaEval challenge: *Social Event Detection Task* (SED)
Introduction

- **MediaEval** challenge: *Social Event Detection Task* (SED)
- **Objective**: propose system able to detect/retrieve and group event-related images
Semantic Social Event Detection

Introduction

- **MediaEval** challenge: *Social Event Detection Task (SED)*
- **Objective**: propose system able to detect/retrieve and group event-related images
- **Challenges**:
  - **Challenge 1**: *Find all soccer events taking place in Barcelona (Spain) and Rome (Italy) in the test collection. For each event provide all photos associated with it*
  - **Challenge 2**: *Find all events that took place in May 2009 in the venue named Paradiso (in Amsterdam, NL) and in the Parc del Forum (in Barcelona, Spain). For each event provide all photos associated with it.*
Introduction

- **MediaEval** challenge: *Social Event Detection Task* (SED)
- **Objective**: propose system able to detect/retrieve and group event-related images
- **Challenges**:
  - **Challenge 1**: Find all soccer events taking place in Barcelona (Spain) and Rome (Italy) in the test collection. For each event provide all photos associated with it
  - **Challenge 2**: Find all events that took place in May 2009 in the venue named Paradiso (in Amsterdam, NL) and in the Parc del Forum (in Barcelona, Spain). For each event provide all photos associated with it.

- **Dataset**:
  - 73K pictures gathered from Flickr.
Introduction

- **MediaEval** challenge: *Social Event Detection Task (SED)*

- **Objective**: propose system able to detect/retrieve and group event-related images

- **Challenges**:
  - **Challenge 1**: *Find all soccer events taking place in Barcelona (Spain) and Rome (Italy) in the test collection. For each event provide all photos associated with it*
  - **Challenge 2**: *Find all events that took place in May 2009 in the venue named Paradiso (in Amsterdam, NL) and in the Parc del Forum (in Barcelona, Spain). For each event provide all photos associated with it.*

- **Dataset**:
  - 73K pictures gathered from Flickr.
  - surrounded metadata – i.e., tags, description, timestamp, location information, etc.
Introduction

- **MediaEval** challenge: *Social Event Detection Task (SED)*
- **Objective**: propose system able to detect/retrieve and group event-related images
- **Challenges**:
  - **Challenge 1**: *Find all soccer events taking place in Barcelona (Spain) and Rome (Italy) in the test collection. For each event provide all photos associated with it*
  - **Challenge 2**: *Find all events that took place in May 2009 in the venue named Paradiso (in Amsterdam, NL) and in the Parc del Forum (in Barcelona, Spain). For each event provide all photos associated with it.*
- **Dataset**:
  - 73K pictures gathered from Flickr.
  - surrounded metadata – i.e., tags, description, timestamp, location information, etc.
  - 80% of the pictures has been removed the gps information
Proposed Approach - Modules

Semantic Social Event Detection

Query Expansion

Locational terms of the query are expanded:
- DBPedia (challenge 1)
- LastFM (challenge 2)

Output: set of queries \{Q_1, ..., Q_N\}

Q_i = \{q_{i1}, ..., q_{iM}\}, subset of queries for each venue

q_{ij} = \{T, g\}, name of the venue and geographical location

Search

Solr search engine for indexing and searching

Spatial (latitude, longitude) + textual search

Title, Description and Tags as testual metadata

Output: result lists grouped by venue names
Proposed Approach - Modules

Query Expansion

- **Locational** terms of the query are expanded:
  - DBPedia (challenge 1)
  - LastFM (challenge 2)

- **Output**: set of queries \( \{Q_1, \ldots, Q_N\} \)
  - \( Q_i = \{q_{i1}, \ldots, q_{iM}\} \), subset of queries for each venue
  - \( q_{ij} = \{T, g\} \), name of the venue and geographical location
Query Expansion

- **Locational** terms of the query are expanded:
  - DBPedia (challenge 1)
  - LastFM (challenge 2)

- **Output**: set of queries \(\{Q_1, \ldots, Q_N\}\)
  - \(Q_i = \{q_{i1}, \ldots, q_{iM}\}\), subset of queries for each venue
  - \(q_{ij} = \{T, g\}\), name of the venue and geographical location

Search

- **Solr** search engine for indexing and searching
- Spatial (latitude, longitude) + textual search
- Title, Description and Tags as testual metadata

- **Output**: result lists grouped by venue names
Semantic Social Event Detection

Proposed Approach - Modules

- Clustering + Semantic Merge
- Results grouped by temporal tag (Qt Clustering)
- Semantic merge: based on entity names representing artist and event name
- Output: pictures grouped by temporal tag and venue

Refinement
- Refinement query for each cluster:
  - Top-k most frequent tags
  - Top-k most frequent entity names

Categorization: filter over the query result
Proposed Approach - Modules

Clustering + Semantic Merge

- Results grouped by **temporal tag** (Qt Clustering)
- **Semantic merge**: based on **entity names** representing **artist** and **event name**
- **Output**: pictures **grouped** by **temporal tag** and **venue**
Proposed Approach - Modules

Clustering + Semantic Merge

- Results grouped by **temporal tag** (Qt Clustering)
- **Semantic merge**: based on **entity names** representing **artist** and **event name**
- **Output**: pictures **grouped** by **temporal tag** and **venue**

Refinement

- **Refinement** query for **each cluster**:
  - Top-k most frequent **tags**
  - Top-k most frequent **entity names**
- **Categorization**: **filter** over the **query result**
Results and Discussions

**Challenge 1**
- **Run 1**: Categorization with only Tag
- **Run 2**: Categorization with all textual metadata

**Challenge 2**
- **Run 1**: No Refinement step
- **Run 2**: Refinement with top-100 tags
- **Run 3**: Refinement with entity names
Summary: Contributions vs Research Questions

- Tag metadata more representative
- Better performance using of entity names in event cluster refinement
- Refinement block useful for better completeness

[RQ3] How can raw geographical and temporal data be modelled and cooperate to improve and extend the retrieval performance of a tag-based search of event-related resources?
Outline

1. Introduction and Motivation
2. Research Questions
3. Part I - Mining and Detection of Events in Media Sharing Applications
4. Part II - Exploring Geo-Temporal Distribution of Tags in Social Media
5. Conclusion and Future Work
Part II: Overview

- **Tag-based Search** of Event-related pictures: extending *query expansion*, leveraging on *temporal* and *geographical* distribution of terms
Part II: Overview

- **Tag-based Search** of Event-related pictures: extending *query expansion*, leveraging on *temporal* and *geographical* distribution of terms

- **Case 1**: *Timestamped* Queries
Part II: Overview

- **Tag-based Search** of Event-related pictures: extending *query expansion*, leveraging on **temporal** and **geographical** distribution of terms

- **Case 1**: **Timestamped** Queries
  - [C3.1] *Analyzing effectiveness of textual metadata for retrieval purpose*
Part II: Overview

- **Tag-based Search** of Event-related pictures: extending query expansion, leveraging on **temporal** and **geographical** distribution of terms

- **Case 1**: Timestamped Queries
  - [C3.1] *Analyzing effectiveness of textual metadata for retrieval purpose*
  - [C3.2] *Improving tag-based search of event-related images of timestamped queries*
Part II: Overview

- **Tag-based Search** of Event-related pictures: extending *query expansion*, leveraging on *temporal* and *geographical* distribution of terms

  - **Case 1**: Timestamped Queries
    - [C3.1] *Analyzing effectiveness of textual metadata for retrieval purpose*
    - [C3.2] *Improving tag-based search of event-related images of timestamped queries*

  - **Case 2**: Not Timestamped Queries
Part II: Overview

Tag-based Search of Event-related pictures: extending query expansion, leveraging on temporal and geographical distribution of terms

Case 1: Timestamped Queries
- [C3.1] Analyzing effectiveness of textual metadata for retrieval purpose
- [C3.2] Improving tag-based search of event-related images of timestamped queries

Case 2: Not Timestamped Queries
- [C2.1] Definition of novel features from Geographical Distribution of tags for extraction of locational tags
Part II: Overview

- **Tag-based Search** of Event-related pictures: extending *query expansion*, leveraging on *temporal* and *geographical* distribution of terms

  - **Case 1**: *Timestamped* Queries
    - [C3.1] *Analyzing effectiveness of textual metadata for retrieval purpose*
    - [C3.2] *Improving tag-based search of event-related images of timestamped queries*

  - **Case 2**: *Not Timestamped* Queries
    - [C2.1] *Definition of novel features from Geographical Distribution of tags for extraction of locational tags*
    - [C2.2] *Definition of novel tag relatedness similarity measure improving retrieval effectiveness in QE framework*
Part II: Overview

- **Tag-based Search** of Event-related pictures: extending **query expansion**, leveraging on **temporal** and **geographical** distribution of terms

  - **Case 1**: **Timestamped** Queries
    - [C3.1] Analyzing **effectiveness** of **textual metadata** for retrieval purpose
    - [C3.2] Improving **tag-based search** of event-related images of **timestamped** queries

  - **Case 2**: **Not Timestamped** Queries
    - [C2.1] **Definition of novel features** from **Geographical Distribution of tags** for extraction of locational **tags**
    - [C2.2] **Definition of novel tag relatedness similarity measure** improving retrieval **effectiveness** in **QE framework**
    - [C3.3] Improving **tag-based search** of event-related images of **not timestamped** queries
Baseline Query Expansion Model

1. **Selecting/Scoring** expansion terms
   - Kullback-Liebler *divergence* based approach

\[
KL(e) = P_{Rel}(e) \log \left( \frac{P_{Rel}(e)}{P_{Coll}(e)} \right)
\]
Baseline Query Expansion Model

1. **Selecting/Scoring** expansion terms
   - Kullback-Liebler *divergence* based approach
     \[
     KL(e) = P_{Rel}(e) \log \left( \frac{P_{Rel}(e)}{P_{Coll}(e)} \right)
     \]
   - **Higher score**: terms with *lower* probability in the *whole* collection and *higher* probability in the top-k retrieved documents
Baseline Query Expansion Model

1. **Selecting/Scoring** expansion terms
   - Kullback-Liebler divergence based approach
     \[
     KL(e) = P_{Rel}(e) \log \left( \frac{P_{Rel}(e)}{P_{Coll}(e)} \right)
     \]
   - **Higher score**: terms with lower probability in the whole collection and higher probability in the top-k retrieved documents

2. **Re-weighting/Expanding** step
Baseline Query Expansion Model

1. **Selecting/Scoring** expansion terms
   - Kullback-Liebler divergence based approach
     \[ KL(e) = P_{Rel}(e) \log \left( \frac{P_{Rel}(e)}{P_{Coll}(e)} \right) \]
   - **Higher score**: terms with lower probability in the whole collection and higher probability in the top-k retrieved documents

2. **Re-weighting/Expanding** step
   - Rocchios Beta Equation
     \[ \hat{w}(t_q) = \frac{tf_{qt_q}}{\max tf_q} + \beta \frac{w(t_q)}{\max w} \]
Extended QE Model - *Temporal Awareness*
Event Related Image Search: Timestamped Query

Extended QE Model - *Temporal Awareness*

- Hypothesis:
Event Related Image Search: Timestamped Query

Extended QE Model - *Temporal Awareness*

- **Hypothesis:**
  - *Photo related to the same event: temporally clustered*
Hyphothesis:

- **Photo** related to the same event: *temporally clustered*
- **Good** candidate **expansion terms** $e$ must **co-occur** as much as possible with $q_j \in Q$, in documents **temporally close** to the user query
Event Related Image Search: Timestamped Query

Extended QE Model - *Temporal Awareness*

- **Hypothesis:**
  - *Photo* related to the same event: temporally clustered
  - *Good* candidate *expansion terms* $e$ must co-occur as much as possible with $q_j \in Q$, in documents *temporally close* to the user query

- **Temporal Proximity** Aware KL-Score (**KLTEMP**)
Event Related Image Search: Timestamped Query

Extended QE Model - *Temporal Awareness*

- **Hypothesis:**
  - Photo related to the same event: temporally clustered
  - Good candidate expansion terms $e$ must co-occur as much as possible with $q_j \in Q$, in documents temporally close to the user query

- **Temporal Proximity** Aware KL-Score ($KL_{\text{TEMP}}$)
  
  $$
  sc^T_Q(e) = \sum_{t_i \in Q} KL(\theta_L[e,t_i] \parallel \theta_{\text{Coll}}[e,t_i])
  $$
Event Related Image Search: Timestamped Query

Extended QE Model - *Temporal Awareness*

- **Hypothesis:**
  - Photo related to the same event: temporally clustered
  - Good candidate expansion terms $e$ must co-occur as much as possible with $q_j \in Q$, in documents temporally close to the user query

- **Temporal Proximity** Aware KL-Score ($KLTEMP$)
  
  $$sc_T^Q(e) = \sum_{t_i \in Q} KL(\theta^L_{[e,t_i]} || \theta^{Coll}_{[e,t_i]})$$

- Linearly **Combining** $KL(e)$ and $sc_T^Q(e)$ scores
Event Related Image Search: Timestamped Query

Extended QE Model - *Temporal Awareness*
Extended QE Model - *Temporal Awareness*

Original query: [atmedia, london, ajax]
Event Related Image Search: Timestamped Query

Extended QE Model - Temporal Awareness

Original query: [atmedia, london, ajax]

![Figure: Ranked candidate score](image)
Event Related Image Search: Timestamped Query

Extended QE Model - *Geo-Temporal Awareness*
Event Related Image Search: Timestamped Query

Extended QE Model - *Geo-Temporal Awareness*

- Hypothesis:
Event Related Image Search: Timestamped Query

Extended QE Model - *Geo-Temporal Awareness*

- Hypothesis:
  - Photo related to the **same event**: limited geographical areas
Event Related Image Search: Timestamped Query

Extended QE Model - *Geo-Temporal Awareness*

- **Hypothesis:**
  - Photo related to the *same event*: **limited geographical** areas
  - *World* divided in *tiles* $T_k$
Extended QE Model - *Geo-Temporal Awareness*

- Hypothesis:
  - Photo related to the **same event**: limited geographical areas
  - **World** divided in tiles $T_k$

- **Spatio-Temporal Aware** ranking KL-Score (**KLST**):
  \[
  sc^ST_Q(e, T_k) = \sum_{t_i \in Q} KL(\theta^C_{[e,t_i,T_k]} \| \theta^{Coll}_{[e,t_i,T_k]})
  \]
Event Related Image Search: Timestamped Query

Extended QE Model - Geo-Temporal Awareness

- **Hypothesis:**
  - Photo related to the **same event**: **limited geographical** areas
  - **World** divided in tiles $\mathcal{T}_k$

- **Spatio-Temporal Aware** ranking KL-Score ($\text{KLST}$):

$$sc^\text{ST}_Q(e, \mathcal{T}_k) = \sum_{t_i \in Q} KL(\theta^L_{[e, t_i, \mathcal{T}_k]} \parallel \theta^\text{Coll}_{[e, t_i, \mathcal{T}_k]})$$

- Calculated for **each tile**
Event Related Image Search: Timestamped Query

Extended QE Model - *Geo-Temporal Awareness*

- **Hypothesis:**
  - Photo related to the **same event**: limited geographical areas
  - World divided in tiles $T_k$

- **Spatio-Temporal Aware** ranking KL-Score (**KLST**):
  \[
  sc^{ST}_Q(e, T_k) = \sum_{t_i \in Q} KL(\theta^L_{[e, t_i, T_k]} \parallel \theta^{Coll}_{[e, t_i, T_k]})
  \]

  - Calculated for **each tile**
  - **Heuristic:** get $T_k$ maximizing $sc^{ST}_Q(e, T_k)$
Extended QE Model - *Geo-Temporal Awareness*

- **Hypothesis:**
  - Photo related to the **same event**: **limited geographical** areas
  - **World** divided in **tiles** $T_k$

- **Spatio-Temporal Aware** ranking KL-Score (**KLST**):

  $$sc_Q^{ST}(e, T_k) = \sum_{t_i \in Q} KL(\theta_{[e,t_i,T_k]}^{L} \parallel \theta_{[e,t_i,T_k]}^{Coll})$$

  - Calculated for **each tile**
  - **Heuristic**: get $T_k$ maximizing $sc_Q^{ST}(e, T_k)$

- Linearly **Combining** $KLTEMP(Q, e)$ and $sc_Q^{ST}(e, T_k)$ scores
Event Related Image Search: Timestamped Query

**Dataset**

*Upcoming*\(^TM\) **Dataset**
- \(~270K\) pictures
- Each photo related to a specific Upcoming Event
- \(9.5K\) unique events

**World Dataset**
- \(~88M\) pictures
- Flickr API

Merged in a single dataset
Experimental Setup

- Tag Indexing + retrieving: Terrier v3.5
- Test Queries:
  - 100 randomly selected pictures from Upcoming Dataset
  - short queries (less than 3 tags)
- Baseline Models:
  - Vector Space Model (TFIDF), Okapi BM25, Hiemstra LM
  - Default Terrier parameters for BM25 and LM
- Evaluation Metrics:
  - Main Average Precision (MAP)
  - R-Precision (RP)
- Query Expansion Setup
  - different number of $K$ documents and terms $n$
Event Related Image Search: Timestamped Query

Results and Discussion

<table>
<thead>
<tr>
<th></th>
<th>30_8</th>
<th>60_8</th>
<th>90_8</th>
<th>120_8</th>
<th>30_18</th>
<th>60_18</th>
<th>90_18</th>
<th>120_18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.40</td>
<td>0.42</td>
<td>0.44</td>
<td>0.46</td>
<td>0.48</td>
<td>0.50</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>KL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KL TEMP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KLST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparison of MAP and RP values of KLST against other query expansion models, as function of the values of $K$ and $n$ (expressed as $\{K\} \{n\}$), using BM25.
Event Related Image Search: Timestamped Query

Results and Discussion

- **Scalability** of the method
Results and Discussion

- **Scalability** of the method
Results and Discussion

- **Scalability** of the method
- Improving retrieval **effectiveness** of KLST and KLTEMP over baseline method
Results and Discussion

- **Scalability** of the method
- Improving retrieval **effectiveness** of KLST and KLTEMP over baseline method

Comparison of MAP and RP values of KLST against other query expansion models, as function of the values of $K$ and $n$ (expressed as $\{K\}_{-\{n\}}$), using **BM25**
Event Related Image Search: Timestamped Query

**Summary: Contributions vs Research Questions**

- **Extend** existing QE model incorporating **spatial** and **temporal** term distribution
- Experiments on **real data** (dataset of 88M of photos)
- **Short** queries

[RQ2] *Can a raw geographical and temporal term profile improve existing retrieval and query expansion models for the task of tag-based search of event-related resources?*

[RQ3] *How can raw geographical and temporal data be modelled and cooperate to improve and extend the retrieval performance of a tag-based search of event-related resources?*
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Research Questions</th>
<th>Part I</th>
<th>Part II</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

Event Related Image Search: Timestamped Query

**NEXT QUESTION**: How to improve retrieval effectiveness in event related search of **not timestamped** query?
Overview

- [C2.1] *Definition of novel features from Geographical Distribution of tags for extraction of locational tags*
- [C2.2] *Definition of novel tag relatedness similarity measure improving retrieval effectiveness in QE framework*
Point Process Theory Extended

**Picture Point Process**

*Point Process modeling the Spatial Distribution of Pictures taken in a 2-dimensional study region*

**Picture Point Pattern**

*Any realization D of the Picture Point Process*

**Tag-Point Pattern**

*subset of a Picture Point Pattern D. It is composed by the spatial point related to the geographical positions of the pictures annotated with a specific tag*
Pattern Analysis: Reduced second order moment

- **Point Pattern Analysis Objective**: determine if a given set of spatial points (Spatial Point Pattern) exhibits clustering, regularity or are randomly distributed within an area A subspace of $R^2$.

- Ripley’s $K$ function summarizing a spatial point pattern:
  - **Univariate** $\lambda K(h) = E(\# \text{points within a distance } h \text{ of an arbitrary point})$
  - **Bivariate** $\lambda_j K_{ij}(h) = E(\# \text{points } j \text{ within a distance } h \text{ of an arbitrary point } j)$

- Information about clustering and dispersion at different scale

- Complete Spatial Randomness Test (CSR):
  - $K(h) > \pi h^2$ clustering at scale $h$
  - $K(h) < \pi h^2$ dispersion at scale $h$
Objectives and Challenges
Objectives and Challenges

- **Main Objective**: Derive **indicators** estimating the **clustering tendency** of a Tag-Point Pattern
## Objectives and Challenges

- **Main Objective**: Derive *indicators* estimating the *clustering tendency* of a Tag-Point Pattern
- **Applications:**
Objectives and Challenges

- **Main Objective**: Derive indicators estimating the *clustering tendency* of a Tag-Point Pattern

- **Applications**:
  - ranking and extracting social *tags* indicating geographical point of interests (POI)
Objectives and Challenges

- **Main Objective**: Derive indicators estimating the **clustering tendency** of a Tag-Point Pattern

- **Applications**:
  - ranking and extracting social tags indicating geographical point of interests (POI)
  - query expansion framework in combination with other metadata
Event Related Image Search: Geographical Features Extraction

Objectives and Challenges

- **Main Objective**: Derive *indicators* estimating the *clustering tendency* of a Tag-Point Pattern

- **Applications**:
  - *ranking* and *extracting* social *tags* indicating geographical point of interests (POI)
  - *query expansion* framework in *combination* with other metadata

- Deal with Media Sharing Application *real data*
Objectives and Challenges

- **Main Objective**: Derive *indicators* estimating the *clustering tendency* of a Tag-Point Pattern

- **Applications**:
  - ranking and extracting social *tags* indicating geographical point of interests (*POI*)
  - *query expansion* framework in *combination* with other metadata

- Deal with Media Sharing Application *real data*
  - size
  - inhomogeneity
**Objectives and Challenges**

- **Main Objective**: Derive indicators estimating the *clustering tendency* of a Tag-Point Pattern

- **Applications**:
  - *ranking* and extracting social *tags* indicating geographical point of interests (POI)
  - *query expansion* framework in combination with other metadata

- Deal with Media Sharing Application *real data*
  - size
  - inhomogeneity
Event Related Image Search: Geographical Features Extraction

Objectives and Challenges: Inhomogeneity

Figure: Example of point pattern of the tag night
Objectives and Challenges: Inhomogeneity

Figure: Related underlying Picture Point Pattern
Objectives and Challenges: Inhomogeneity

Figure: Example of point pattern of the tag night over the underlying distribution
Tackle the Challenges

- **Inhomogeneity of the data:**
  - Case-Control Analysis: *detect the clustering tendency of the case according to the concentration of the control*
    \[
    \hat{D}_{wi}(h) = \hat{K}_{wi}(h) - \hat{K}_{MAX,all}(h)
    \]
  - CSR Testing: \( \hat{D}_{wi}(h) \leq 0 \) indicates clustering

- **Dataset size:**
  - *Subsample similarity*: sub-sampling \( C \)-times the control dataset for estimation of the envelope
Proposed Estimators

- \( I_{\text{MAX}} \): Area underlying \( \hat{D}_{wi}(h) \) in the considered scale interval
- \( I_{\text{SUM}} \): Maximum function value \( \hat{D}_{wi}(h) \) over the scale interval
- (Maximum significant radius of aggregation)
### Results and Discussions

- **Tag ranking** according $I_{SUM}$ and $I_{MAX}$:

<table>
<thead>
<tr>
<th>TAG</th>
<th>#</th>
<th>$I_{MAX}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>royalchelseahospital</td>
<td>339</td>
<td>1</td>
</tr>
<tr>
<td>paternostersquare</td>
<td>209</td>
<td>0.999</td>
</tr>
<tr>
<td>victoriaalbertmuseum</td>
<td>174</td>
<td>0.993</td>
</tr>
<tr>
<td>cannizaropark</td>
<td>116</td>
<td>0.985</td>
</tr>
<tr>
<td>whitetower</td>
<td>115</td>
<td>0.982</td>
</tr>
<tr>
<td>smithfieldmarket</td>
<td>102</td>
<td>0.979</td>
</tr>
<tr>
<td>thebritishmuseum</td>
<td>130</td>
<td>0.977</td>
</tr>
<tr>
<td>heygateestate</td>
<td>163</td>
<td>0.976</td>
</tr>
<tr>
<td>fourthplinth</td>
<td>159</td>
<td>0.976</td>
</tr>
<tr>
<td>zsl</td>
<td>282</td>
<td>0.972</td>
</tr>
</tbody>
</table>

**Table:** Top-10 tags extracted, ordering the list by $I_{MAX}$

<table>
<thead>
<tr>
<th>TAG</th>
<th>#</th>
<th>$I_{SUM}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>royalchelseahospital</td>
<td>339</td>
<td>1</td>
</tr>
<tr>
<td>fourthplinth</td>
<td>159</td>
<td>0.985</td>
</tr>
<tr>
<td>victoriaalbertmuseum</td>
<td>174</td>
<td>0.972</td>
</tr>
<tr>
<td>paternostersquare</td>
<td>209</td>
<td>0.969</td>
</tr>
<tr>
<td>leadenhallmarket</td>
<td>184</td>
<td>0.942</td>
</tr>
<tr>
<td>limestreet</td>
<td>152</td>
<td>0.937</td>
</tr>
<tr>
<td>turbinehall</td>
<td>221</td>
<td>0.936</td>
</tr>
<tr>
<td>wellingtonarch</td>
<td>125</td>
<td>0.935</td>
</tr>
<tr>
<td>oneandother</td>
<td>150</td>
<td>0.935</td>
</tr>
<tr>
<td>thebritishmuseum</td>
<td>130</td>
<td>0.933</td>
</tr>
</tbody>
</table>

**Table:** Top-10 tags extracted, ordering the list by $I_{SUM}$
Results and Discussions

- **Manual evaluation of top-100 extracted tags**: at rank \( n \)

<table>
<thead>
<tr>
<th></th>
<th>( \hat{i}_{\text{SUM}} )</th>
<th>( \hat{i}_{\text{ord\ SUM}} )</th>
<th>( \hat{i}_{\text{MAX}} )</th>
<th>( \hat{i}_{\text{ord\ MAX}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P@5 )</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>( P@10 )</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>( P@20 )</td>
<td>0.95</td>
<td>1.0</td>
<td>0.90</td>
<td>1.0</td>
</tr>
<tr>
<td>( P@50 )</td>
<td>0.92</td>
<td>1.0</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>( P@100 )</td>
<td>0.92</td>
<td>0.94</td>
<td>0.92</td>
<td>0.92</td>
</tr>
</tbody>
</table>

**Table**: Precisions values using the different estimators
Results and Discussion

- Comparison between the proposed estimators
### Event Related Image Search: Geographical Features Extraction

#### Summary: Contributions vs Research Questions

- Model the **Geo-Tagged resources** in the field of Spatial Point Pattern
- Indicators able to
  - rank/extract social Tags related to geographical hot-spot
  - define a similarity between the geographical distribution of two terms
- Analyse clustering tendency at different scales
- Deal with large dataset and inhomogeneity of real data
- **Comparison** and analysis of behaviours of proposed indicators

**[RQ5]** *How can a geographical and temporal term profile be used to infer semantics to the tags?*
NEXT QUESTION: How to apply these features for improving retrieval effectiveness in event-related search of not timestamped query?
Event Related Image Search: Not Timestamped Query

Main Idea

- Incorporate spatial features in a retrieval framework
- Machine learning-based query expansion model
- Exploring the candidate term expansion over more than one textual dimension
- Improve ability of term expansion selection
- Good expansion term: improving the retrieval performance of the initial query in terms of average precision (AP) gain

Textual Features
Temporal Features
Geographical Features
Main Idea

- Incorporate **spatial features** in a retrieval framework
Main Idea

- Incorporate **spatial features** in a retrieval framework
- **Machine learning-based** query expansion model
Main Idea

- Incorporate **spatial features** in a retrieval framework
- **Machine learning-based** query expansion model
- Exploring the candidate term expansion over **more than one textual dimension**
Main Idea

- Incorporate **spatial features** in a retrieval framework
- **Machine learning-based** query expansion model
- **Exploring** the candidate term expansion over *more than one textual dimension*
- Improve ability of term expansion selection
Main Idea

- Incorporate **spatial features** in a retrieval framework
- **Machine learning-based** query expansion model
- **Exploring** the candidate term expansion over **more than one textual dimension**
- Improve ability of term expansion selection
  - **good expansion term**: term **improving** the **retrieval performance** of the initial query in term of average precision (AP) gain
Main Idea

- Incorporate **spatial features** in a retrieval framework
- **Machine learning-based** query expansion model
- **Exploring** the candidate term expansion over **more than one textual dimension**
- Improve ability of term expansion selection
  - **good expansion term**: term **improving** the **retrieval performance** of the initial query in term of average precision (AP) gain
  - **Binary classification** problem (discriminating good expansion terms from bad ones)
Main Idea

- Incorporate **spatial features** in a retrieval framework
- **Machine learning-based** query expansion model
- **Exploring** the candidate term expansion over **more than one textual dimension**
- Improve ability of term expansion selection
  - **good expansion term**: term improving the **retrieval performance** of the initial query in term of average precision (AP) gain
  - **Binary classification** problem (discriminating good expansion terms from bad ones)
  - expansion terms as **features vector**
Main Idea

- Incorporate **spatial features** in a retrieval framework
- **Machine learning-based** query expansion model
- **Exploring** the candidate term expansion over **more than one textual dimension**
- Improve ability of term expansion selection
  - **good expansion term**: term **improving** the **retrieval performance** of the initial query in term of average precision (AP) gain
  - **Binary classification** problem (discriminating good expansion terms from bad ones)
  - expansion terms as **features vector**

\[
e 
\]

- **Textual Features**
- **Temporal Features**
- **Geographical Features**
Event Related Image Search: Not Timestamped Query

Feature Set
Event Related Image Search: Not Timestamped Query

**Feature Set**

- **Textual/Term Features** $x'(e_i, Q)$
  - Used as *baseline* features.
  - Traditional *statistical* term features as TF, IDF, DF
  - *Both* related to *feedback* document set and *whole* collection
Event Related Image Search: Not Timestamped Query

Feature Set

- **Textual/Term Features** $\mathcal{X}(e_i, Q)$
  - Used as *baseline* features.
  - Traditional *statistical* term features as TF, IDF, DF
  - *Both* related to *feedback* document set and *whole* collection

- **Temporal Features** $\mathcal{Y}(e_i, Q)$
  - Capture the *contribution* of the tag *over the time*
  - *Peakdness* of the tag time series: *kurtosis*
  - *Randomness* of a term over the time: *autocorrelation*
  - Temporal *similarity* of two series: *cross-correlation*
Feature Set

- **Textual/Term Features** $\mathcal{X}(e_i, Q)$
  - Used as **baseline** features.
  - Traditional **statistical** term features as TF, IDF, DF
  - **Both** related to feedback document set and **whole** collection

- **Temporal Features** $\mathcal{Y}(e_i, Q)$
  - Capture the **contribution** of the tag over the time
  - **Peakdness** of the tag time series: *kurtosis*
  - **Randomness** of a term over the time: *autocorrelation*
  - Temporal **similarity** of two series: *cross-correlation*

- **Spatial Features** $\mathcal{Z}(e_i, Q)$ [24, 89]
  - Good expansion term $e$ are **spatially correlated** with some query terms $q_i$
  - Calculated for each tile $T_{q_i,e}$
  - $T_{q_i,e}$ selection process:
    - Tile containing picture annotated with $q_i$ and $e$
    - Top-$K$ pictures for each tile
    - Tile as a document
    - Ranked according TF-IDF score
  - **single term** + **term-to-term** spatial features
Event Related Image Search: Not Timestamped Query

Experimental Setup

- Tag Indexing + retrieving: Terrier v3.5
- Test Queries:
  - 150 randomly selected pictures from Upcoming Dataset: 100 as training set, 50 as test set
  - short queries (less than 3 tags)
- Baseline Models:
  - **Retrieval Models**: Vector Space Model (TFIDF), Okapi BM25
  - **QE Models**: KL divergence (KL), machine learning (KLML) with baseline features [89], Mixture Model (MM), Relevance Model (RM)
  - **Geo-temporal similarity** model (ZKYC) [189]
- Evaluation Metrics:
  - Main Average Precision (MAP)
Results and Discussions

Event Related Image Search: Not Timestamped Query

Retrieval effectiveness: KLST outperform baseline methods. Document number breakpoint for QE: higher, exploring the candidate over all the three dimensions. Feature Analysis features related to spatial/temporal correlation of distribution of pictures including both $q_i$ and $e_i$ (i.e., $KURT_{12}$, $AC_{12}$ and $RDMD_i (Q_{+} e)$, $RDPA_i (Q_{+} e)$). Features related to correlation between spatial/temporal distribution of $q_i$ and spatial/temporal distribution of $e_i$ (i.e., $CC_{12}$ and $RDMD_i (Q, e)$, $RDPA_i (Q, e)$).
Results and Discussions

- Retrieval effectiveness: *KLST* outperform baseline methods
Results and Discussions

- Retrieval effectiveness: \textit{KLST} outperform baseline methods.
- Document number \textbf{breakpoint} for QE: \textbf{higher}, exploring the candidate over all the \textbf{three} dimensions.
Results and Discussions

- Retrieval effectiveness: *KLST* outperform baseline methods
- Document number breakpoint for QE: *higher*, exploring the candidate over all the *three* dimensions
- Feature Analysis
  - features related to *spatial/temporal correlation* of distribution of *pictures* including both $q_i$ and $e$ 
    \(i.e., KURT12, AC12\) and \(D_{RDMD}(Q + e), D_{RDPA}(Q + e)\)
  - features related to *correlation between spatial/temporal* distribution of $q_i$ and spatial/temporal distribution of $e$ 
    \(i.e., CC\) and \(D_{RDMD}(Q, e), D_{RDPA}(Q, e)\)
Summary: Contributions vs Research Questions

- Approach for search event related pictures from a general textual query
- Spatial features in combination with other features: effectiveness in the retrieval model
- Use of rigorous spatial statistics
- General query expansion framework incorporating heterogeneous feature

[RQ2] Can a raw geographical and temporal term profile improve existing retrieval and query expansion models for the task of tag-based search of event-related resources?

[RQ3] How can raw geographical and temporal data be modelled and cooperate to improve and extend the retrieval performance of a tag-based search of event-related resources?

[RQ4] Can a geographical and temporal tag profile improve tag co-occurrence and tag similarity?
Outline

1. Introduction and Motivation
2. Research Questions
3. Part I - Mining and Detection of Events in Media Sharing Applications
4. Part II - Exploring Geo-Temporal Distribution of Tags in Social Media
5. Conclusion and Future Work
List of Publications


List of Papers
## Papers vs Contributions

<table>
<thead>
<tr>
<th>Papers</th>
<th>C1.1</th>
<th>C1.2</th>
<th>C2.1</th>
<th>C2.2</th>
<th>C3.1</th>
<th>C3.2</th>
<th>C3.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P4</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P5</td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>P7</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

**Table:** Relations between contributions and publications.
Future Work

- Event Detection and Mining: distributed suffix-tree
- Query Expansion framework: Cross-media
- Query Expansion framework: more general information retrieval setting
- Both: exploration of features related user behaviour
Key References

Key References

Thanks for the Attention!
Questions?
Events in Media Sharing Applications - Media Type

- **Image-based sharing application** - e.g., Flickr, Panoramio, Pixable
  - MediaEval benchmarking initiative, Social Media Event Detection task
  - S.O.A. - Clustering Algorithms + Supervised classifiers

- **Microblogging services** (e.g., Twitter)
  - Real-time and dynamic data
  - Short textual messages

- **Video-sharing applications** (e.g., Youtube, Vimeo)
  - TRECVID multimedia event detection task
  - S.O.A. - Video and events as BOW model
Events in Media Sharing Applications - **Features**

- **Textual Features** - extracted from Tags, Description, Title etc.
  - term frequency (TF) and TF-IDF
  - audio transcription via ASR
  - optical character recognition (OCR)

- **Temporal Features** - extracted from metadata or EXIF

- **Geographical Features** - extracted from textual metadata or EXIF
  - location name
  - latitude and longitude

- **Visual Features** - extracted from video and image
  - local and scale invariant features (e.g., \textit{SIFT} and \textit{SURF})
  - global features (e.g., \textit{color histogram} and \textit{Gabor features})
  - texture descriptors (e.g., \textit{EHD})
## Results and Discussions

<table>
<thead>
<tr>
<th></th>
<th>100 m WEEK</th>
<th>100 m DAY</th>
<th>200 m WEEK</th>
<th>200 m DAY</th>
<th>500 m WEEK</th>
<th>500 m DAY</th>
<th>1000 m WEEK</th>
<th>1000 m DAY</th>
</tr>
</thead>
<tbody>
<tr>
<td>P@5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>P@10</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>P@20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>P@50</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.90</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

**Table:** P@k of event clusters extraction with K = 1

<table>
<thead>
<tr>
<th></th>
<th>100 m WEEK</th>
<th>100 m DAY</th>
<th>200 m WEEK</th>
<th>200 m DAY</th>
<th>500 m WEEK</th>
<th>500 m DAY</th>
<th>1000 m WEEK</th>
<th>1000 m DAY</th>
</tr>
</thead>
<tbody>
<tr>
<td>P@5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>P@10</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>P@20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>1.0</td>
<td>0.95</td>
<td>1.0</td>
<td>0.95</td>
</tr>
<tr>
<td>P@50</td>
<td>0.98</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
</tr>
</tbody>
</table>

**Table:** P@k of event clusters extraction with K = 0.8
Results and Discussions
Results and Discussions

Figure: Part of the ranked list of the extracted event clusters with $K = 1$.
## Results and Discussions

<table>
<thead>
<tr>
<th></th>
<th>$K = 1$</th>
<th>$K = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of Merged Clusters</td>
<td>4.3</td>
<td>5.65</td>
</tr>
<tr>
<td>Average number of Tags</td>
<td>2.95</td>
<td>3.1</td>
</tr>
</tbody>
</table>

**Table:** Summary of the average number of merged clusters and average number of tags for each event cluster detected for different values of $K$. These are results from the top 20 event clusters.
Extended QE Model -
\textit{NaivemainBaselineQueryExpansionModel}

- Temporal-Proximity re-ranking
- push up documents with higher temporal proximity in the ranked list of feedback documents
- combine temporal and textual similarity:
  - not unified measures
  - \textit{rCombMNZ} merging function
Tag Point Pattern Examples

hydepark

hydeparkcorner

kensingtonchelsea
Main Objective: Extract of group of images related to an event from a collection of pictures

Figure: Overview of the system
Proposed Approach - Overview

- Query Expansion
- Search
- Categorization
- Clustering
- Semantic Merge
- Refinement
- Categorization
- SparQL EndPoint
- lost.fm
- flickr
- Clustered List
**Point Process Theory**

**Spatial Point Pattern**
set of spatial point, not regularly distributed in an area \( A \subset R^2 \)

**Point Process**
stochastic process used to model an irregular point pattern

**Poisson Point Process**
represent the theoretical model for Complete Spatial Randomness
- **homogeneous** if intensity \( \lambda \) is constant
- **inhomogeneous** if intensity \( \lambda(x) \) if function of the location \( x \)
## Results and Discussions

<table>
<thead>
<tr>
<th></th>
<th>ANU</th>
<th>ITI</th>
<th>EURECOM</th>
<th>LIA</th>
<th>NTNU</th>
<th>QMUL</th>
<th>VTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>84.86</td>
<td>90.58</td>
<td>97.75</td>
<td>7.49</td>
<td><strong>92.47</strong></td>
<td>76.81</td>
<td>—</td>
</tr>
<tr>
<td>Recall</td>
<td>52.54</td>
<td>67.58</td>
<td>42.38</td>
<td>15.62</td>
<td><strong>43.16</strong></td>
<td>62.11</td>
<td>—</td>
</tr>
<tr>
<td>F-Measure</td>
<td>64.90</td>
<td>77.40</td>
<td>59.13</td>
<td>10.13</td>
<td><strong>58.85</strong></td>
<td>68.68</td>
<td>—</td>
</tr>
<tr>
<td>NMI</td>
<td>0.237</td>
<td>0.618</td>
<td>0.247</td>
<td>0.026</td>
<td><strong>0.475</strong></td>
<td>0.414</td>
<td>—</td>
</tr>
</tbody>
</table>

**Table:** Results of the first challenge among the groups

<table>
<thead>
<tr>
<th></th>
<th>ANU</th>
<th>ITI</th>
<th>EURECOM</th>
<th>LIA</th>
<th>NTNU</th>
<th>QMUL</th>
<th>VTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>70.79</td>
<td>54.31</td>
<td>70.99</td>
<td>14.37</td>
<td><strong>78.85</strong></td>
<td>42.14</td>
<td>73.79</td>
</tr>
<tr>
<td>Recall</td>
<td>43.90</td>
<td>80.61</td>
<td>67.01</td>
<td>37.99</td>
<td><strong>56.83</strong></td>
<td>27.13</td>
<td>64.21</td>
</tr>
<tr>
<td>F-Measure</td>
<td>50.44</td>
<td>64.90</td>
<td>68.95</td>
<td>12.44</td>
<td><strong>66.05</strong></td>
<td>33.01</td>
<td>68.67</td>
</tr>
<tr>
<td>NMI</td>
<td>0.448</td>
<td>0.385</td>
<td>0.617</td>
<td>—0.013</td>
<td><strong>0.645</strong></td>
<td>0.498</td>
<td>0.678</td>
</tr>
</tbody>
</table>

**Table:** Results of the second challenge among the groups
Results and Discussions

- Retrieval effectiveness: KLST outperform baseline methods

<table>
<thead>
<tr>
<th>#Doc</th>
<th>#Term</th>
<th>BM25</th>
<th>KL</th>
<th>KL+ZKYC(189)</th>
<th>KL.T</th>
<th>KLST</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>15</td>
<td>0.448</td>
<td>0.4601</td>
<td>0.4614</td>
<td>0.4752</td>
<td>0.4816</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.448</td>
<td>0.4605</td>
<td>0.4626</td>
<td>0.4755</td>
<td>0.4834</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>0.448</td>
<td>0.4618</td>
<td>0.4638</td>
<td>0.4761</td>
<td>0.4838</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>0.448</td>
<td>0.4618</td>
<td>0.4634</td>
<td>0.4764</td>
<td>0.4833</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>0.448</td>
<td>0.4618</td>
<td>0.4624</td>
<td>0.4761</td>
<td>0.4836</td>
</tr>
<tr>
<td>40</td>
<td>15</td>
<td>0.448</td>
<td>0.4708</td>
<td>0.4744</td>
<td>0.4783</td>
<td>0.4870</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.448</td>
<td>0.4714</td>
<td>0.4738</td>
<td>0.4793</td>
<td>0.4880</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>0.448</td>
<td>0.4705</td>
<td>0.4734</td>
<td>0.4813</td>
<td>0.4885</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>0.448</td>
<td>0.4726</td>
<td>0.4757</td>
<td>0.4943</td>
<td>0.4918</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>0.448</td>
<td>0.4717</td>
<td>0.4745</td>
<td>0.4827</td>
<td>0.4910</td>
</tr>
<tr>
<td>60</td>
<td>15</td>
<td>0.448</td>
<td>0.4665</td>
<td>0.4674</td>
<td>0.4815</td>
<td>0.4848</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.448</td>
<td>0.4685</td>
<td>0.4696</td>
<td>0.4818</td>
<td>0.4906</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>0.448</td>
<td>0.4704</td>
<td>0.4733</td>
<td>0.4856</td>
<td>0.4915</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>0.448</td>
<td>0.4712</td>
<td>0.4721</td>
<td>0.4977</td>
<td>0.4957</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>0.448</td>
<td>0.4703</td>
<td>0.4731</td>
<td>0.4967</td>
<td>0.4957</td>
</tr>
<tr>
<td>80</td>
<td>15</td>
<td>0.448</td>
<td>0.4697</td>
<td>0.4706</td>
<td>0.4803</td>
<td>0.4894</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.448</td>
<td>0.4699</td>
<td>0.4711</td>
<td>0.4847</td>
<td>0.4935</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>0.448</td>
<td>0.4718</td>
<td>0.4733</td>
<td>0.4862</td>
<td>0.4951</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>0.448</td>
<td>0.4712</td>
<td>0.4731</td>
<td>0.4884</td>
<td>0.4979</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>0.448</td>
<td>0.4719</td>
<td>0.4741</td>
<td>0.4890</td>
<td>0.5001</td>
</tr>
<tr>
<td>100</td>
<td>15</td>
<td>0.448</td>
<td>0.4613</td>
<td>0.4621</td>
<td>0.4701</td>
<td>0.4727</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.448</td>
<td>0.4611</td>
<td>0.4619</td>
<td>0.4755</td>
<td>0.4802</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>0.448</td>
<td>0.4634</td>
<td>0.4642</td>
<td>0.4781</td>
<td>0.4840</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>0.448</td>
<td>0.4613</td>
<td>0.4621</td>
<td>0.4803</td>
<td>0.4870</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>0.448</td>
<td>0.4621</td>
<td>0.4631</td>
<td>0.4820</td>
<td>0.4891</td>
</tr>
<tr>
<td>120</td>
<td>15</td>
<td>0.448</td>
<td>0.4592</td>
<td>0.4591</td>
<td>0.4681</td>
<td>0.4709</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.448</td>
<td>0.4589</td>
<td>0.4610</td>
<td>0.4760</td>
<td>0.4814</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>0.448</td>
<td>0.4606</td>
<td>0.4625</td>
<td>0.4771</td>
<td>0.4871</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>0.448</td>
<td>0.4589</td>
<td>0.4596</td>
<td>0.4829</td>
<td>0.4896</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>0.448</td>
<td>0.4595</td>
<td>0.4606</td>
<td>0.4845</td>
<td>0.4914</td>
</tr>
</tbody>
</table>
Document number breakpoint for QE
higher, exploring the candidate over all the three dimensions

![Graph showing ΔMAP(%) vs. #Doc for Best KL, Best KL_T, and Best KL_ST.](image-url)
Results and Discussions

Feature Analysis

- Feature related to spatial/temporal correlation related distribution of pictures including both $q_i$ and $e$ 
  ($i.e., \text{KURT}12$, $\text{AC}12$ and $D_{RDMD}(Q + e), D_{RDPA}(Q + e)$)

- Features related to correlation between spatial/temporal distribution of $q_i$ and spatial/temporal distribution of $e$
  ($i.e., \text{CC}$ and $D_{RDMD}(Q, e), D_{RDPA}(Q, e)$)