

# Information Extraction from Social Sites

Yiannis Kompatsiaris

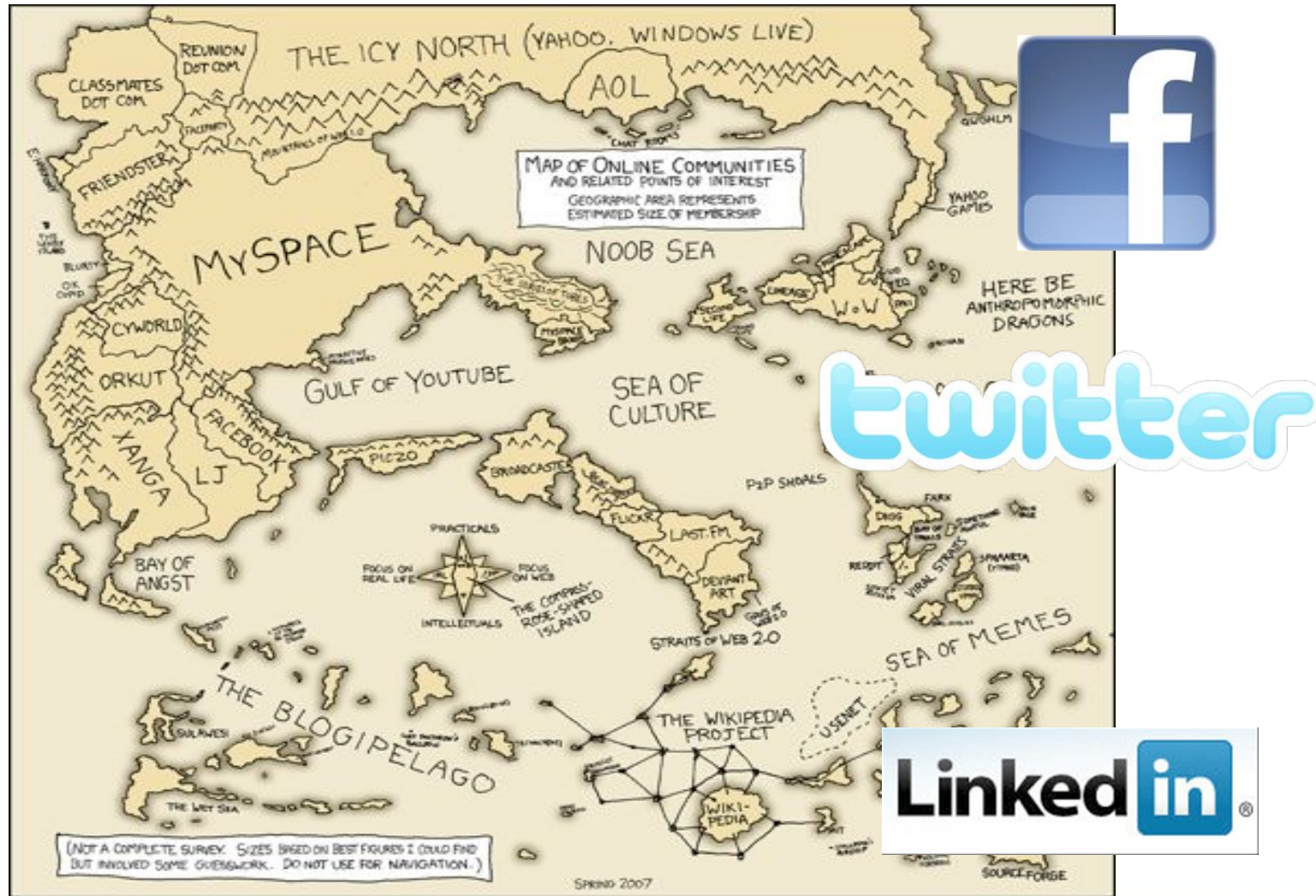
Eirini Yannakidou, Symeon Papadopoulos,  
Spyros Nikolopoulos, Elisavet Chatzilari

Athena Vakali, AUTH

# Contents

- Introduction
- Clustering in Social Media
- Social media “teacher” of the machine
- Community detection in Social Media
- WeKnowIt project
- Conclusions - Issues

# Our world today (already old)



# Web 2.0 content (July 2010)

## flickr

- 3,190 uploads in the last minute
- 3.2 million things geotagged this month
- 4,754,012,299 photos (2 July 2010)

## YouTube

- 24h of video content uploaded every minute
- 2 billion movies watched every day

## facebook

- More than 400 million active users
- More than 200 million users log on at least once each day
- 2.5 billion photos uploaded each month



## Winner



The winner of the WeKnowIt Grand Travel Challenge

# Tags everywhere

tag cloud  
Call for papers  
CIVR2009  
Collective Intelligence Conference  
content popularity  
images Invited  
Talk IVUS  
Multimedia Retrieval  
Multimedia Semantics  
News object detection  
Ontologies  
Patents proceedings  
Project Semantic Multimedia  
Semantics social bookmarking tutorial  
video retrieval  
WeKnowIt  
Workshop  
WWW2009  
more tags

Search, Describe content, Extract knowledge

amsterdam animal april architecture art australia baby barcelona  
beach berlin birthday blackandwhite blue boston bsp busing bw  
california cameraphone canada car cat cats chicago  
china christmas church city clouds concert day in dog england  
europe family festival florida flower flowers food france  
friends fun garden germany graduation graffiti green hawaii  
holiday home house india italy japan june kids london  
light london macro may me mexico moblog  
music nature new newyork newyorkcity newzealand night nyc  
paris park party people photo portrait red  
sanfrancisco school scotland seattle sky  
snow spain spring street summer sunset taiwan thailand tokyo  
travel tree trees trip uk unbound urban usa vacation  
vancouver water wedding white winter


























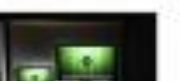


# Very low precision

Search Photos Groups People

Everyone's Uploads   [Full Text](#) | [Tags Only](#)  
[Advanced Search](#)

Sort: **Relevant** | [Recent](#) | [Interesting](#) View: **Small** | [Medium](#) | [Detail](#) | [Slideshow](#)

 From sunnyhung	 From Warm Tr...	 From Hugo...	 From B@rtan@	 From Glenn Waters...	 From fyzor	 From Taxi Lady...
 From ( karen )	 From HAZEL- S b G r B	 From feumpungle	 From Earl - What...	 From Bald Monk	 From rik@flickr	 From jonbradbury
 From rolor2000	 From dave-	 From jaudrus	 From Marchissimo	 From nebarria	 From rikpk	 From photophile
 From anny johanna	 From rnyick	 From fernando780	 From jordanmeric...	 From humedri		

# Very low recall



## Tags

- Property#1
- Canada
- photo
- image
- digital
- urban
- Halifax
- park
- morning
- afternoon
- night
- Pentax K20D
- Sigma 70-300
- early
- Sackville

# Can we improve things?

The screenshot shows a search interface with a 'Search' bar at the top. Below the search bar, there are tabs for 'Photos', 'Groups', and 'People'. A 'SEARCH' button is visible on the right. The main content area is titled 'Tag Clusters' and lists three clusters of photos based on tags:

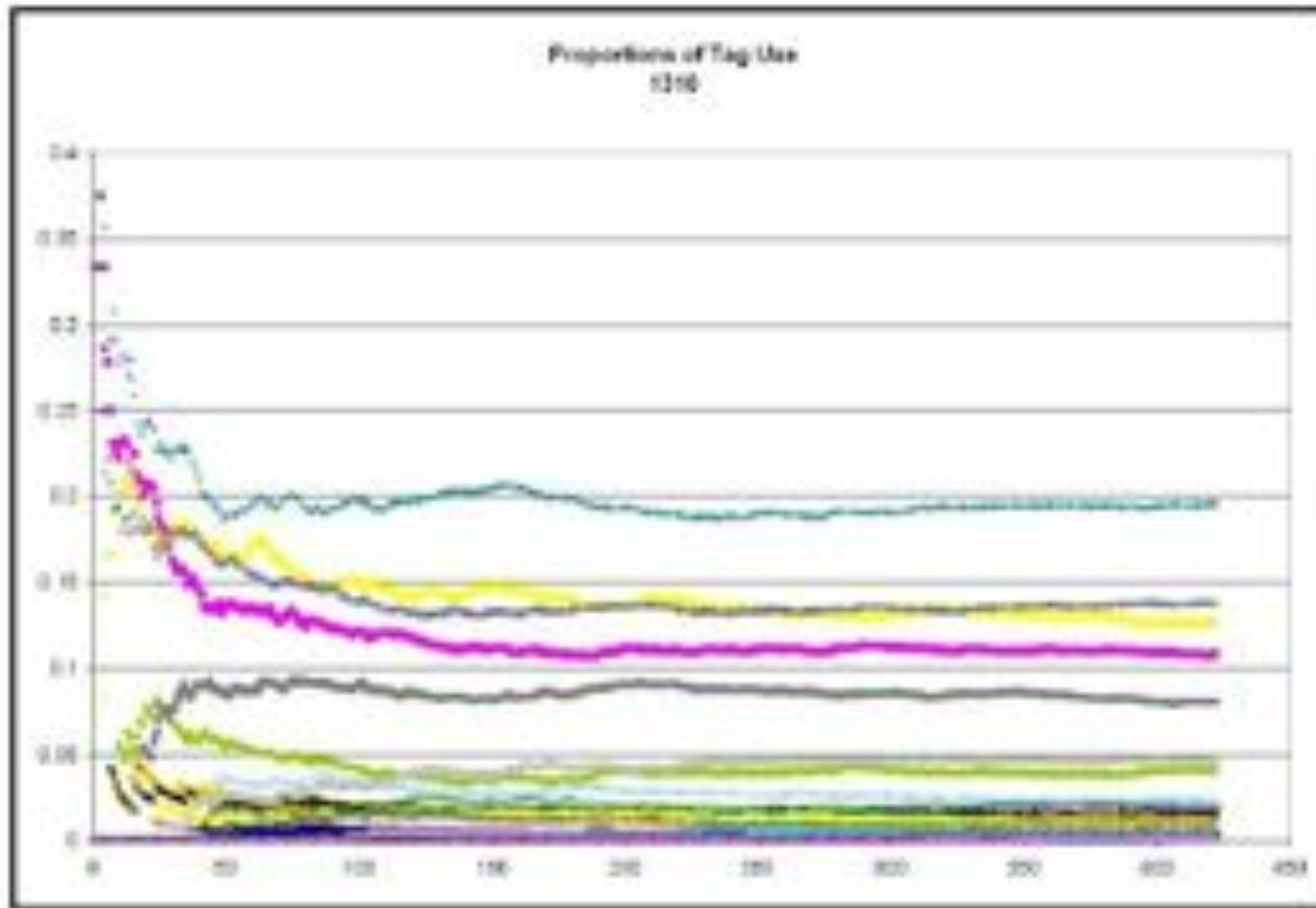
- Photos with tags like nyc, newyork and manhattan
- Photos with tags like fruit, red and green
- Photos with tags like ipod, iphone and music

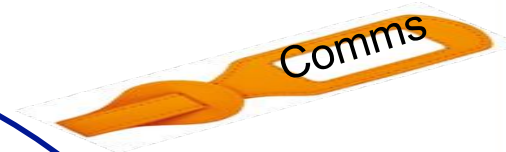
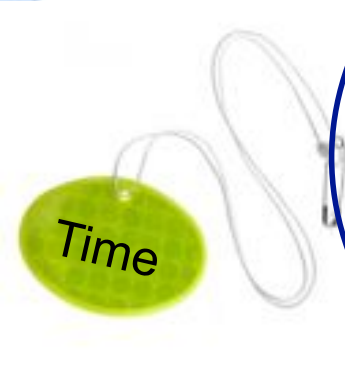
Below the clusters, there are several photo thumbnails with their respective sources. The sources include: From sonnyhung, From War, From ( karen ), From HAZEL-8, b-G-8, From jnradbury, From amy johanna, From nrvica, From fernando780, From jordanmarric..., and From hamedni.

By combining information from many photos - tags, it seems that we can **Stable patterns** in tagging systems over time



# Stable tagging patterns





Deutsches Eck from Ehrenbreitstein  
Fortress, Koblenz, Germany



When you're high up on the hill above Koblenz at Ehrenbreitstein Fortress you can get a great panoramic view of the city and the surrounding area.

flickr®

by [schaengel](#)

121 comments 69 faves

Tagged with [koblenz](#), [ehrenbreitstein](#) ...  
Taken on November 15, 2009, uploaded  
November 17, 2009

See more of [schaengel](#) photos, or visit  
his profile.

## ... and more: Travel trends using flickr



Trace Flickr users from a chronologically ordered set of geographically referenced photos

*Who are the Italians and who are the Americans?*

*MIT SENSEABLE CITY LAB, "The World's eyes"*

## What else we can do?

Tags that are “representative” for a geographical area

- 1. Clustering of photos
  - K-means, based on their location [Kennedy07]
- 2. Rank each cluster’s tags
- 3. Get tags above a certain threshold

Contribute to our understanding of the world

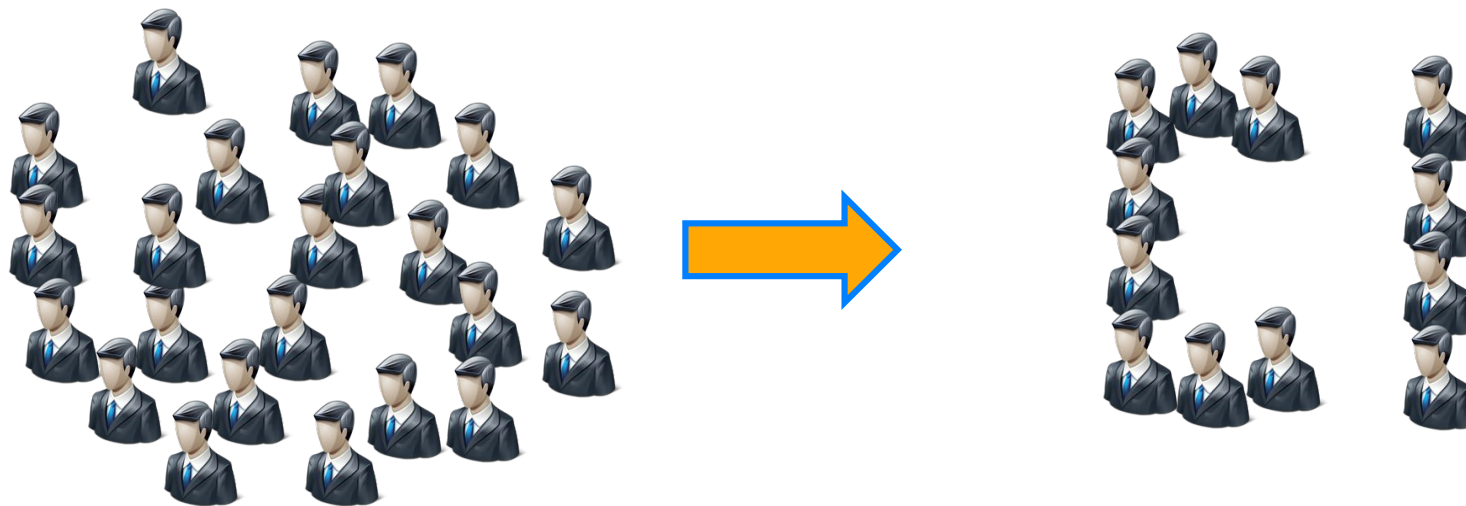


*Representative tags for San Francisco [Kennedy07]*



# Collective Intelligence, PeopleWeb, Crowdsourcing, Wisdom of crowds ...

Collective Intelligence is the Intelligence which emerges from the collaboration, competition and coordination among individuals.



...an Intelligence greater than the sum of the individuals' intelligence

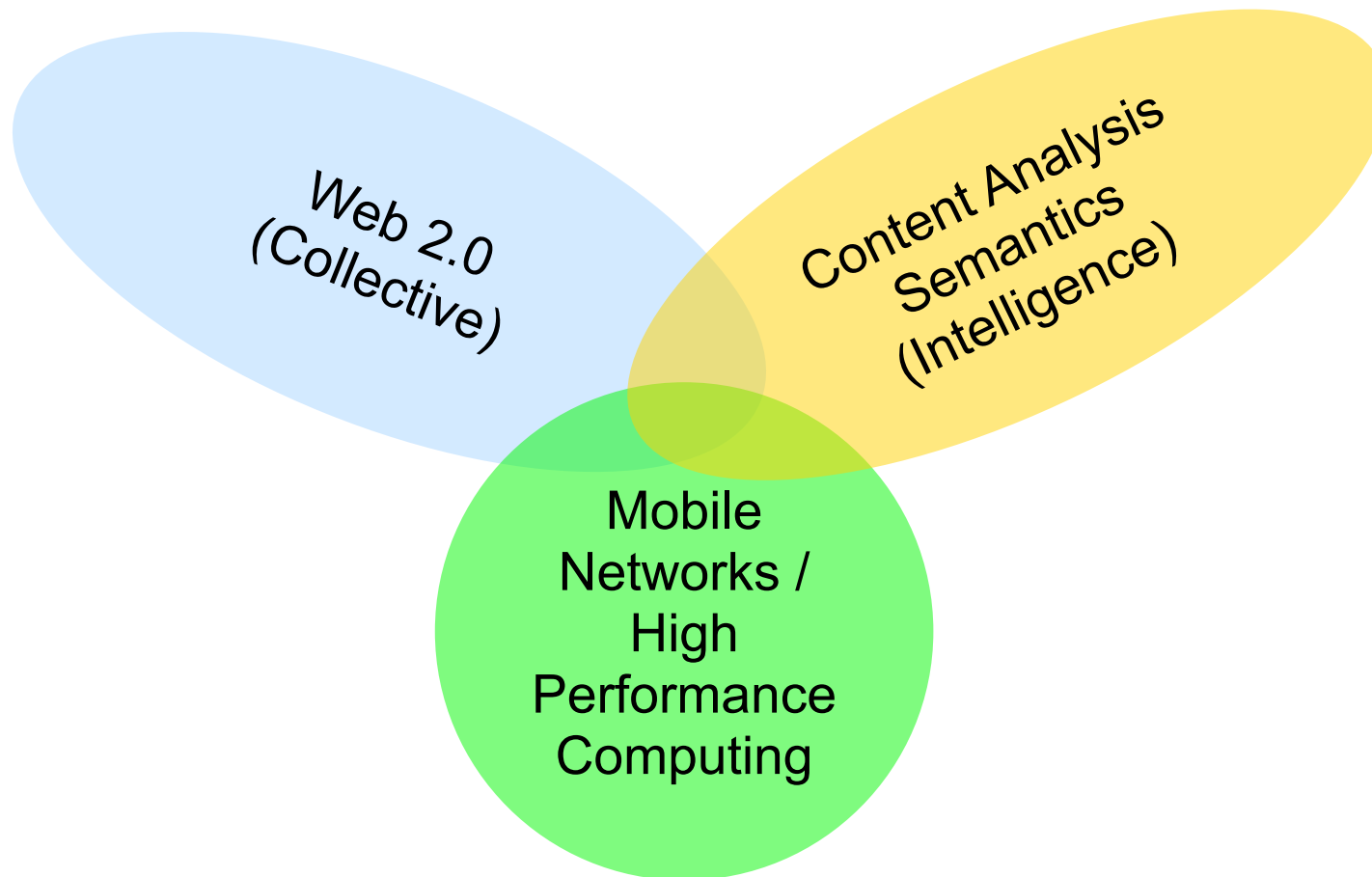
# CI and Web 2.0?

- Analyze user-generated content, such as tags that are manually assigned to photos, and its relation to context over time, space and social connectivity
- Sources
  - Tags
  - Content
  - Social info
  - Time, Location
  - Other sources (e.g. Wikipedia)



<http://www.iyouit.eu>

# Why today?



## A “simple” example

Uses the GPS in cellular phones to gather traffic information, process it, and distribute it back to the phones in real time

- online, real-time data processing
- privacy-preservation
- data efficiency, i.e. not requiring excessive cellular network



*Mobile Century Project: <http://traffic.berkeley.edu/mobilecentury.html>*



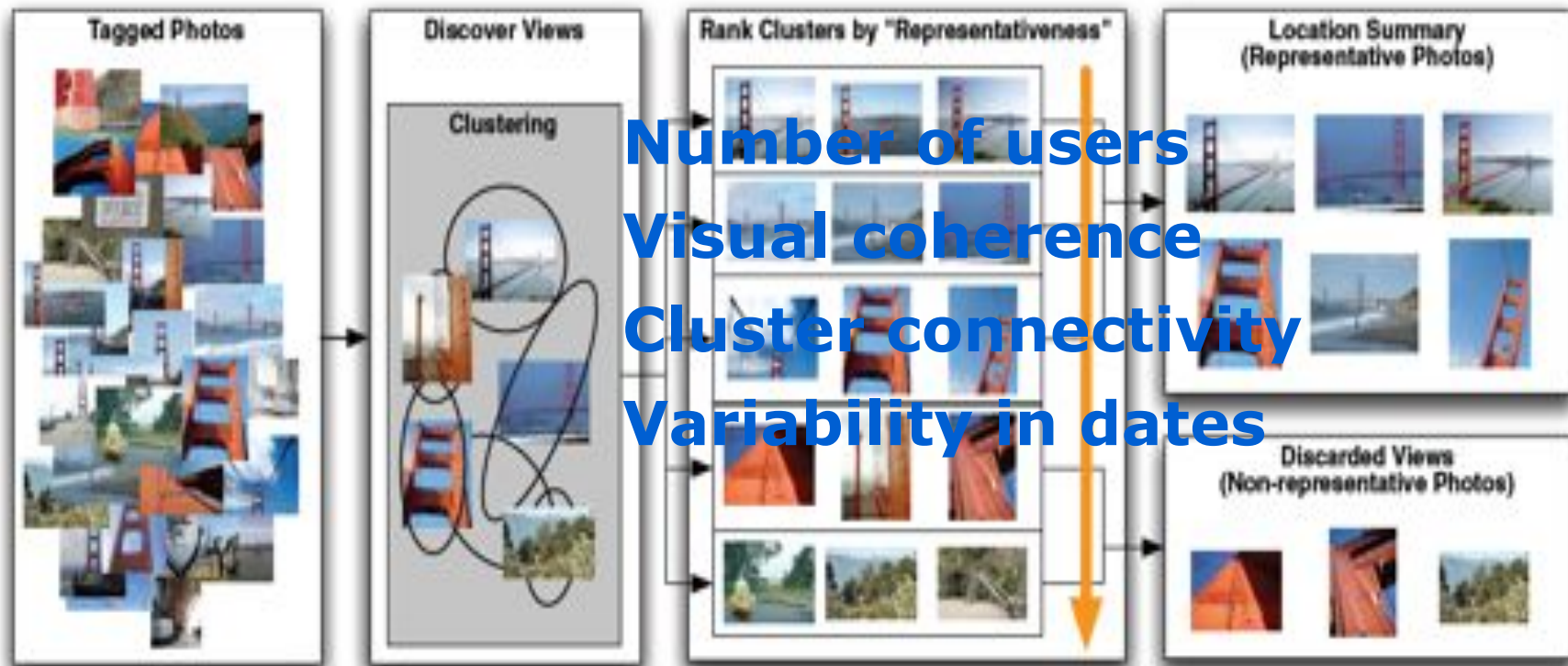
# Image search - Tourism

- Linguistic processing of semi-structured sources
  - Wikipedia, Geoplanet
- Statistical analysis for ranking
  - User Queries
  - Flickr tags



# Generating photo summaries

- **Problem formulation:** Having identified a tag x as representative of a cluster, compute a set of photos that are representative for that tag



Generating photo summaries for geographic objects in [Kennedy07]

# Sample photo summaries of events [Quacko8]

**DATASET:** Divide the earth's surface into square tiles of 200m<sup>2</sup>  
70000 geographic tiles  
220000 geotagged photos from Flickr  
After preprocessing, 73000 photos were assigned to clusters  
Manually labeling of 700 clusters



The most commonly identified event (single day covered by a single photographer)

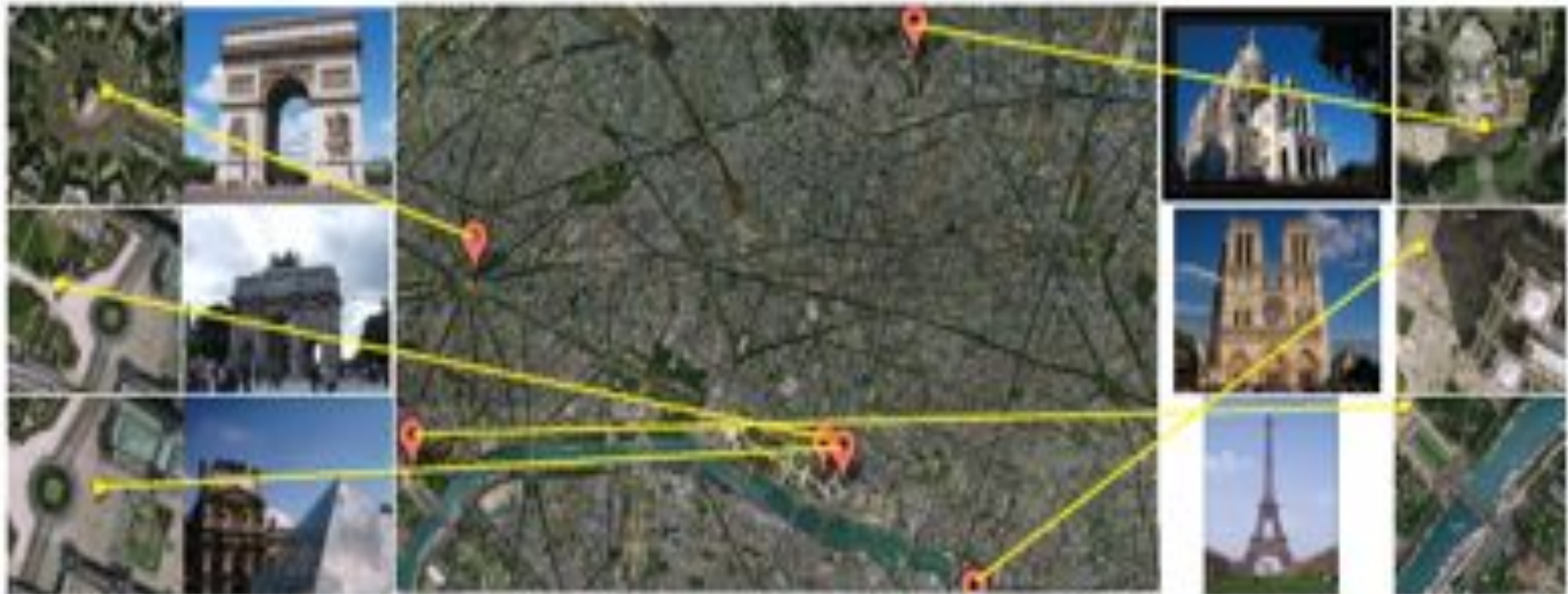
“Oxford Geek nights”

“Movie premiere Italy”

“Exhibition gallery paris”



# Auto annotation & geo-location



# Images	222'757
Size Metadata	1.1 GB
Size Features	111 GB
# Images assigned to clusters	73'236
# Similarities computed	217'330'144
# Similarities > 0	751'457

[Quack08]



# EpiCollect: Science - epidemiology example

A scientist or member of the public collects and records data, photos and videos then sends this information to a central web-based database

- e.g. to document the presence of an animal or plant species that are “representative” for a geographical area
- Location information – maps
- Citizen scientists



*EpiCollect: Linking Smartphones to Web Applications for Epidemiology, Ecology and Community Data Collection, David M. Aanensen, Derek M. Huntley, Edward J. Feil, Fada'a al-Own, Brian G. Spratt*

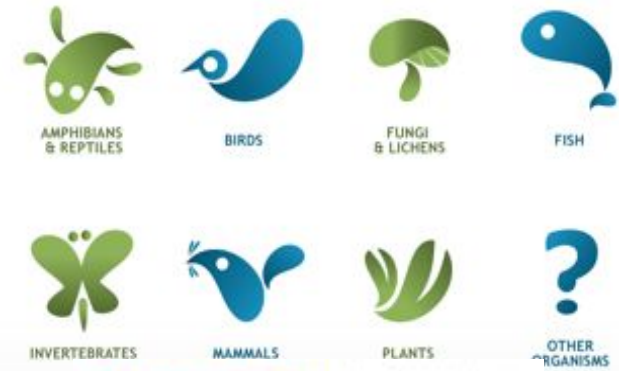
# Relevant activities



<http://traffic.berkeley.edu>



Boston Citizens Connect



Dopplr helps you share your ... travel ... and exchange tips ... presents this **collective intelligence** - the travel patterns and advice ... as the Social Atlas.

# Research Fields and Issues

- Statistical analysis, machine learning, data mining, pattern recognition, social network analysis
- Clustering
- Graph theory
- Image, text, video analysis
- Information extraction
- Fusion techniques
- Trust, security, privacy
- Performance, scalability
  - speed, storage, power, grids, clouds

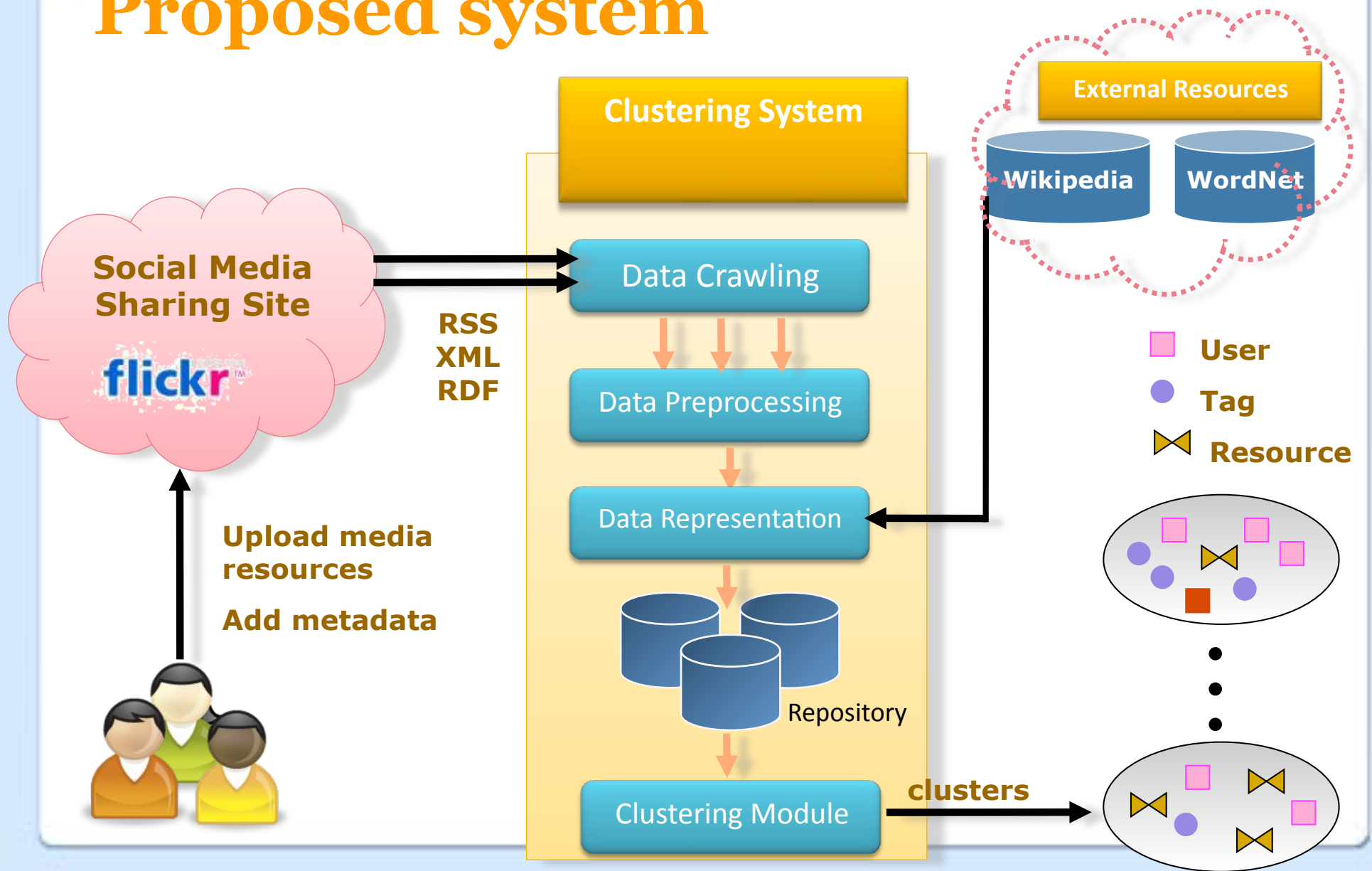
# Clustering for Social Media



# Clustering Approaches

- Tag-Based
- Content-Based
- Time-based

# Proposed system



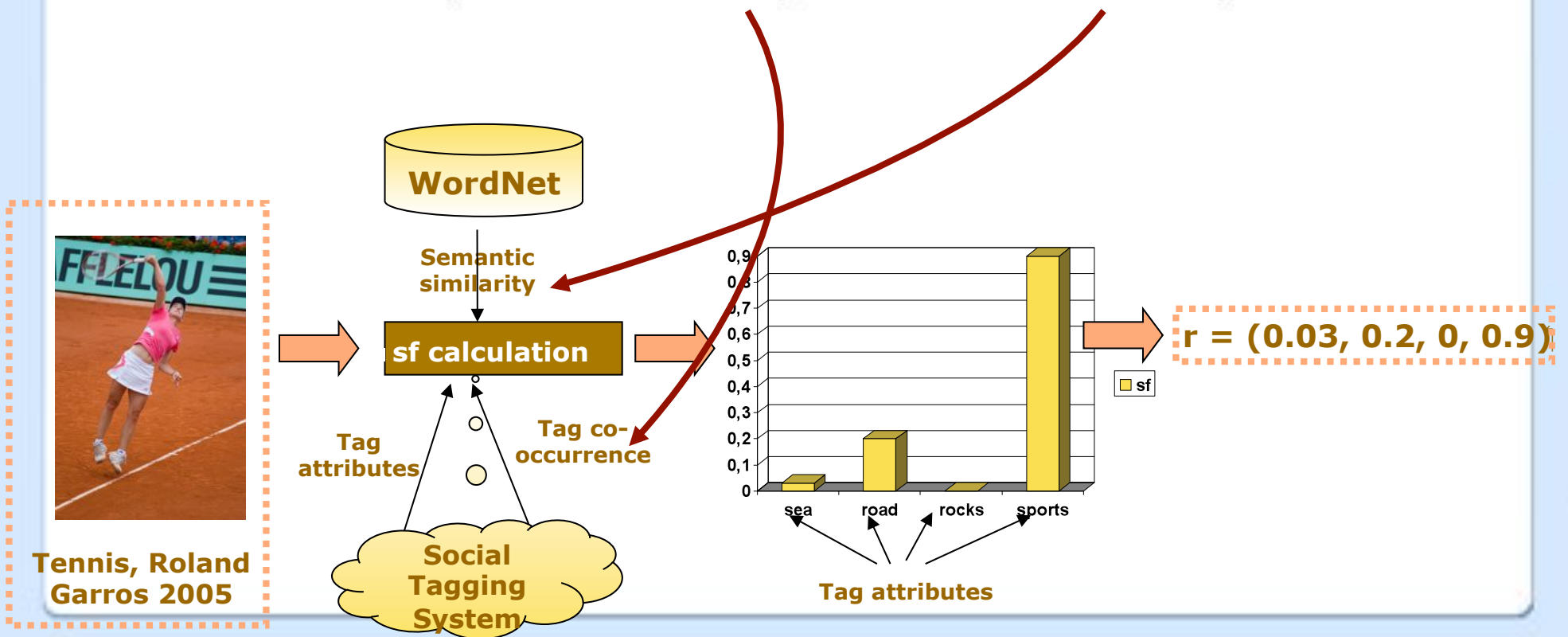
# Tag-based Clustering (I)

- **1. Vector data model**
- Assume **n** resources and **d** attribute-tags
  - **d**: a representative set of tags
- A resource representation in vector space (**sf**) is based on **semantic similarity** and **tag co-occurrence** between the resource's tags and the attribute-tags
- A resource **r<sub>i</sub>** is represented by a **d**-dimensional vector **r<sub>i</sub> = (sf<sub>1</sub>, sf<sub>2</sub>, ..., sf<sub>d</sub>)**
- All resources can be represented by an **n x d** matrix

# Tag-based Clustering (II)

- 2. Clustering on  $n$  (resources,  $r$ )  $\times$   $d$  (attributes) matrix (K-means, Hierarchical, COBWEB)

$$SS(t_x, t_y) = w * SoS(t_x, t_y) + (1 - w) * SeS(t_x, t_y)$$



Tennis, Roland Garros 2005

# Tag-based Clustering - Experimental Results

- **Dataset:** 3000 images downloaded from Flickr

- Meaningful subdomains of **roadside:**

**buildings, roof, street, road**



(a)

**cars, vehicles, race**



(b)

**people, street, festival**



(c)

- Different clusters for the **ambiguous tag** *wave, rock*:

**wave, sea, ocean**



(a)

**wave, person, hand**



(b)

**rocks, stone, rockside**



**rock, music, band**





# Tag & Content-based Clustering

- After performing tag-based clustering, low-level features of resources are used for cluster refinement
- Outlier Detection (mahalanobis distance)
- For each resource the following visual descriptors are extracted:
  - Scalable Color, *SC*
  - Color Structure, *CS*
  - Color Layout, *CL*
  - Edge Histogram, *EH*
  - Homogenous Texture, *HT*
- A single image feature vector per each resource is produced, encompassing all descriptors normalized in  $[0,1]$
- Feature extraction and distances between image feature vectors are according to MPEG-7 XM.

# Evaluation Method

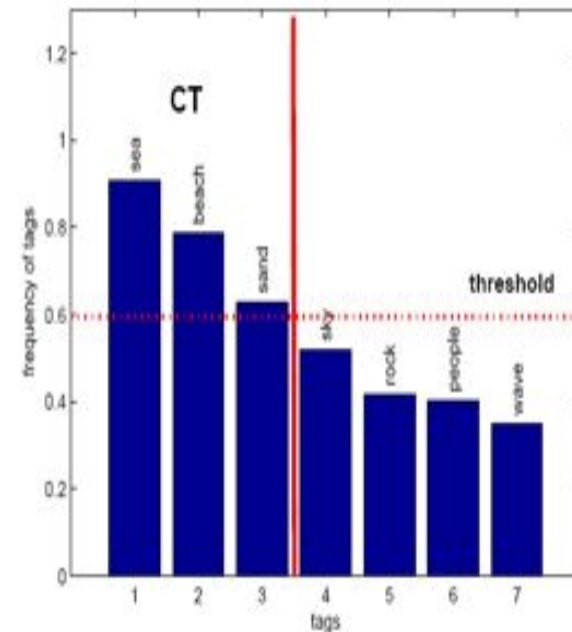
- **Definition:** Cluster Topic, CT, are the tags that have frequency in cluster's resources annotation over a threshold T.

- **Evaluation Metrics**

- Precision  $Pr(C_j) = \frac{|C_j \cap RR(C_j)|}{|C_j|}$

- Recall  $R(C_j) = \frac{|RR(C_j) \cap C_j|}{|RR(C_j)|}$

- F-Measure  $F(C_j) = \frac{2 * Pr(C_j) * R(C_j)}{Pr(C_j) + R(C_j)}$

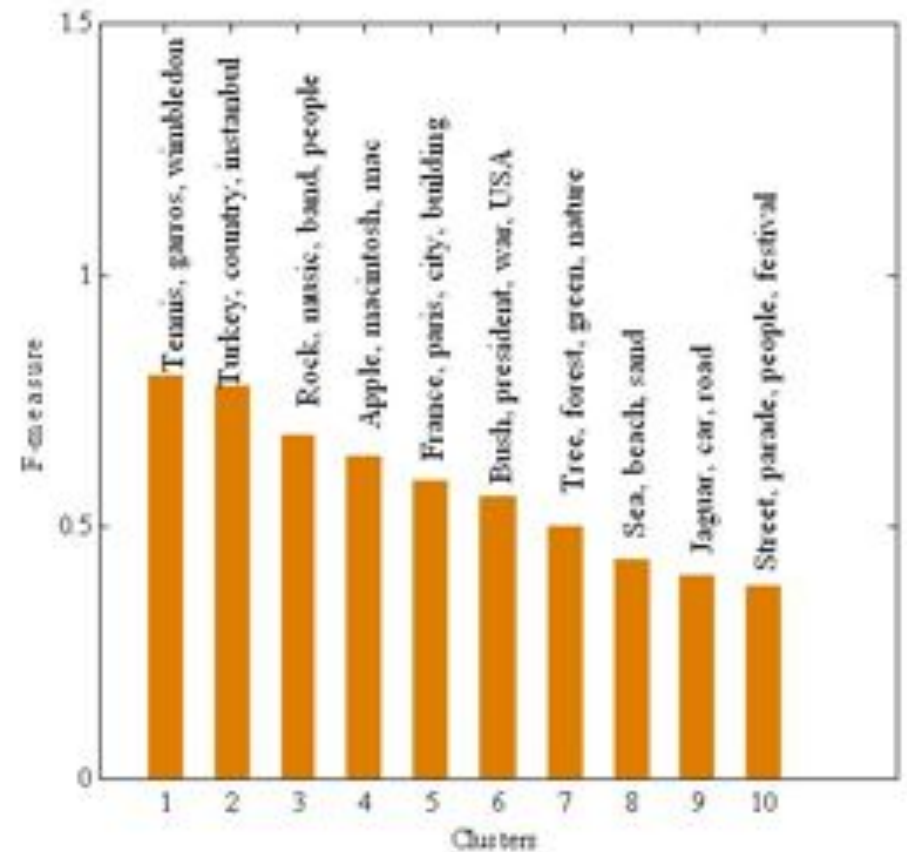
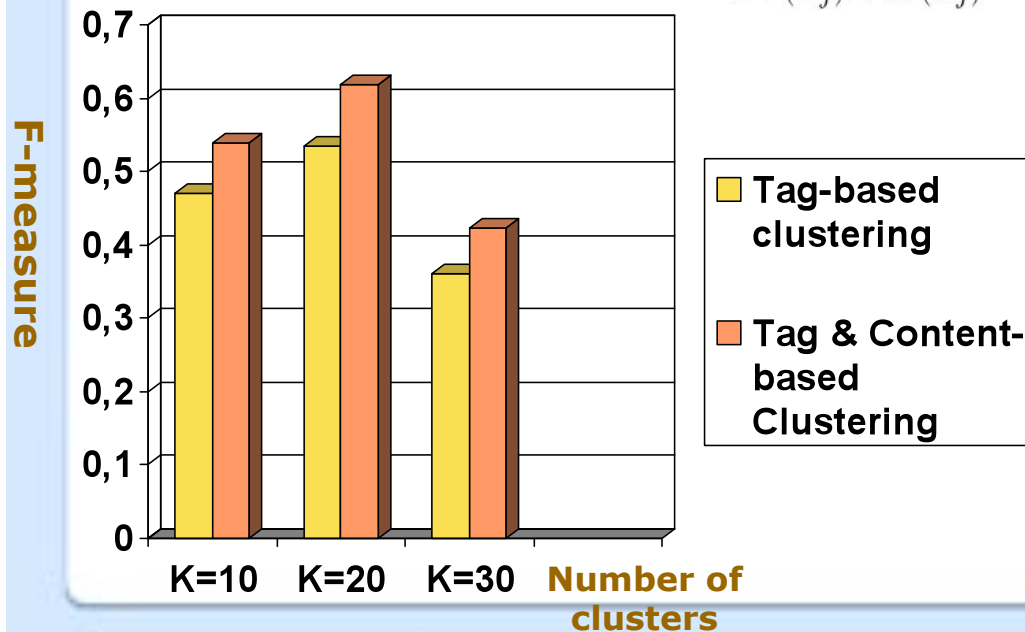


# Tag & Content-based Clustering – Experimental Results

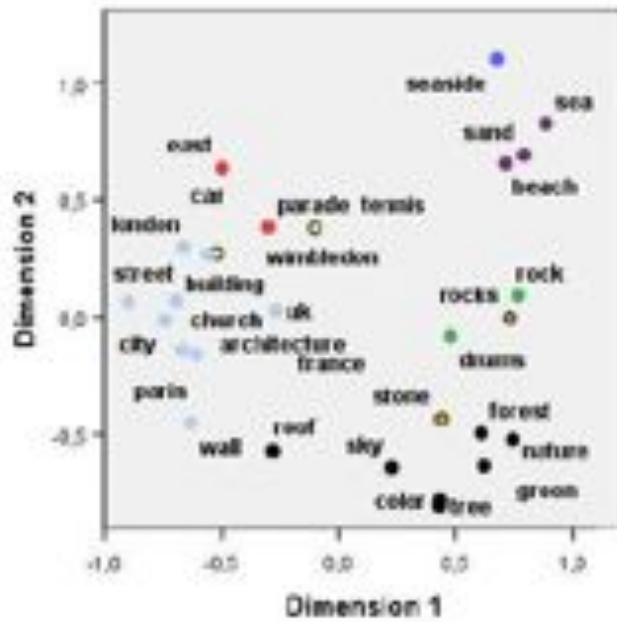
**Dataset:** 10000 images (with their tags) downloaded from Flickr

**Evaluation:** Manual annotation and use of F-Measure.

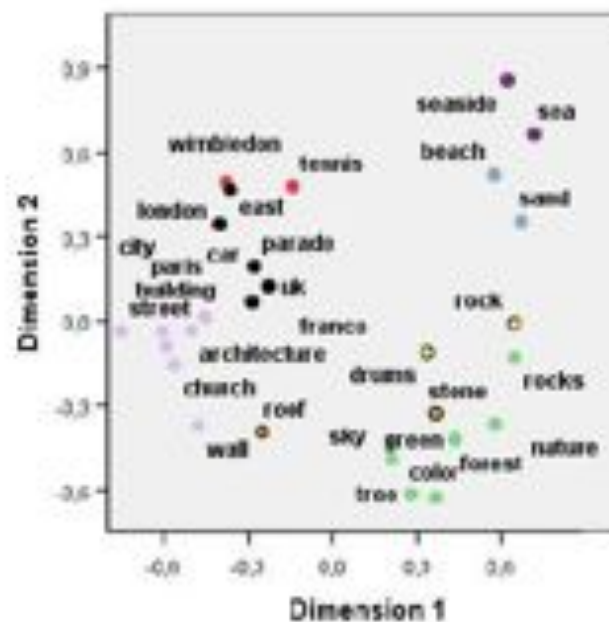
$$F(C_j) = \frac{2 * Pr(C_j) * R(C_j)}{Pr(C_j) + R(C_j)}$$



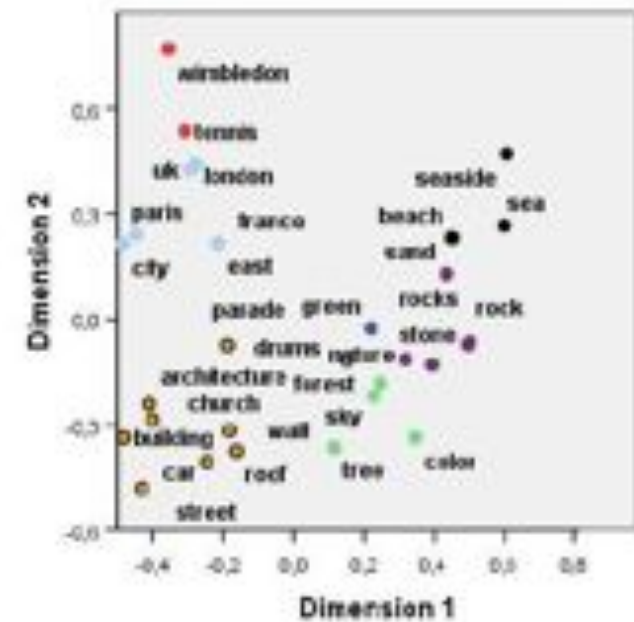
# Experimental Results (II)



(a)  $w = 0.2$



(b)  $w = 0.5$



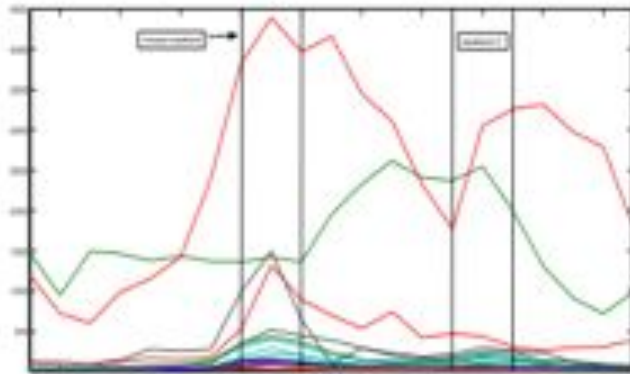
(c)  $w = 0.8$

Attributes Assignment to  $k=8$  clusters,

$w$ : weighting factor of semantic similarity against similarity derived from tag co-occurrence

# Why consider time?

- Most approaches analysis of “static” views of users-tags
- Events, Trends change user interests
- Users Tagging Behavior changes over time
- Time is a fundamental dimension in analysis of users and tags in a social tagging system

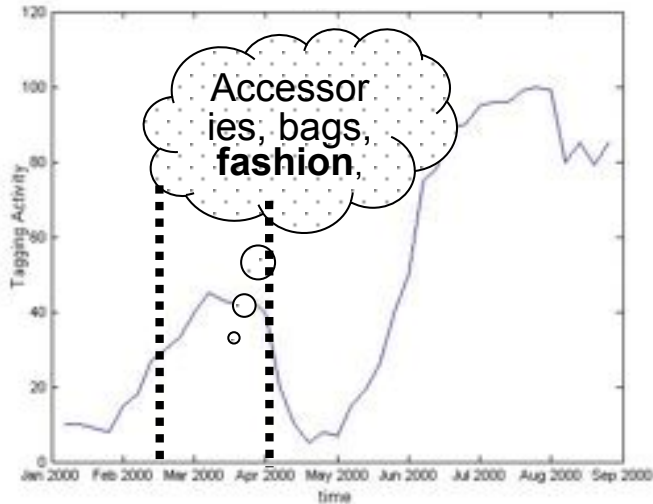


E.g. : Prediction of first weekend box-office revenues using tweets



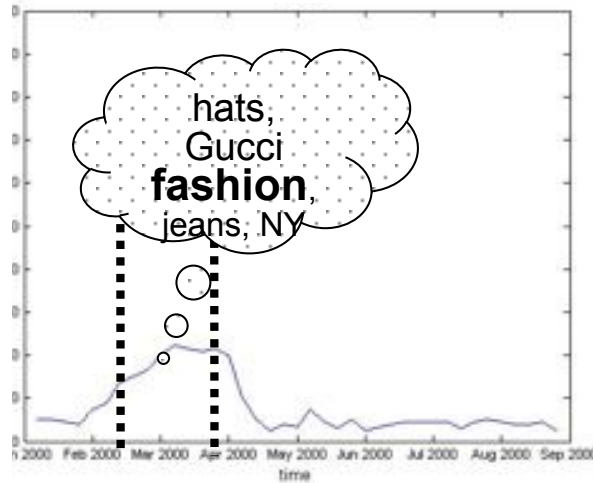
# Many times, a user's targeted interest is hidden in the general tagging activity...

User 1



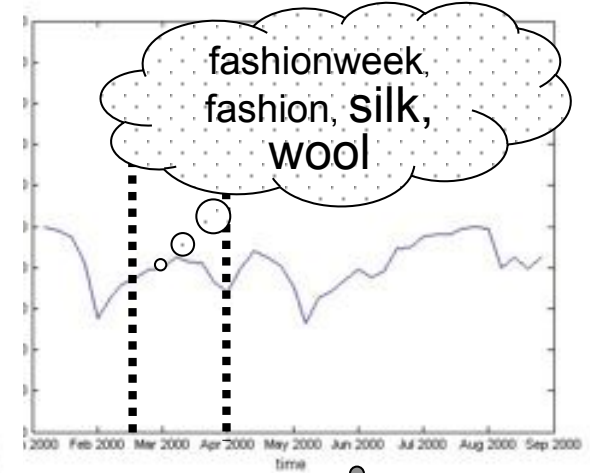
Cars, football, holidays, horses, sea, turkey, fashion

User 2



New York, hat, trousers, fashion, Gucci

User 3



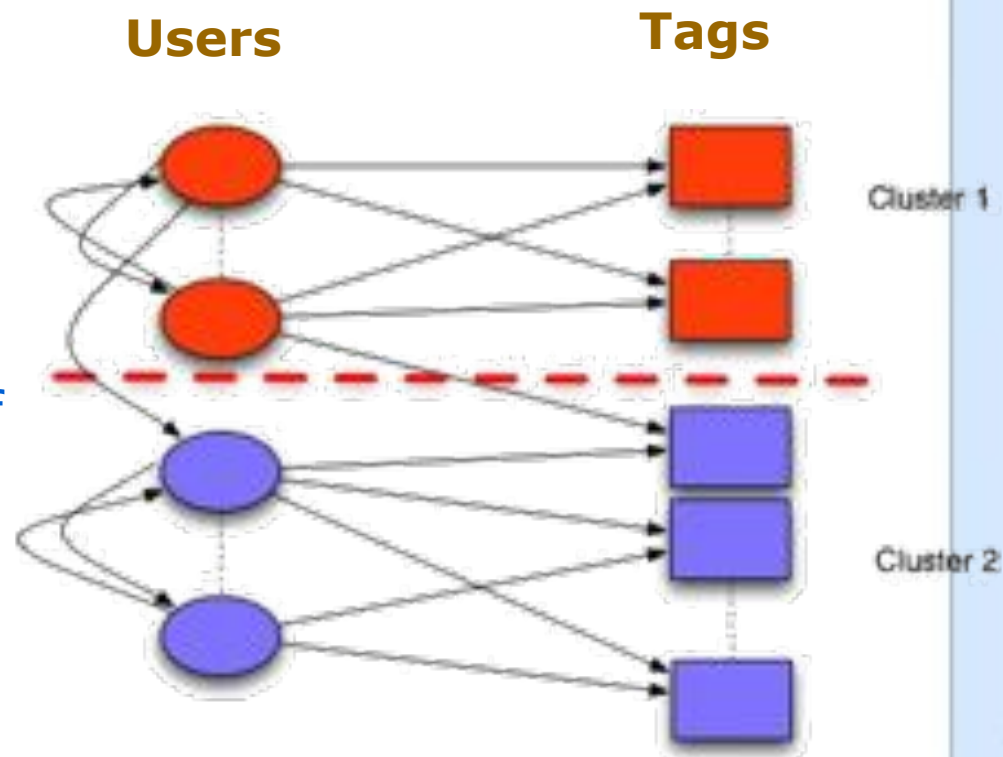
animals, elephants, nature sea, turkey, bags

# Time-aware user/tag clustering

Static user/tag clusters	Time-aware user/tag clusters
Find user/tags groups that relate to a topic	Find user/tags groups that relate to a topic <b>at specific time periods</b> (e.g. people interested in fashion every August and March, that new collections are announced)
Group together users that use similar tags during the entire time span	Discriminate between users' regular interests (spread over the entire time span) and occasional interests (highlighted in specific time periods)

# Clustering vs Co-clustering

- Given a multi-dimensional data matrix, co-clustering refers to **simultaneous** clustering along multiple dimensions
- In a two-dimensional case it is simultaneous clustering of rows and columns
- Most traditional clustering algorithms cluster along a single dimension



# Co-Clustering Example

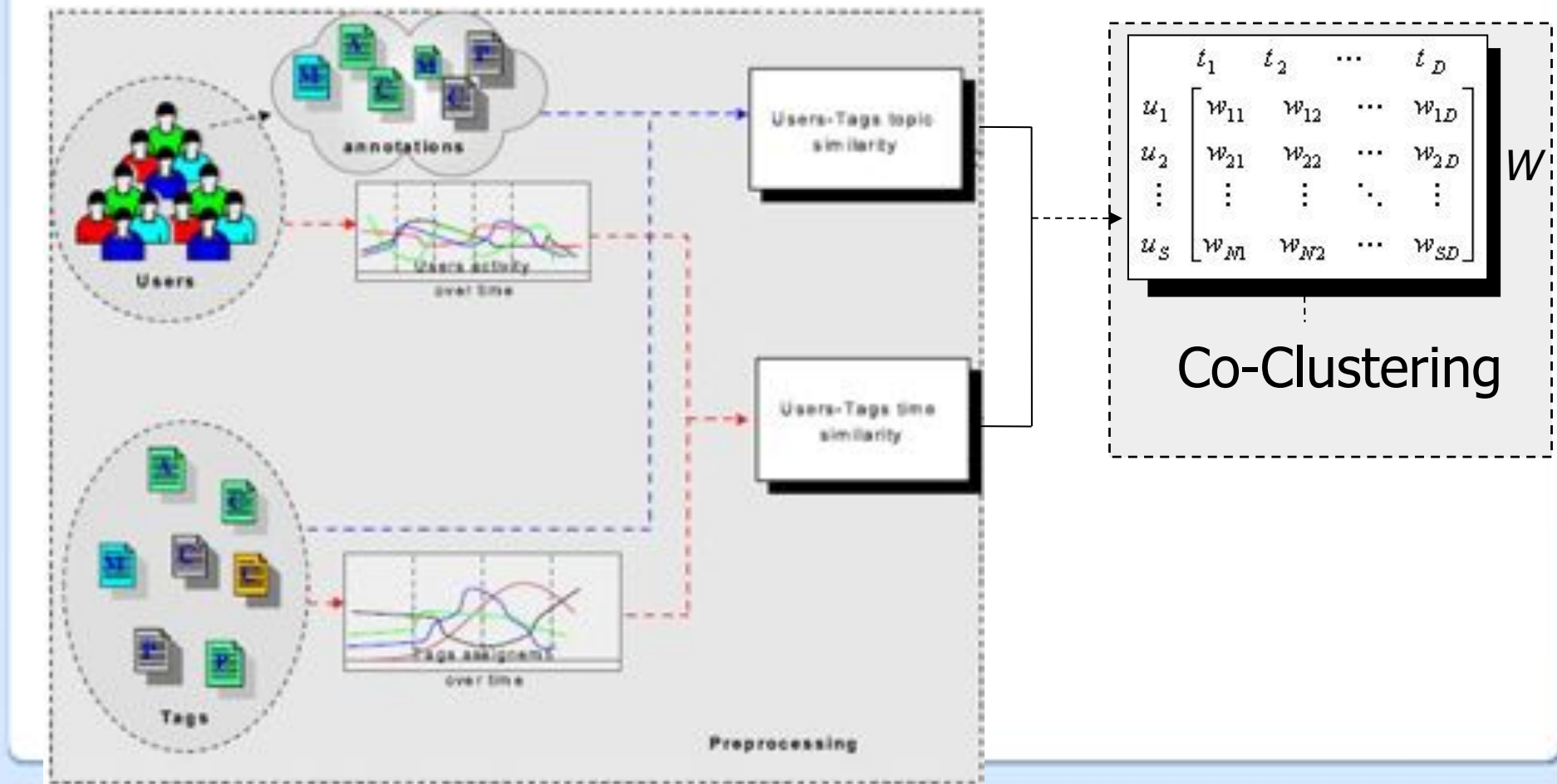
- Text is represented as a Matrix  $D$ 
  - rows denote documents
  - columns denote the words
  - matrix elements  $D_{ij}$  denote occurrence of word  $j$  in document  $i$
- Co-clustering is applied to discover blocks in matrix  $D$ 
  - correspond to a group of documents (rows) characterized by a group of words (columns)
- In our case we want a user – tag time dependent matrix:  $D$  (user, tag)

# The proposed approach (I): Overview

- Build a matrix of user activity over time:  $U(\text{user}, \text{time})$
- Build a matrix of tag activity over time:  $T(\text{tag}, \text{time})$
- Combine these two matrixes:  $TeS(\text{user}, \text{tag})$ 
  - Temporal connection is introduced, but
  - Tag – User connection is not taken into account
- Build a matrix of users and tags based on tag semantics  $SeS(\text{user}, \text{tag})$
- Combine  $TeS$  and  $SeS$  and apply co-clustering:  $Sim(\text{user}, \text{tag})$ 
  - Both temporal and tag – user connections are introduced



# The proposed approach (II): The Co-Clustering algorithm



# The proposed approach (III): Details

## Focus on time locality

- Division of total time in **timeframes** of size  $\tau$
- Representation of users and tags activity in each timeframe (**vector model**)
  - Number of tags a **user** has assigned and number of times a **tag** has been used, during each timeframe
- Combination of the two matrixes: Inner product

## Focus on tag – user similarity

- Compute similarities between users and tags based on **tag semantics**
- Similarity metric: WordNet *Wu & Palmer*

## Joint use of tag and time similarity

- Similarity metric: Dot product  $\rightarrow (u_i, t_j) = SemSim(u_i, t_j) * TemSim(u_i, t_j)$

[I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *7th SIGKDD*]

# Experimentation – Input parameters

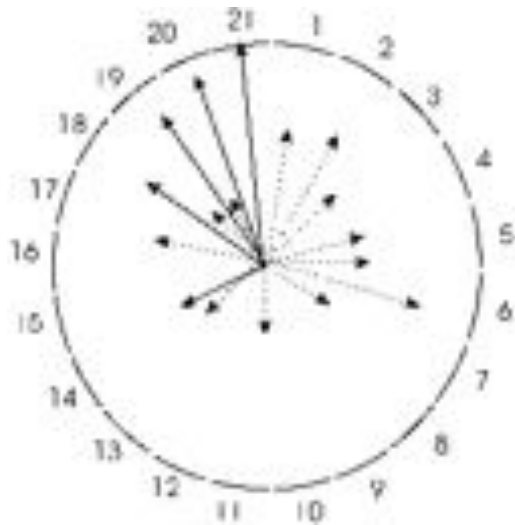
- **Data workload:** 6764 images from Flickr over a time period of 1 year (Sept 2007-Aug 2008) that referred to 4 topics (ancient Greece, Olympics, earthquake and weddings)
  - Pre-processing
    - Remove invalid tags
    - Remove tags with frequency  $< 1$
    - Keep compound valid tags
      - 1218 users, 2496 tags, 210 days
- **Size of timeframes:**  $\tau = 1, 10, 30$  days
- **Number of clusters:**  $k = 7, 10, 12$

# Experimentation – the $\tau$ parameter

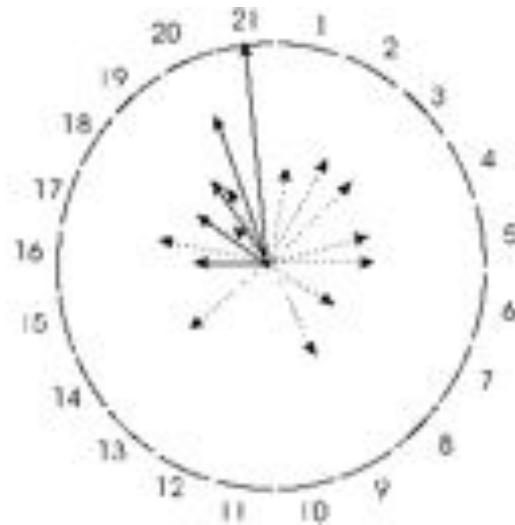
- By changing the value of  $\tau$ , we can discriminate between users **occasional** and **regular** interests

User	$\tau = 30$	$\tau = 100$
User1	olympics2008, beijing, flame, opening ceremony	ancientgreece, acropolis, parthenon, archaeology, ancient-civilizations
User2	earthquake, china, disaster, ruin, disasterassistanceresponseteam	wedding organizing, party

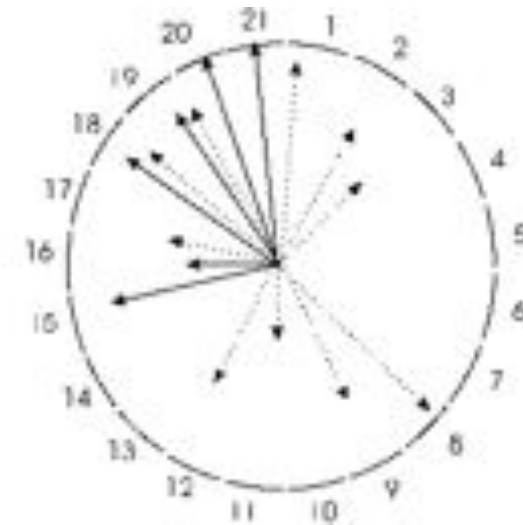
# Experimentation – Clusters Visualization



Tags distribution in a cluster



User1' s tags distribution

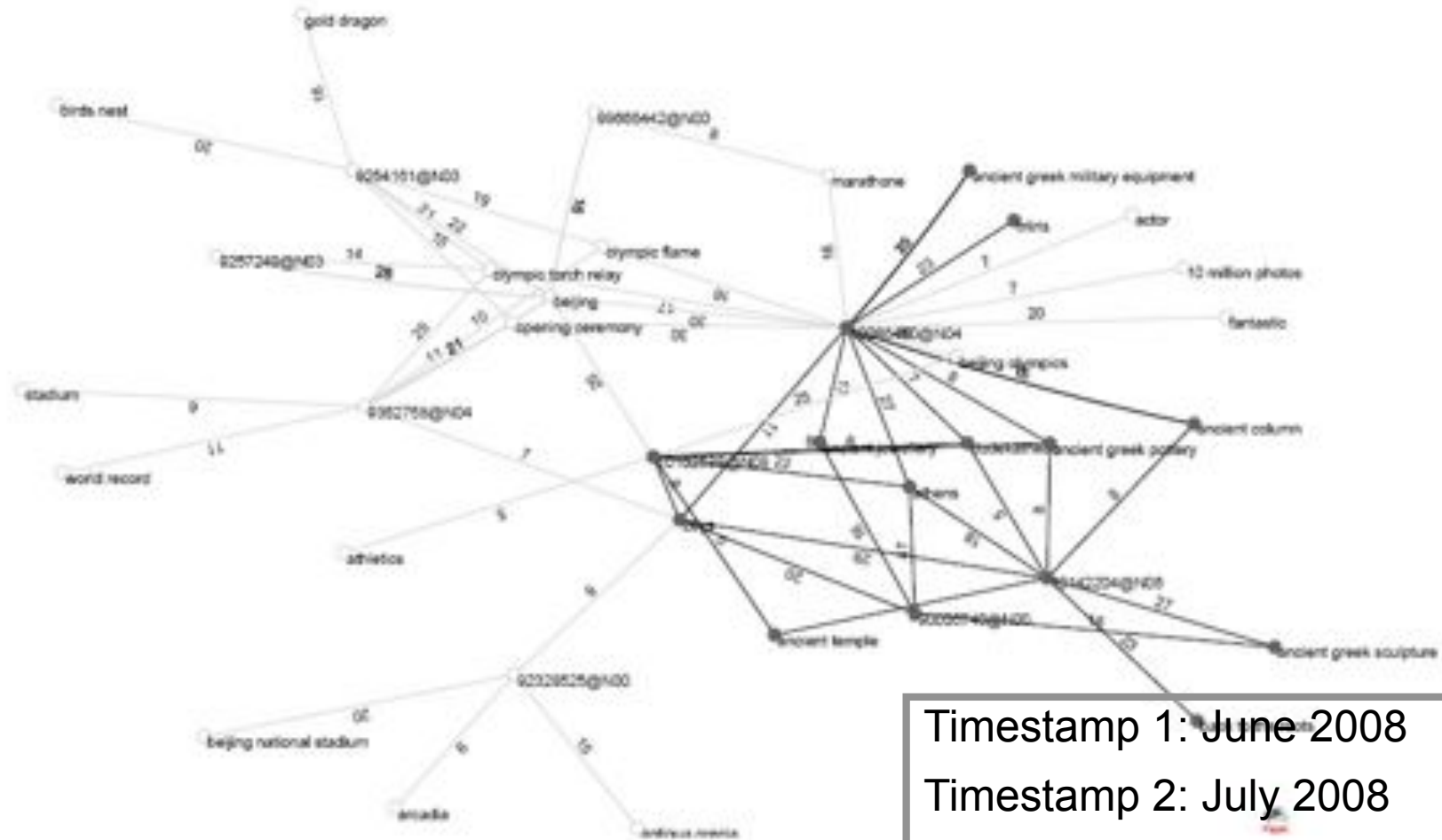


User2' s tags distribution

- Olympics –related tags
- .....→ Ancient Greece –related tags



# Cluster evolution (timestamp 1)







# Use Cases

- Capturing trends, interests, periodic activities of users in specific time periods
- Community-based tag recommendation
- Personalization (time-aware user profiles)
- Fighting spam on social web sites (by discriminating regular and occasional users)

# **Social Media “teacher” of the machine**

# Exploiting clustering for machine learning

*Objective: Develop a framework able to create strongly annotated training samples from weakly annotated images*

## Tagged images



sand, wave, rock, sky



sea, sand



sand, sky

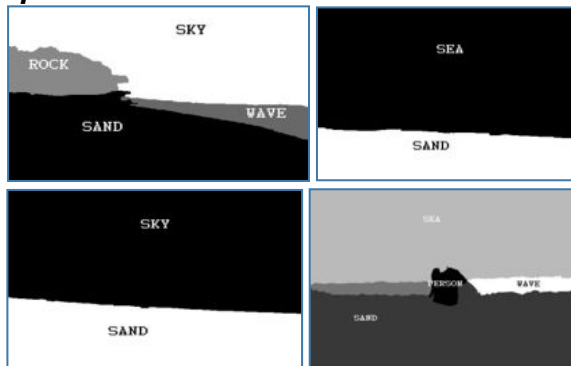


person, sand, wave, sea

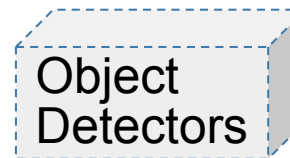
Social information +

Image analysis

## Region-detail annotated



Machine Learning



## Problems:

- ❖ Object detection schemes require region-detail annotations
- ❖ Manual annotation is laborious and time consuming

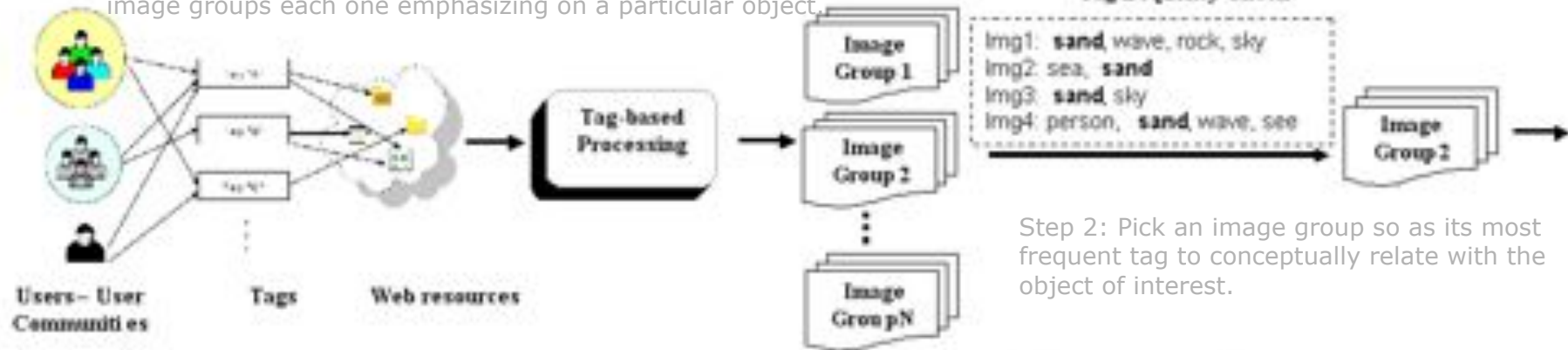
[Chatzilari09]

## Solutions:

- ❖ Exploit user tagged images from social sites like flickr
- ❖ Combine techniques operating on tag and visual information space

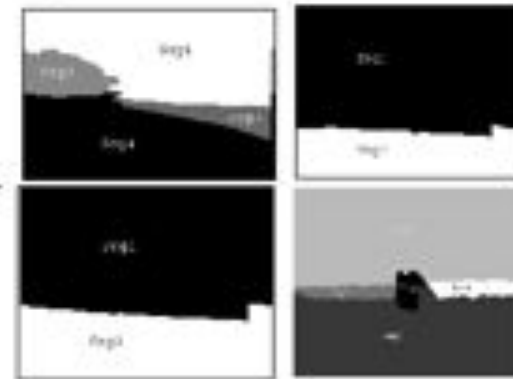


Step 1: Process image tag information in order to acquire image groups each one emphasizing on a particular object.



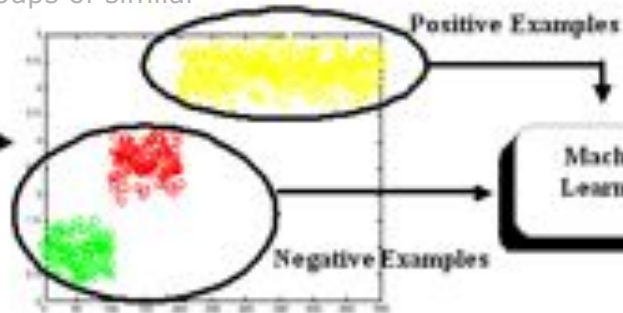
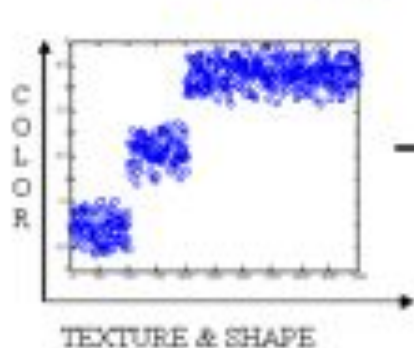
Step 2: Pick an image group so as its most frequent tag to conceptually relate with the object of interest.

Step 3: Segment all images in the selected image group into regions.

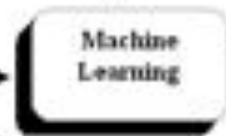


Step 4: Extract the visual features of these regions.

Step 5: Perform feature-based clustering so as to create groups of similar regions



Step 6: Use the visual features extracted from the regions belonging to the most populated cluster, to train a machine learning-based object detector.



# Tag-based processing

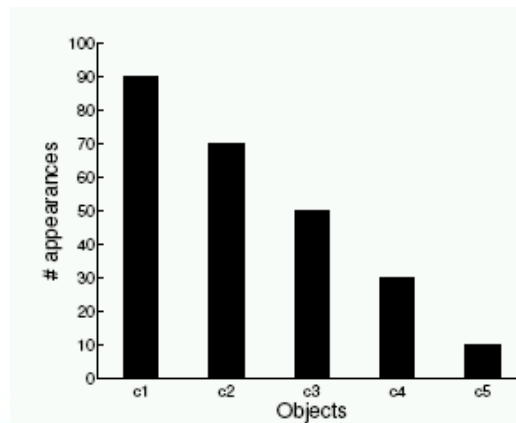
[Giannakidou08]

**SEMSOC**, vector space model where each image is projected onto a space defined by the most prominent tags

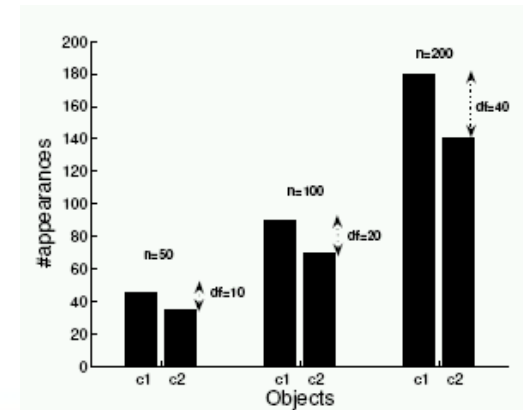
SEMSOC output example



Distribution of objects based on their frequency rank



Absolute difference between 1<sup>st</sup> and 2<sup>nd</sup> most highly ranked objects increases as n increases



# Segmentation & Visual Descriptors

- Segmentation

- K-means with connectivity constraint (KMCC)

*[Mezaris et al., 2004]*

- Visual Descriptors

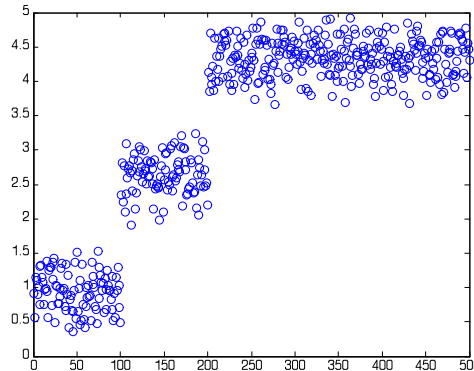
- MPEG-7 standard

- *Dominant Color , Color Layout, Color Structure, Scalable Color, Edge Histogram, Homogeneous Texture, Region Shape.*

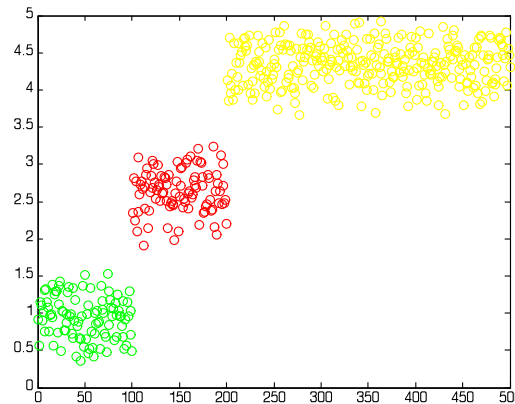
*[Bober et al., 2001], [Manjunath et al., 2001].*

# Region-based Clustering & Cluster Selection

## Region clustering

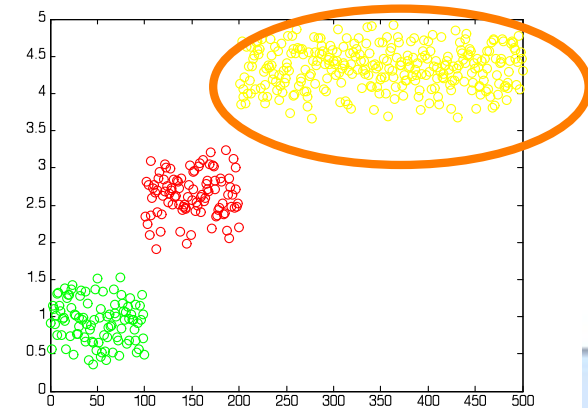


- ❖ Perform segmentation and visual feature extraction from all images in an image group (Identified by SEMSOC)

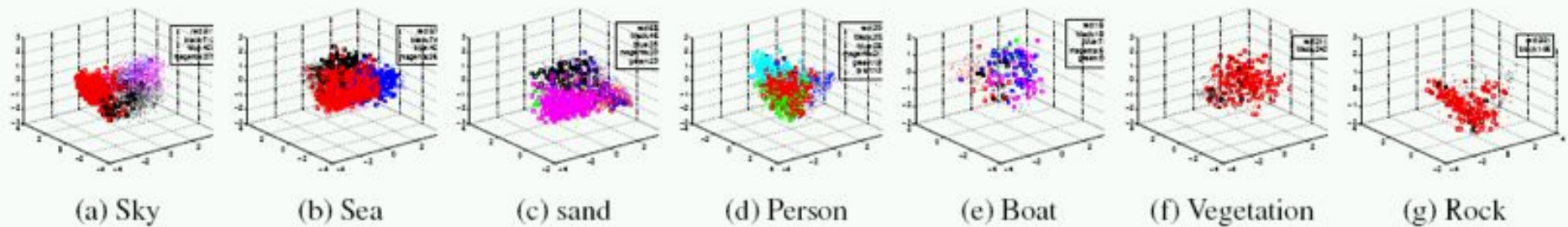


- ❖ Perform clustering based on visual features to gather together regions depicting the same object

- ❖ Pick the most populated cluster as the one representing the most frequently appearing tag of the group



# Experimental Results – Cluster Selection



## **Setting:**

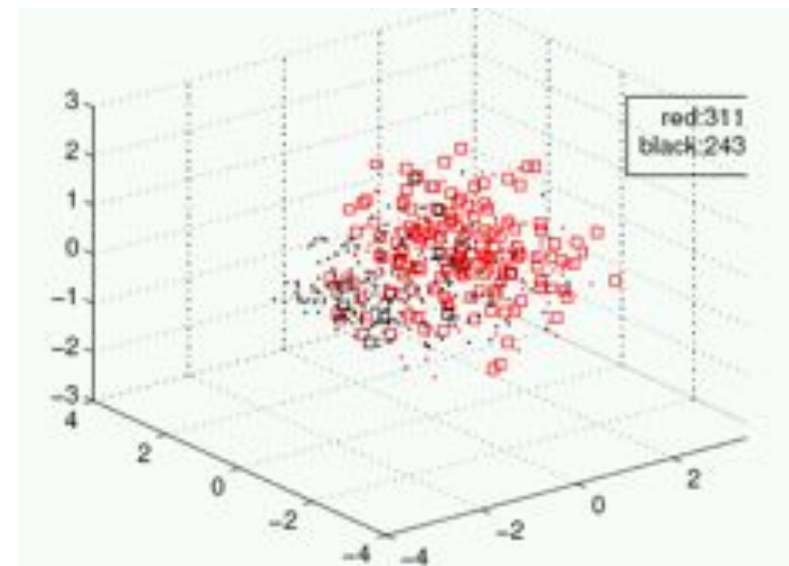
- Visualise the way regions are distributed among clusters
- Use shape-code (squares) to indicate the regions of interest and color-code to indicate a cluster's rank (largest cluster: red)
- Ideally all squares should be painted red and all dots should be painted differently

## **Goal:**

- Validate our theoretical claim that the most populated cluster contains the majority of regions depicting the object of interest

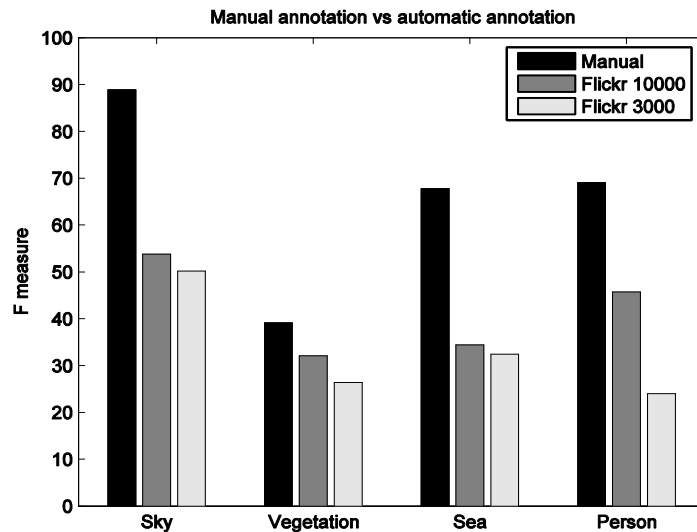
## **Conclusions:**

- Our claim is valid in 5 (i.e., sky, sea, person, vegetation, rock) and not valid in 2 (i.e., boat, sand) cases



**Vegetation in magnification**

# Experimental Results - Man. vs Autom. trained object detectors



## Observations:

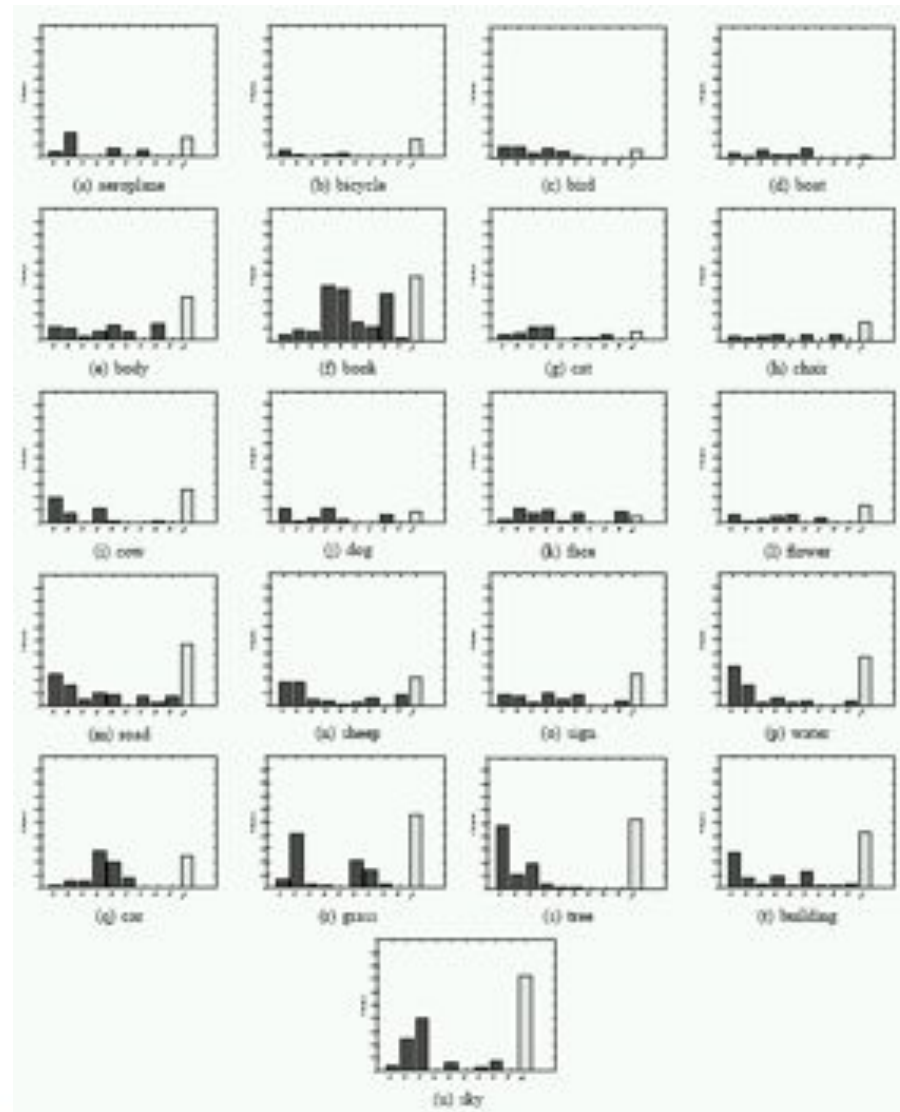
- Performance lower than manually trained detectors
- Consistent performance improvement as the dataset size increases



# Experimental Results – MSRC Dataset (21 objects)

## Observations:

- In 5 cases the objects were too diversiform to be described by the employed feature space (not even the manual annotations performed well)
- In 5 cases the annotation we got from Flickr groups were not appropriate
- In 6 cases, our method has failed to select the appropriate cluster
- In 5 cases our method worked well



# Experimental Results - MSRC vs Flickr groups

## Target object: Tree

Tree object



*Good example: Semantic objects are correctly assigned to clusters and the most-populated cluster corresponds to the target object)*

# Experimental Results - MSRC vs Flickr groups

## Target Object: Sky

Sky object

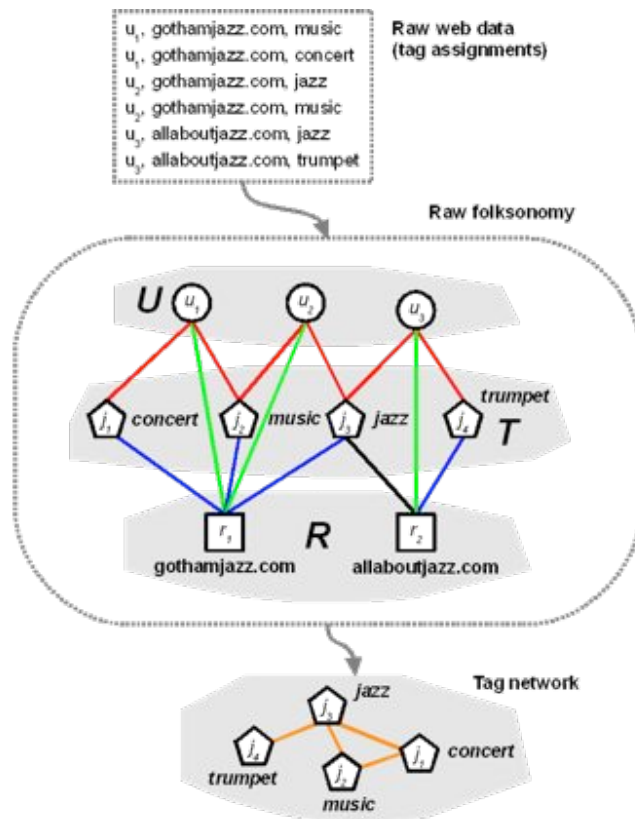


*Bad example: Sky regions are split in many clusters and the most populated cluster contains noise regions*

# Community Detection

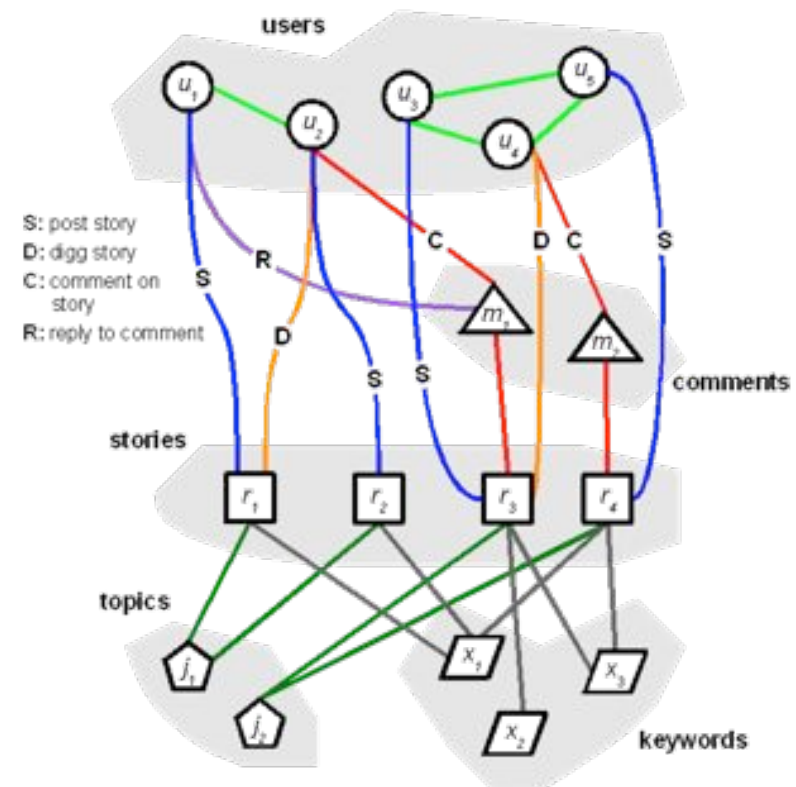
# Examples of Social Media networks

Folksonomy (Delicious)



Mika, P. (2005) Ontologies Are Us: A Unified Model of Social Networks and Semantics. Proceedings of the 4th International Semantic Web Conference (ISWC 2005), Springer Berlin / Heidelberg, pp. 522-536

MetaGraph (Digg)



Lin, Y., Sun, J., Castro, P., Konuru, R., Sundaram, H., and Kelliher, A. (2009) MetaFac: community discovery via relational hypergraph factorization. Proceedings of KDD '09, ACM, pp. 527-536



# Challenges in Social Media network mining

No prior assumptions about structure:

Complex & evolving structure

No possibility for knowing structural features (e.g. number of clusters on a graph) in advance

→ Unsupervised

Scale

Tens of millions of active users frequently contributing loads of content links + metadata (tags, comments, ratings)

→ Efficient - scalable

Quality

Spam is very common. Only a portion of user contributions is worth further analysis.

→ Noise resilient



# What is a community in a network?

Group of vertices that are more densely connected to each other than to the rest of the network.

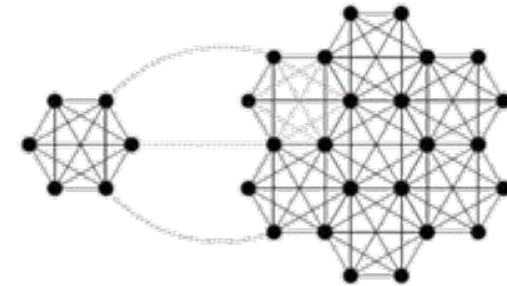
Multiple definitions to quantify communities:

Fortunato S. (2010) Community detection in graphs. Physics Reports 486: 75-174

Global: N-cut, conductance, modularity

Local: Local modularity,  $(\mu, \epsilon)$ -cores

Ad hoc: Label propagation, dynamic synchronization



Related to clustering, but: (a) not necessary to know number of communities, (b) computationally more efficient

In Social Media, we focus on local definitions, because of the properties of Social Media networks: efficiency-scalability and noise resilience.

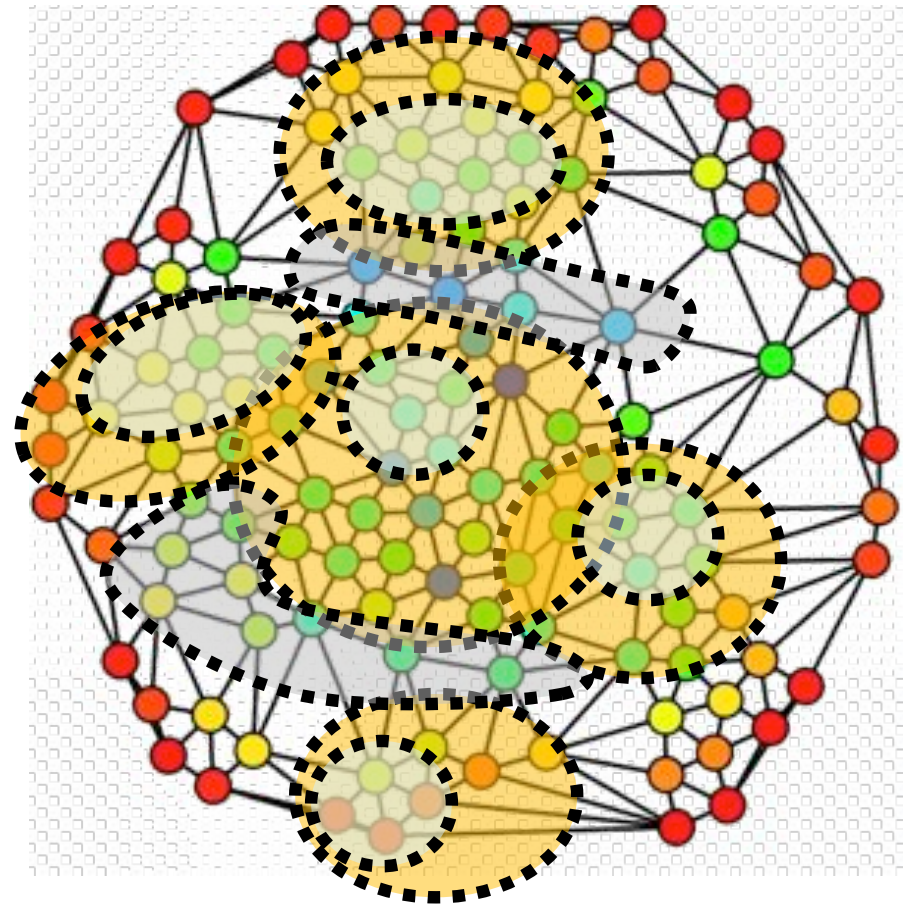
# Global vs. Local

- **Global:** Process the whole graph to derive a partition into communities
  - + Abundant research
  - + Good results (community quality, algorithm efficiency)
    - Not practical for huge graphs or for real-time applications
- **Local:** Incremental process of the graph and output communities (streaming)
  - Relatively little research
  - Great potential for demanding applications

# Approach illustration

Two-step process:

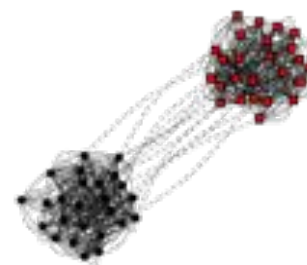
- 1<sup>st</sup> step:  
 $(\mu, \varepsilon)$  – core detection
- 2<sup>nd</sup> step:  
Local expansion
- 3<sup>rd</sup> step:  
Characterization of remaining vertices as *hubs* or *outliers*



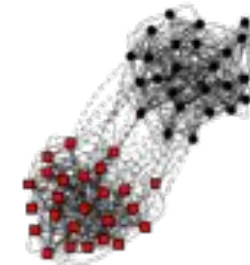
# Experiments on Synthetic Community Networks

- Synthetic networks according to method of Newman and Girvan.

$$S_{PAR} = \{N, K, z_{tot}, p_{out}, s_{var}\}$$



(a)  $p_{out} = 0.01$



(b)  $p_{out} = 0.08$

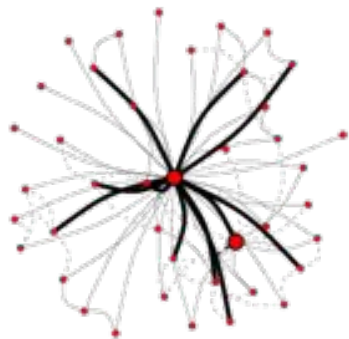
Change complexity of underlying communities.

	$F_C$			NMI		
$p_{out}$	BB	BB'	GN	BB	BB'	GN
0.01	100	100	100	1.0	1.0	1.0
0.05	100	100	100	1.0	1.0	1.0
0.1	100	100	50	1.0	1.0	0.86
0.15	100	99	50	1.0	.98	0.86
0.20	99	74	50	0.98	0.84	0.86
0.25	24	24	0	0.54	0.56	0.02

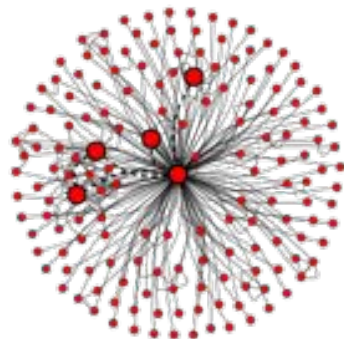
Change relative sizes of underlying communities.

	$F_C$			NMI		
$s_{var}$	BB	BB'	GN	BB	BB'	GN
1.1	100	100	100	1.0	1.0	1.0
1.5	100	100	100	1.0	1.0	1.0
1.6	99.5	100	100	0.99	1.0	1.0
1.7	88	98	100	0.82	0.96	1.0
1.8	85.5	97	100	0.79	0.95	1.0
1.9	58.5	87	90	0.68	0.82	0.88
2.0	12.5	80	82	0.45	0.73	0.81
2.5	0	62	75	0.45	0.63	0.72

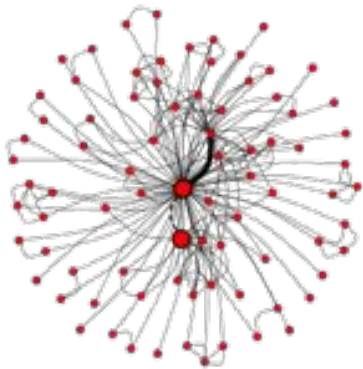
# LYCOS iQ Tag Network



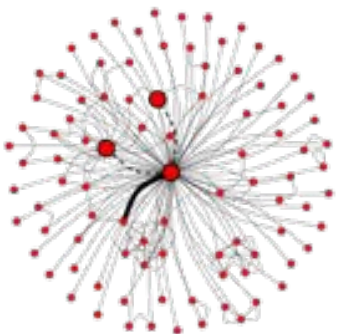
(a) Music



(b) Science



(c) Film



(d) Animals



**Computers:**  
A densely interconnected community

**History:**  
A star-shaped community



# Hybrid Photo Clustering

## Goal:

- Group large photo collections into clusters based on how much they are related to each other
- Assist browsing and navigation by means of a map-based application
- Detect landmark and event clusters.

## Combine both visual features *and* tags

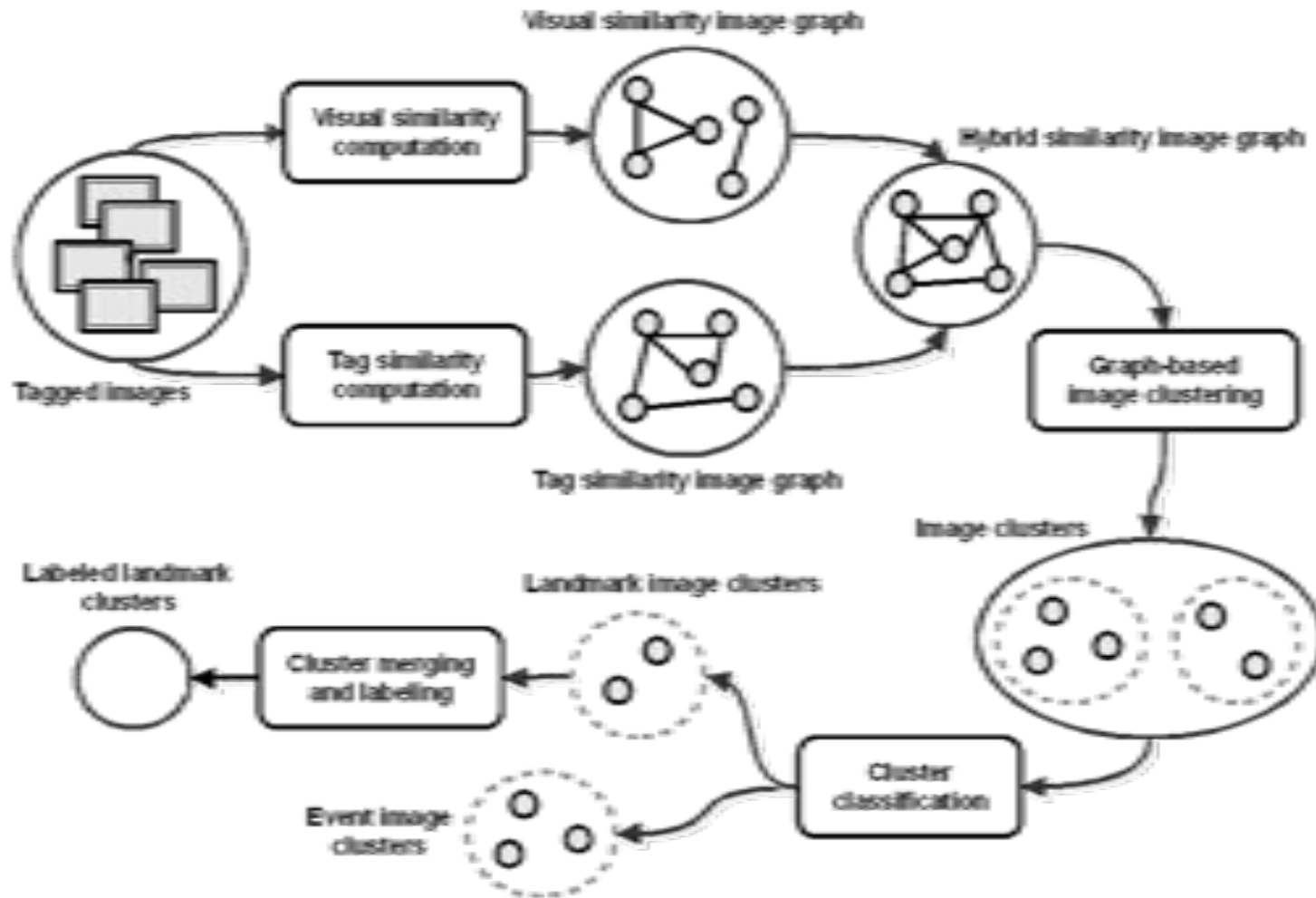
- Two kinds of similarity (visual and tag networks) are complementary to each other
- Many times one photo has missing tags or is hard to interpret visually
- Graph-based approach - superimpose visual and tag graphs
- Use photo cluster features for classification to landmarks/events

## Results

- Higher quality clusters by use of both visual and tag similarity instead of only each one of them.
- Clusters can be used for landmark and event detection.
- Integrated in CSG prototype and ClustTour stand-alone demo.



# Overview of approach



# Photo clustering results (1)

## User study (involving 20 people)

Users were shown photo clusters and they were asked to judge how relevant the photos of each cluster were related to each other

Each cluster was produced by different notion of similarity (tag-only, visual-only, hybrid). Obviously, users were not aware of this information

Hybrid clusters were found to be of superior quality (highest F-measure)

Algorithm	Precision	Recall	F-measure	$\kappa$ -statistic
SCAN-VIS	0.980	0.178	0.301	<b>0.925</b>
SCAN-TAG	0.910	0.197	0.323	0.688
SCAN-HYB	0.898	<b>0.246</b>	<b>0.387</b>	0.637
EXP-VIS	<b>0.985</b>	0.178	0.301	0.895
EXP-TAG	0.929	0.201	0.331	0.709

# Photo clustering results (2)

Geographic localization of results was also found to be very high. Most clusters correspond to landmarks or events.



## EVENTS



# Sample results: [Visual] vs. [Tag] vs. [Visual + Tag]

VISUAL



HYBRID



TAG

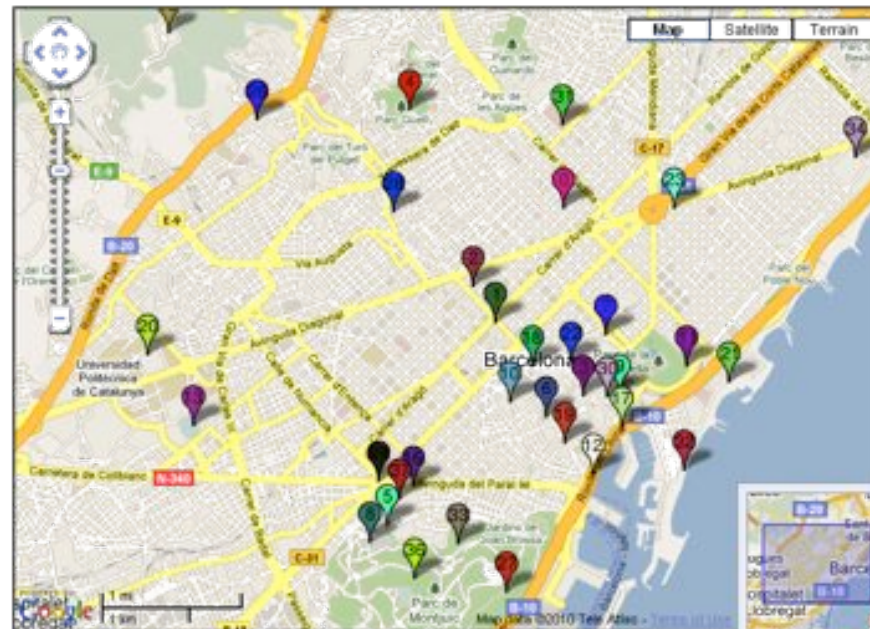




# ClustTour demo: City exploration by means of photo clusters

ClustTour

weknowit 



Landmarks  
Tags: sab, bmx, juanma

- Tag list
- barris gòtics (3)
  - museu (3)
  - yellow (3)
  - cathedral (2)
  - churches (2)
  - espanya (2)
  - catedral (2)
  - casa (2)
  - passeig de gracia (2)
  - el coll (2)
  - mercado (2)
  - arte (2)
  - verde (2)
  - placa (2)
  - port vell (2)
  - plata (2)
  - domènec i montaner (2)
  - diagonal (2)
  - agua (2)
  - apple (2)
  - plants (2)
  - ass (2)
  - sagrada familia (1)
  - sagrada (1)
  - familia (1)
  - catholic (1)
  - construction (1)
  - spire (1)
  - kirche (1)
  - casa batlló (1)

Time filter

From 31/10/2002 up to 1/8/2009

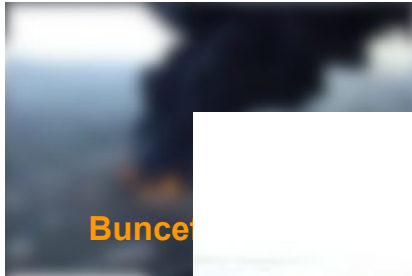


<http://www.clusttour.gr>

# WeKnowIt and CI



## Personal Intelligence



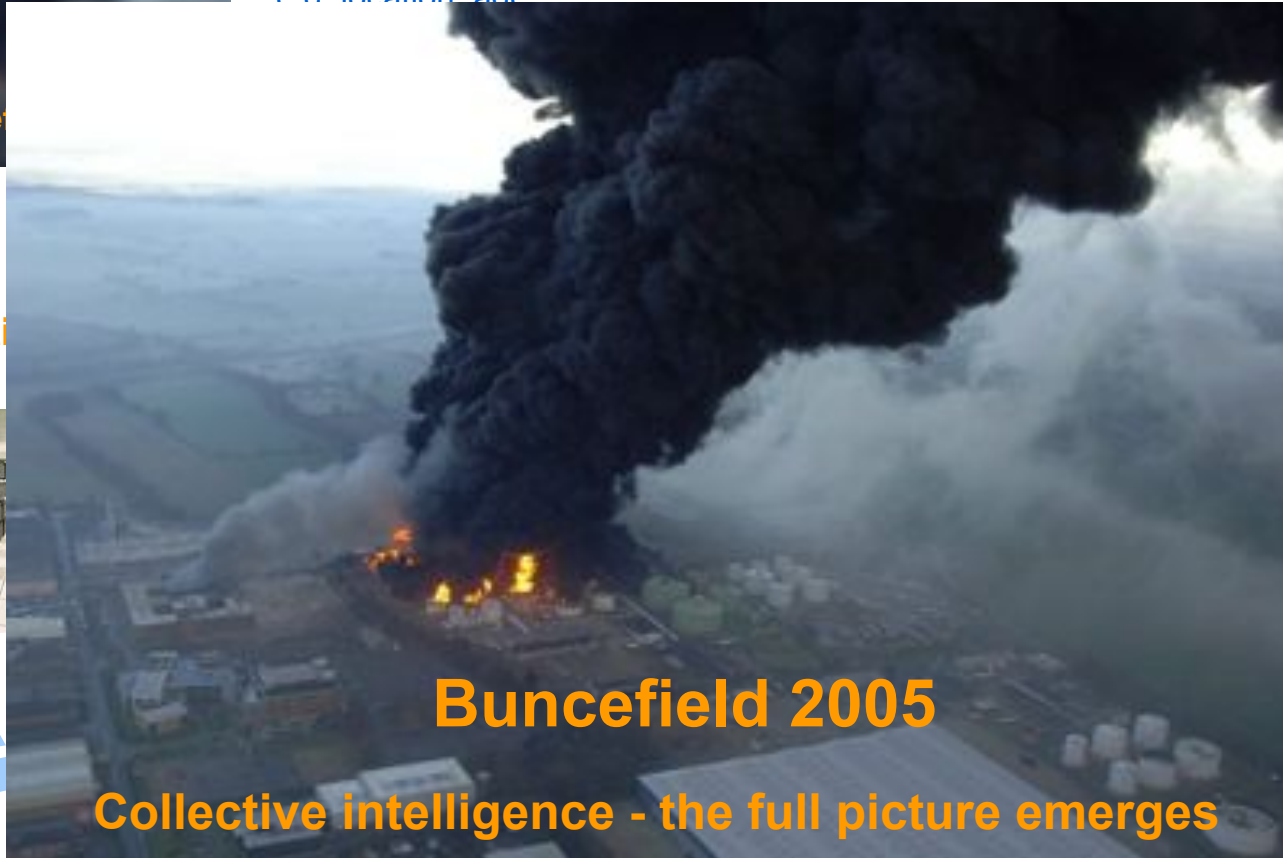
Bunce

### Profile of contributor

>> What to send where,  
e.g. location, age

## Media Intelligence

## Organizational Intelligence

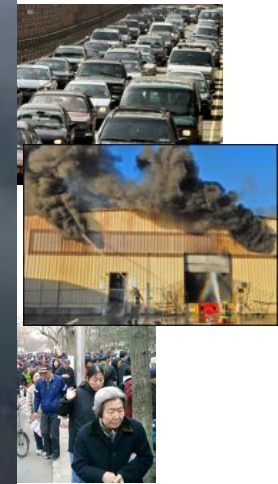


**Buncefield 2005**

**Collective intelligence - the full picture emerges**

### Trust and feedback

>> Determine trustworthiness  
and hub-structures by SNA



# Further Issues

- Not all data always available (e.g. User queries, fb)
- Long tail is forgotten (e.g. flu trends in 3<sup>rd</sup> world countries)
- “More data, less analysis”,.....
- Applications and commercialization
- Efficiency of semantics and analysis
- Real integration
  - not just sum of different analysis
  - formal framework and approach
  - representation
- User interaction – Interfaces

# Thank you!



# References

[Chatzilari09] Elisavet Chatzilari, Spiros Nikolopoulos, Eirini Giannakidou, Athena Vakali and Ioannis Kompatsiaris, "Leveraging Social Media For Training Object Detectors", 16th International Conference on Digital Signal Processing (DSP'09), Special Session on Social Media, 5-7 July 2009, Santorini, Greece.

[Fortunato07a] Santo Fortunato and C. Castellano, "Community structure in graphs", arXiv:0712.2716v1, Dec 2007.

[Freeman77] L. C. Freeman : A set of measures for centrality . Resolution limit in community detection. PNAS, 104(1). pp. 36-41

[Giannakidou08] E. Giannakidou, I. Kompatsiaris, A. Vakali, "SEMSOC: SEMantic, SOcial and Content-based Clustering in Multimedia Collaborative Tagging Systems", In Proc. 2nd IEEE International Conference on Semantic Computing (ICSC' 2008), IEEE Computer Society, August 4-7, 2008 Santa Clara, CA, USA

[Kennedy07] Lyndon S. Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, Tye Rattenbury: How flickr helps us make sense of the world: context and content in community-contributed media collections. ACM Multimedia 2007: 63

[Kumar99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. Computer Networks, 31(11-16):1481-1493, 1999.

[Lindstaedt09] S. Lindstaedt, R. Mörzinger, R. Sorschag, V. Pammer, G. Thallinger. Multimed Tools Appl (2009) 42:97–113, DOI 10.1007/s11042-008-0247-7

[Quack08] Till Quack, Bastian Leibe, Luc Van Gool. World-scale mining of objects and events from community photo collections, In Proceedings of the 2008 international conference on Content-based image and video retrieval, Jul-08

[Zhang06] Y. Zhang, J. Xu Yu, J. Hou : Web Communities: Analysis and Construction, Springer 2006.  
Telematics and Informatics

[Crespoa09] Angel García-Crespoa, Javier Chamizoa, Ismael Riverab, Myriam Menckea, Ricardo Colomo-Palacios and Juan Miguel Gómez-Berbísa, "SPETA: Social pervasive e-Tourism advisor", Volume 26, Issue 3, August 2009, Pages 306-315, Mobile and wireless communications: Technologies, applications, business models and diffusion.