Extracting Collective Intelligence from Social Content Analysis

Yiannis Kompatsiaris
Senior Researcher
CERTH - Informatics and Telematics Institute
Thessaloniki, Greece
http://mklab.iti.gr

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

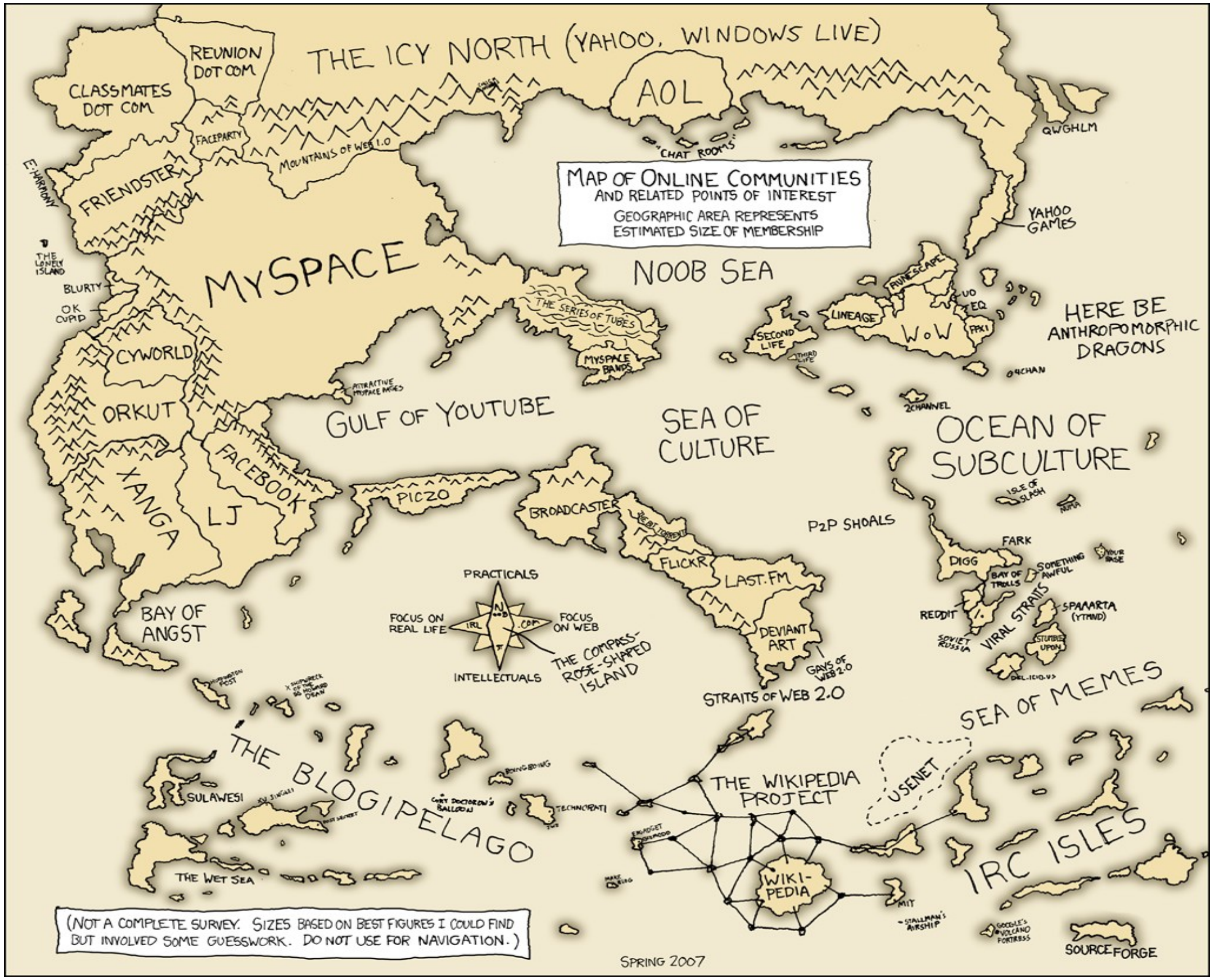
Contents

- Defining Collective Intelligence
- Clustering in Social Media
  - Clustering in social content
  - Applications of clustering
- Community detection in Social Media
- Collective Intelligence in WeKnowIt
- Conclusions - Issues

Introduction – Defining Collective Intelligence

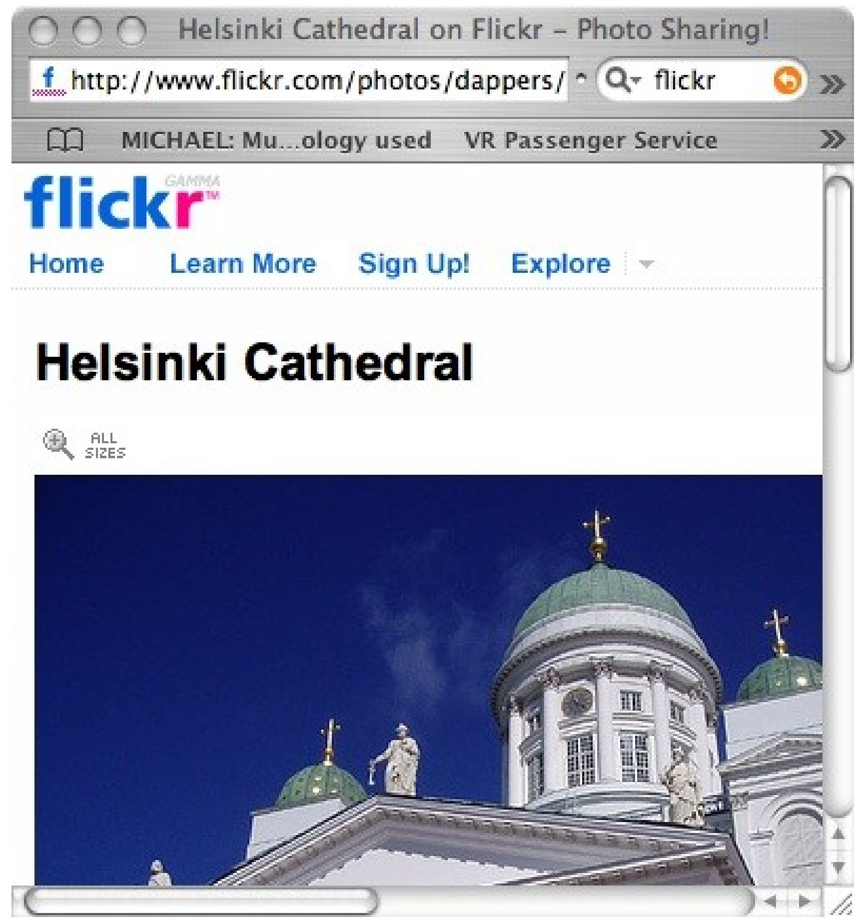
SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Web 2.0 Map (already old)



Evolution of Content and HPC…

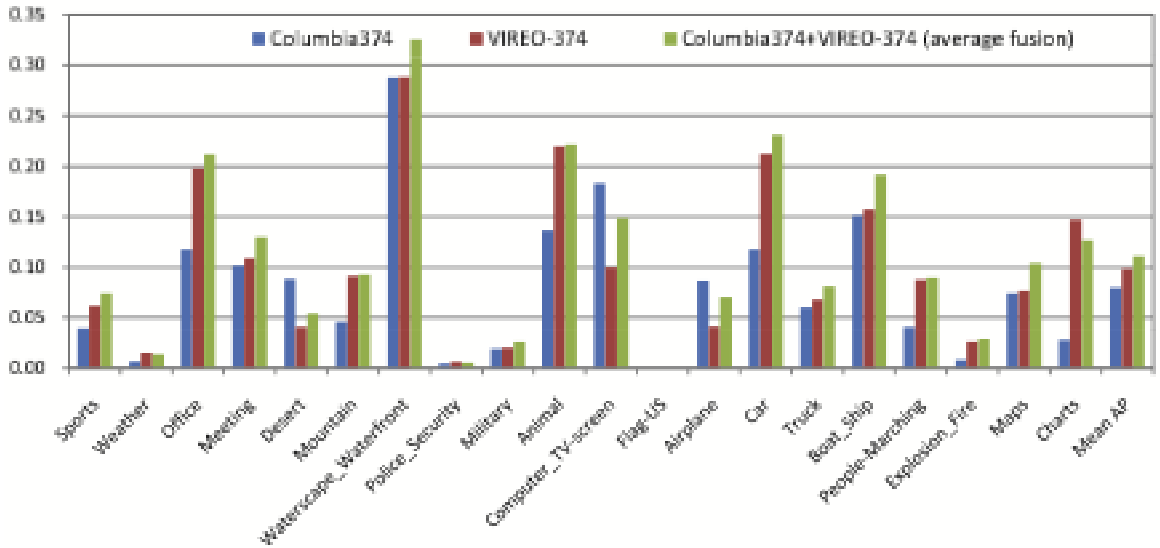
- 20h of video content uploaded every minute at YouTube (2009)
- 3,024,780,142 photos in Flickr @ 11:52, 12 Nov 2008
- 2 million geotagged photos uploaded each month (2008)



Facebook:

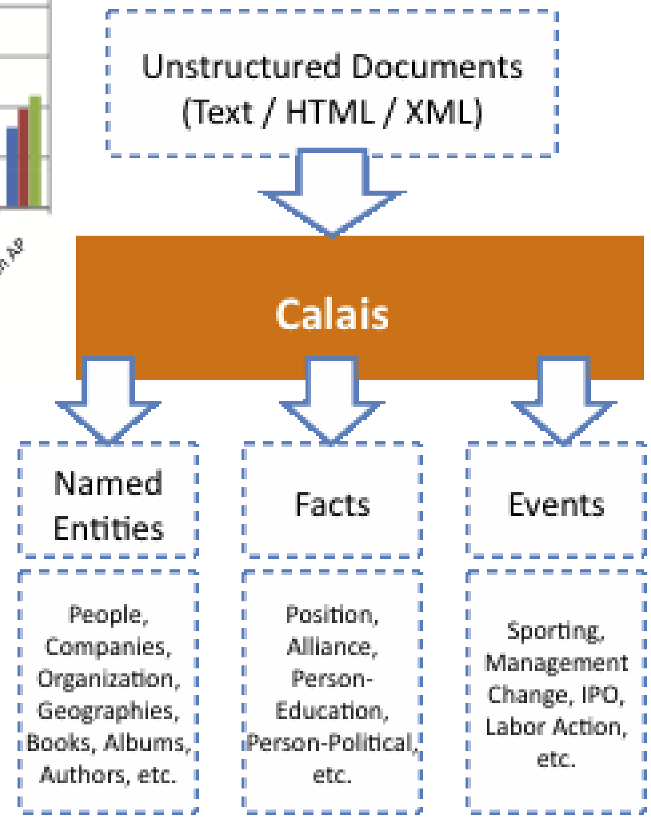
- More than 250 million active users
- More than 120 million users log on to Facebook at least once each day
- More than 1 billion photos uploaded to the site each month

Content Analysis (Text, Visual, ...)



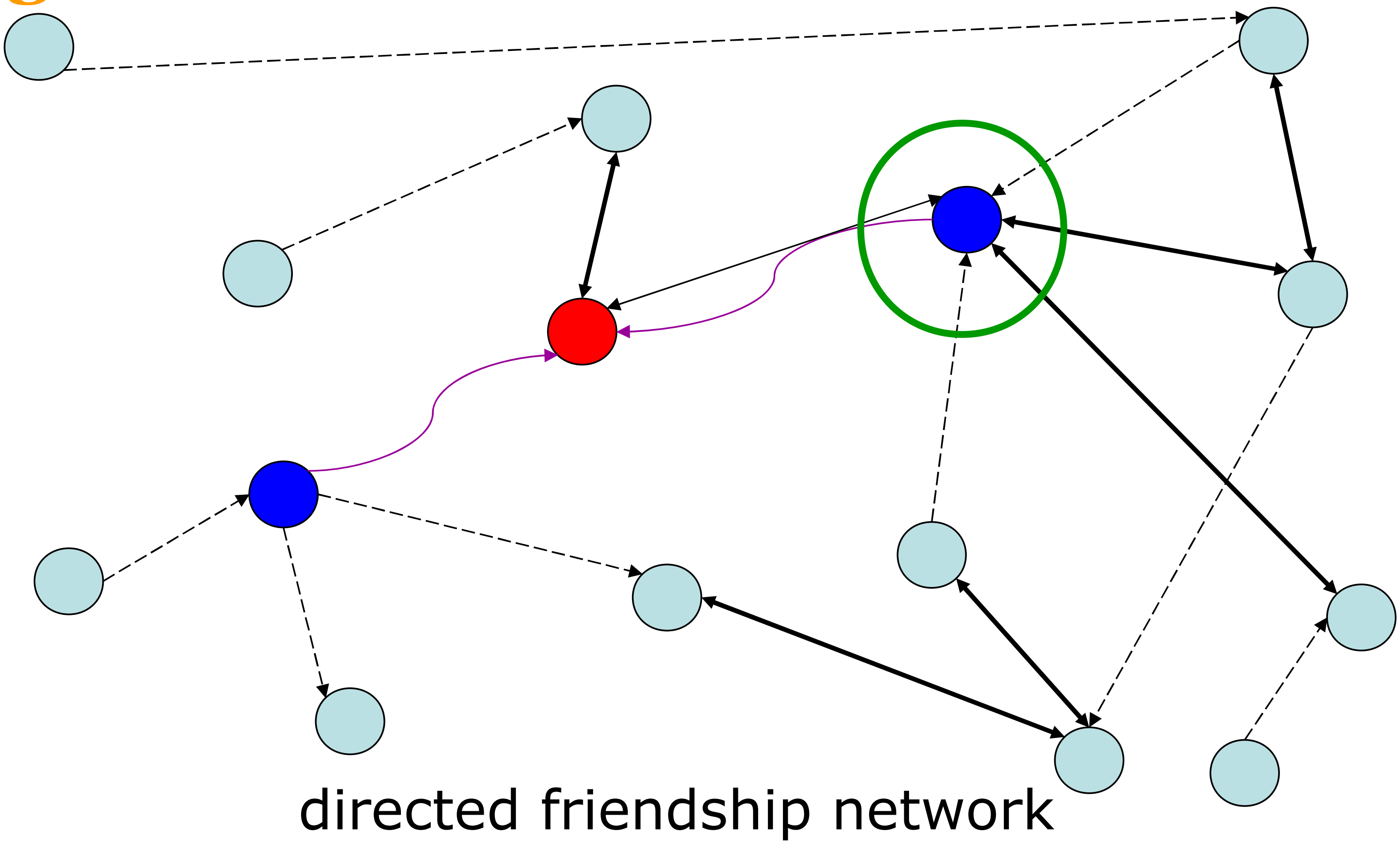
Concept analysis on TRECVID 2007 test data

OpenCalais Web Service automatically creates rich semantic metadata

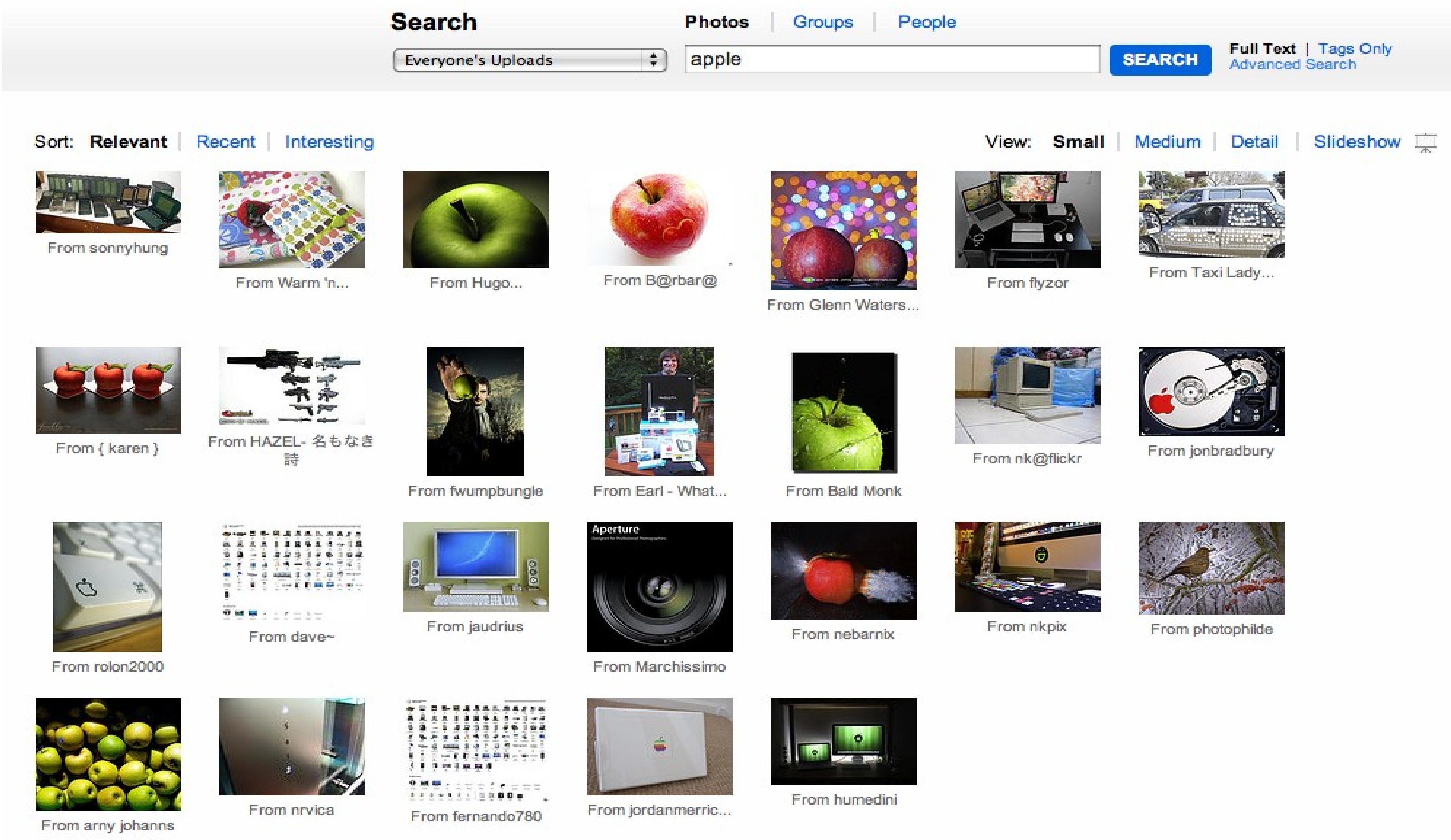


SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

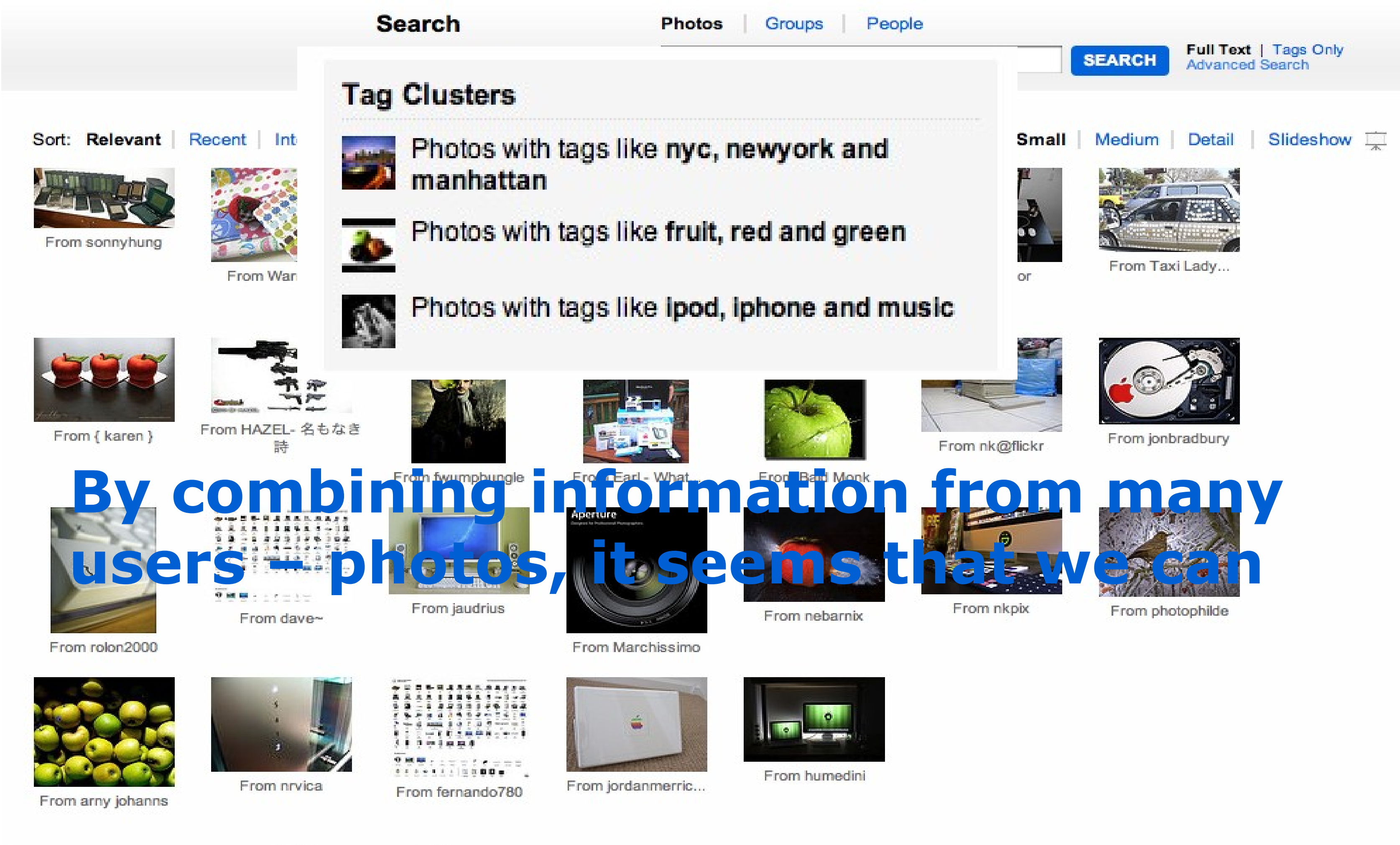
Social Networks Analysis
e.g. which source to trust?



Can we really access all this info?



Can we generate knowledge?



By combining information from many users – photos, it seems that we can

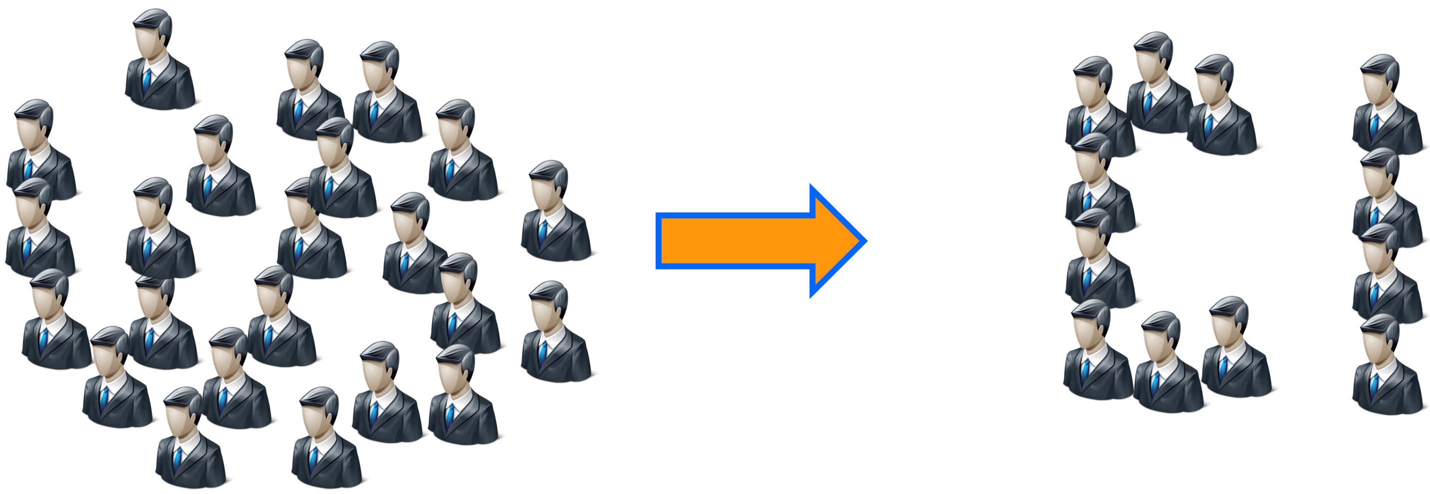
weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Defining Collective Intelligence

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

weknowit

SEVENTH FRAMEWORK PROGRAMME

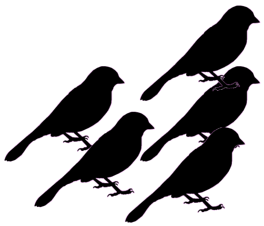
co-funded by the European Union

What is collective?

Human Collectives

(e.g. communities, groups, organizations, families)

Groups of intelligent agents in computer environments



Animal Collectives

(e.g. ants, birds, bees)

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Who is intelligent?

There are too many different definitions out there.

Defining intelligence is controversial and elusive activity.

Characteristics, capacities, functions that can be ascribed to intelligence

problem solving

decision making

reasoning

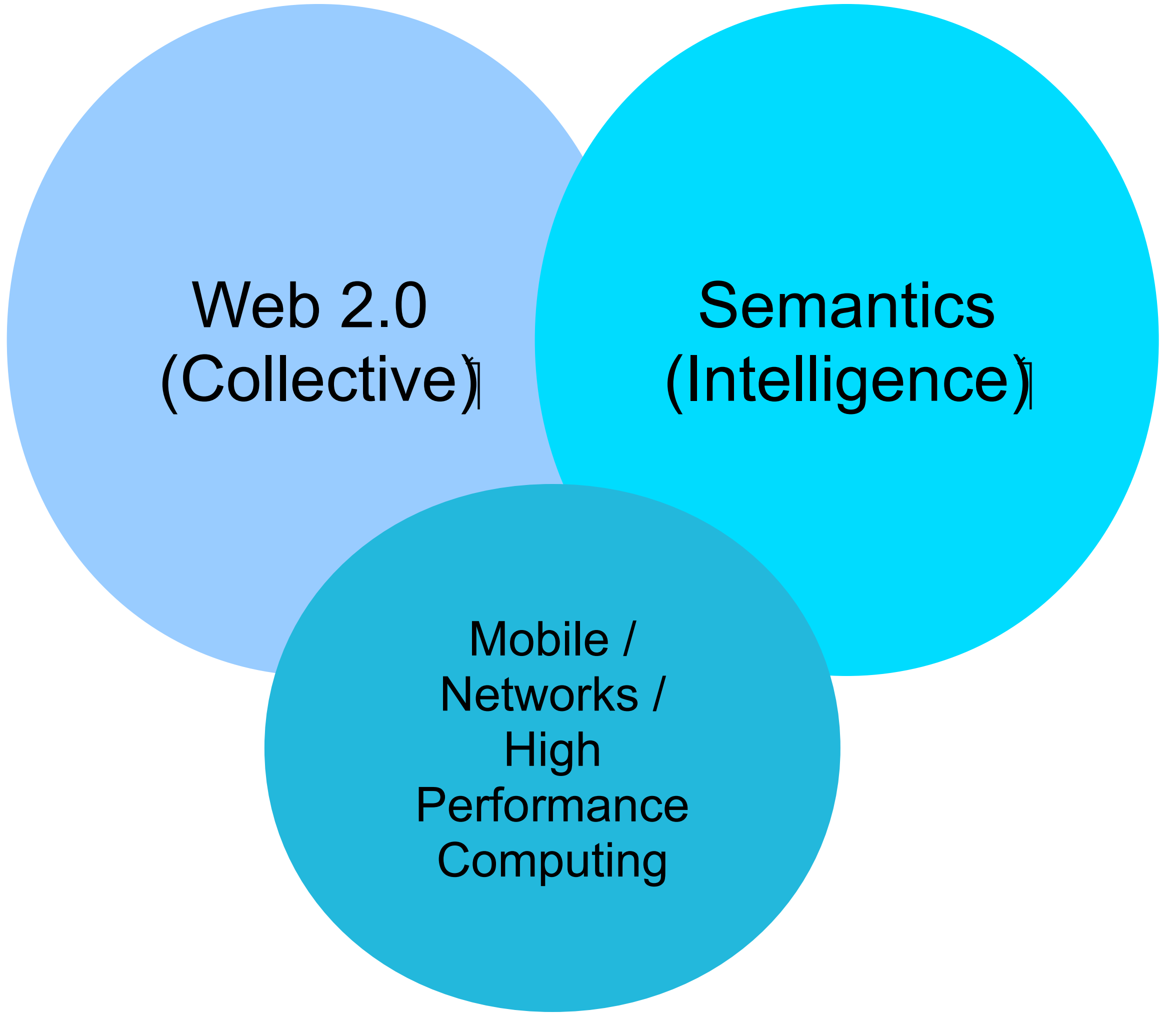
applying knowledge

integration, synthesis

information gathering,
sorting and categorization

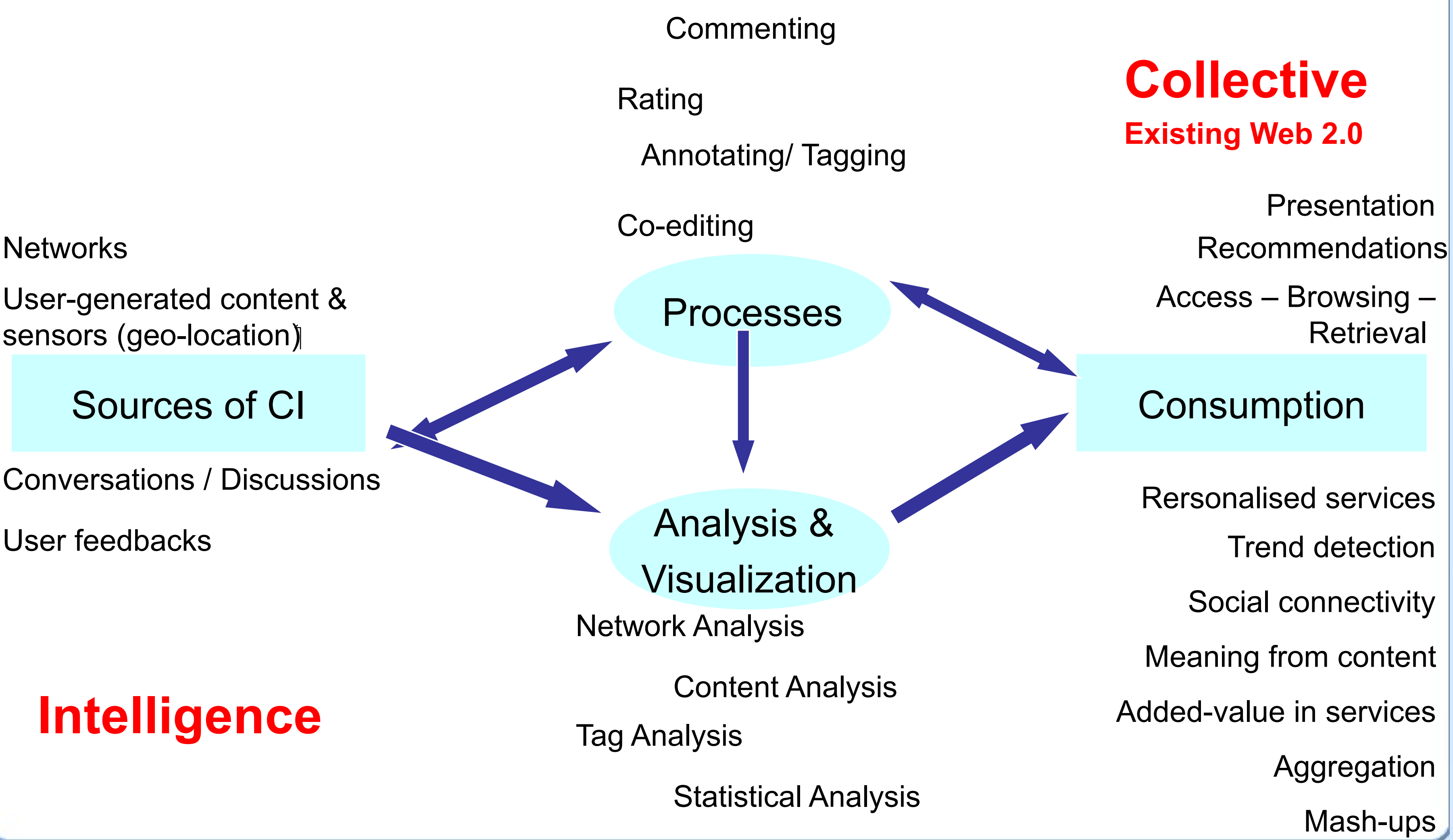
evolution

Why today?



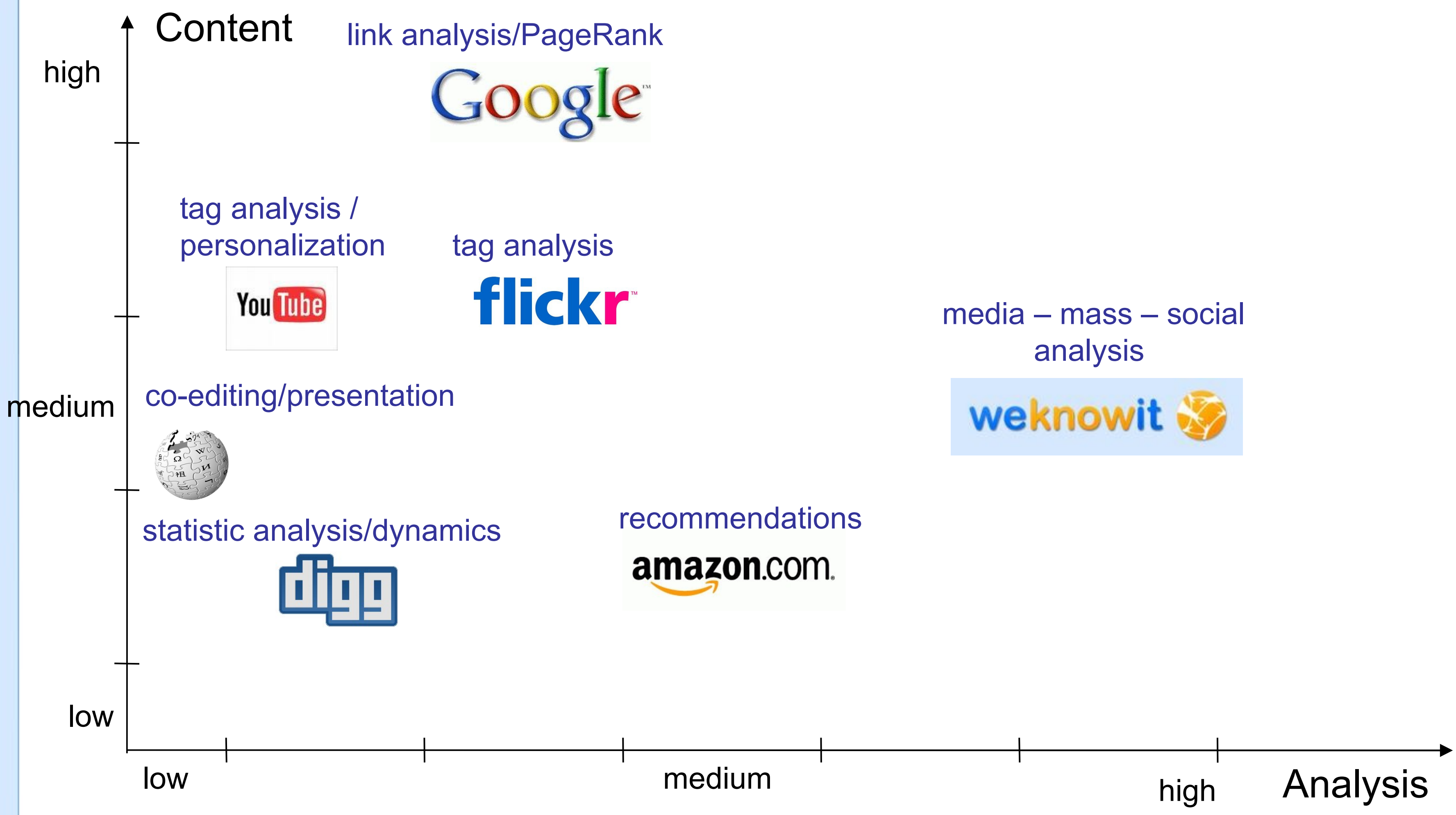
Web 2.0
(Collective)
Semantics
(Intelligence)
Mobile /
Networks /
High
Performance
Computing

Collective Intelligence Overview



co-funded by the European Union

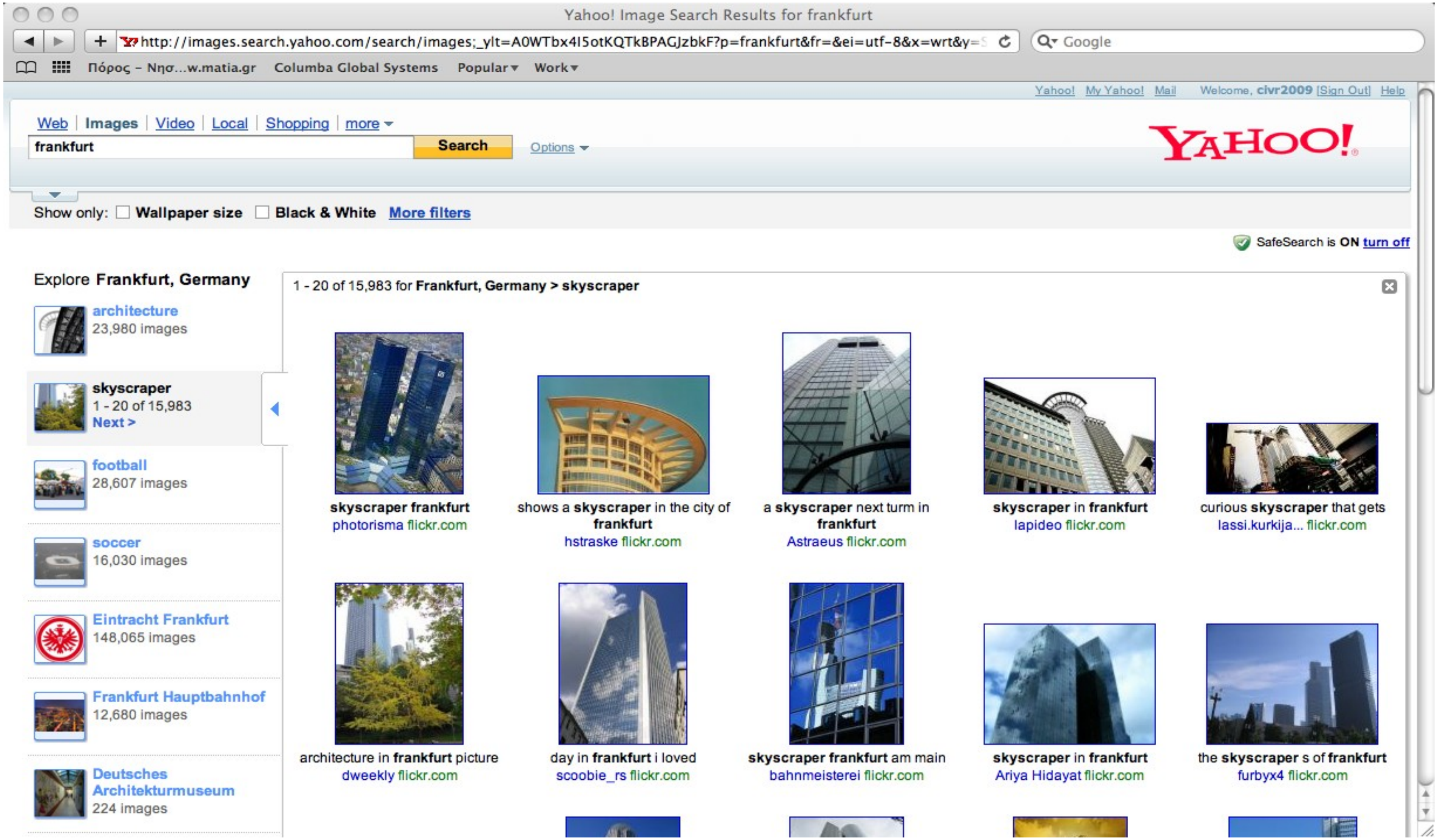
Analysis? What analysis?



weknowit

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Example of analysis



weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Analysis Example

Extracting Meaning from Millions of Pages

- A software engine that pulls together **facts** by **combing** through more than **500 million Web pages** has been developed by researchers at the University of Washington. The tool extracts information from billions of lines of text by **analyzing basic relationships between words.**

http://www.technologyreview.com/computing/22773/

weknowit

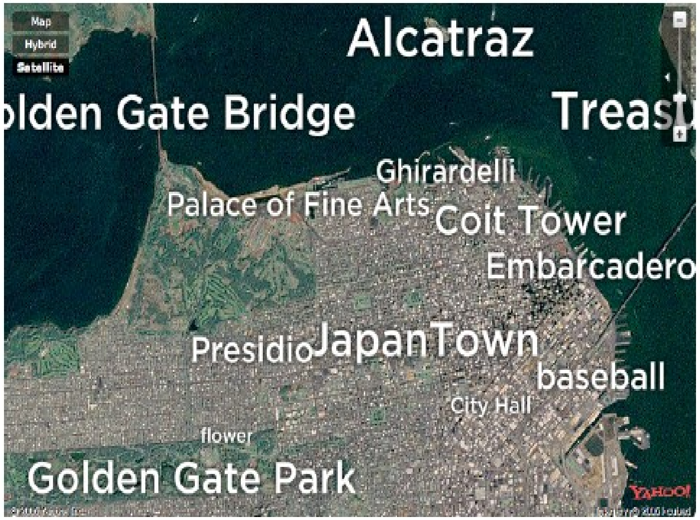
SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Visualization (and analysis) example

Tags that are "representative" for a geographical area

Contribute to our understanding of the world

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



Representative tags for San Francisco [Kennedy07]

weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Clustering for Social Media

Social Tagging & Multimedia Content Clustering

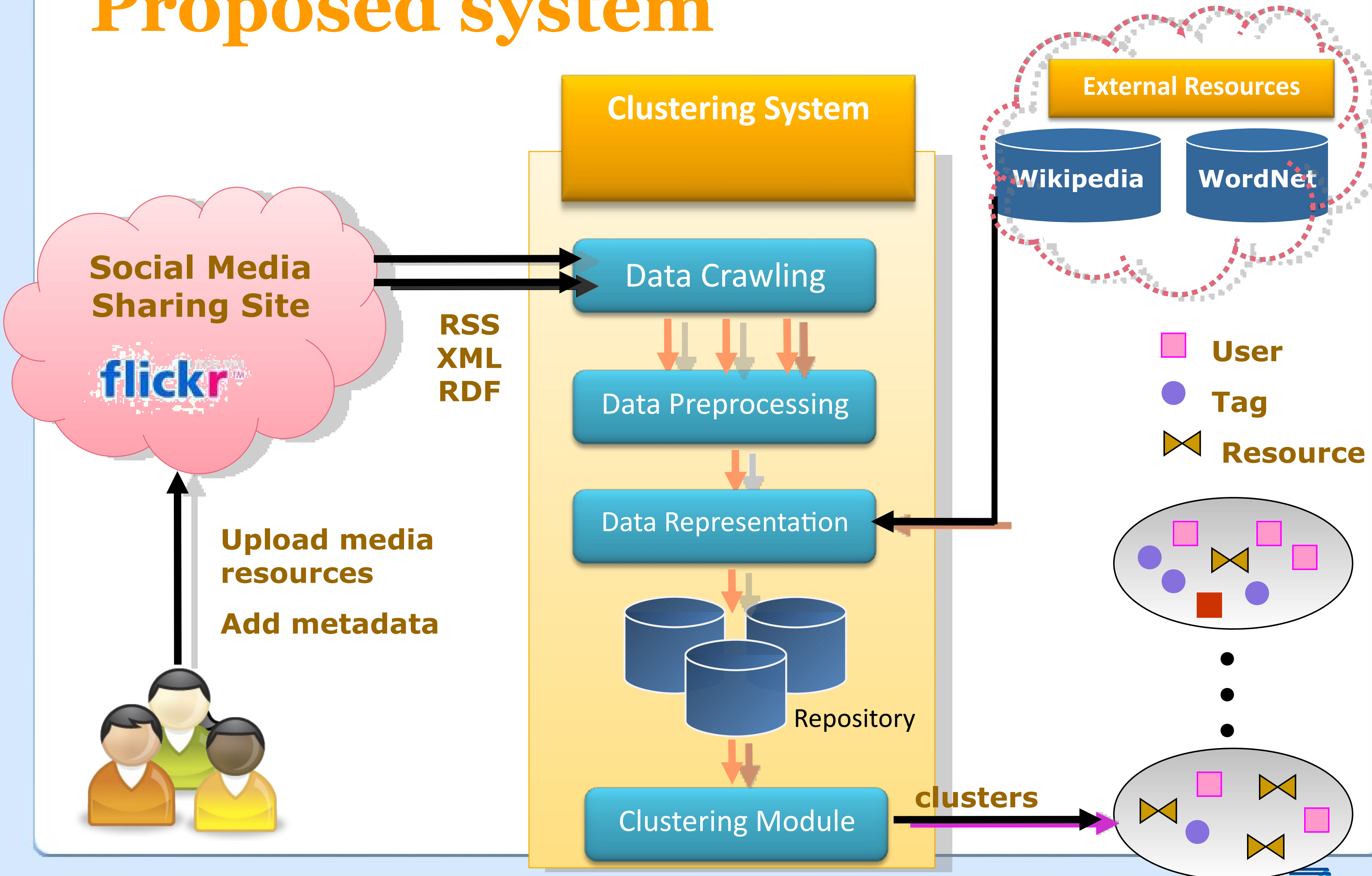
- **Background**
  - High availability of multimedia content in social media sharing sites as source of CI
  - Plenty of user-generated metadata
  - Stable patterns in tagging systems over time
- **Motivation**
  - Poor IR (lack of structure of information, tag polysemy/ambiguity, chaotic environment)
  - Questionable tag validity
  - Produce clean image-video databases
- **Problem Formulation**
  - Overcome of limitations and exploitation of (hidden) knowledge harvested in social media sharing sites through **clustering**.

Clustering Approaches

- Tag-Based
- Content-Based
- Co-clustering
  - Tags - resources
  - Time-based: users and tags

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Proposed system



SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

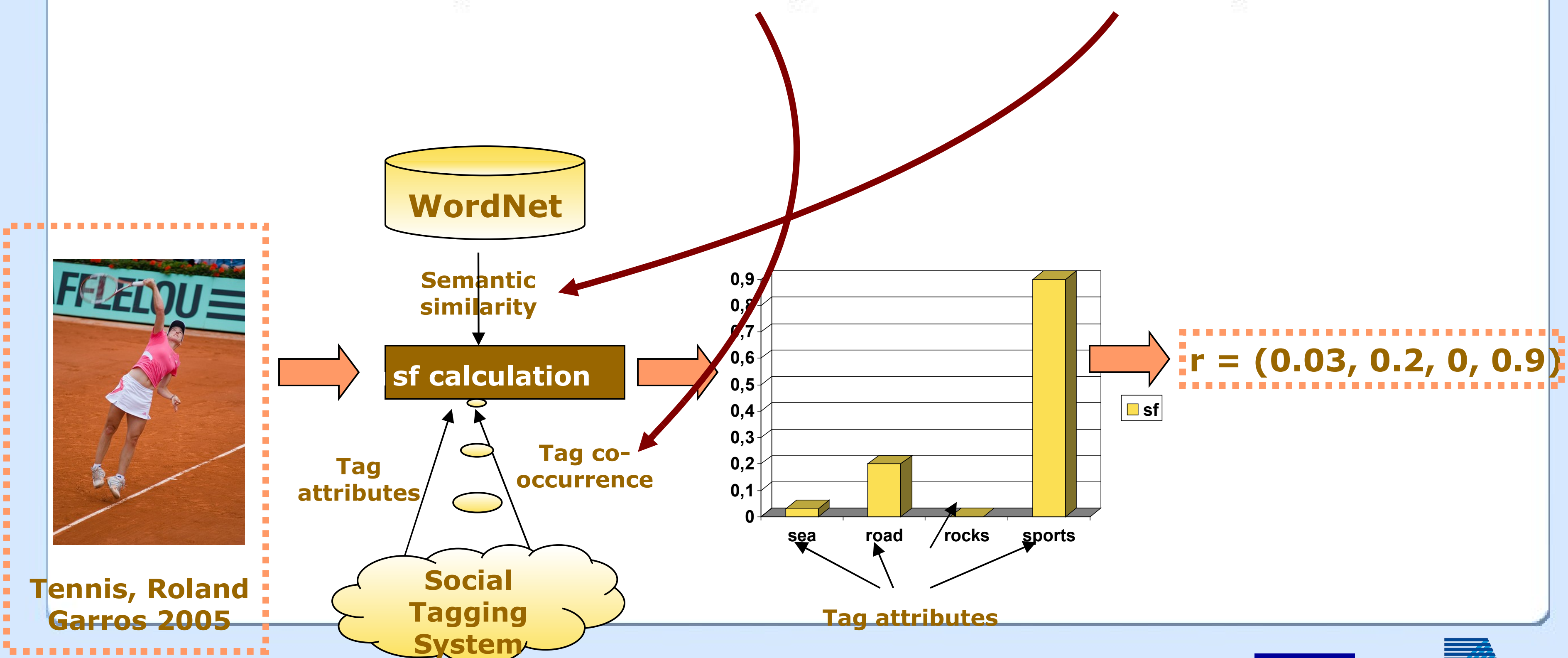
Tag-based Clustering (I)

- **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 $\mathbf{r}_i$ is represented by a **d**-dimensional vector $\mathbf{r}_i = (\mathbf{sf}_1, \mathbf{sf}_2, \ldots, \mathbf{sf}_d)$
- All resources can be represented by an **n** x **d** matrix

Tag-based Clustering (II)

- **2. Clustering on n (resources, r) x d (attributes) matrix (K-means, Hierarchical, COBWEB)**

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



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

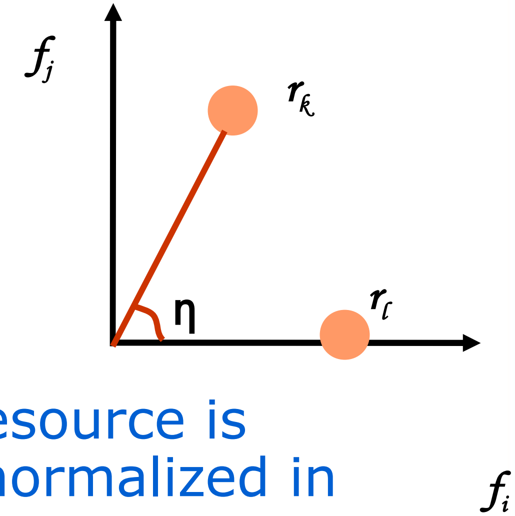
(b)

rocks, stone, rockyside

rock, music, band

Tag & Content-based Clustering

- **Method:** After performing tag-based clustering, low-level features of resources are used for cluster refinement (outlier detection)
- **Vector data model**
- For each resource the following visual descriptors are extracted:
  - Scalable Color, $\mathcal{SC}$
  - Color Structure, $\mathcal{CS}$
  - Color Layout, $\mathcal{CL}$
  - Edge Histogram, $\mathcal{EH}$
  - Homogenous Texture, $\mathcal{HT}$

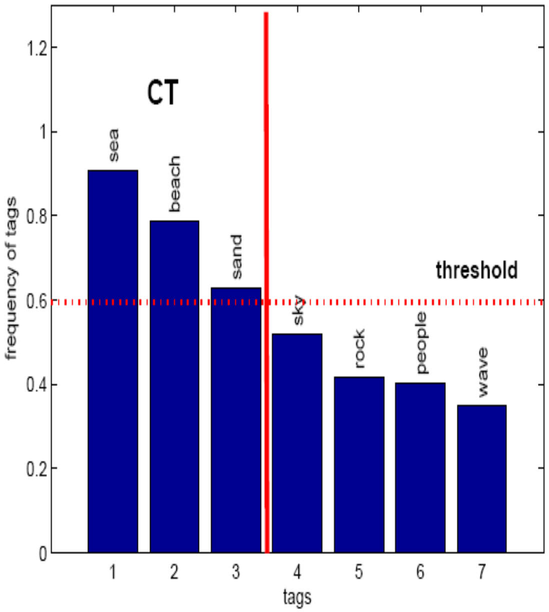


- A single image feature vector per each resource is produced, encompassing all descriptors normalized in [0,1]
- Feature extraction and distances between image feature vectors are according to MPEG-7 XM.

Evaluation Method

- **Definition:** Cluster Topic, CT, are the tags that have frequency in cluster's resources annotation over a threshold **τ.**

- **Evaluation Metrics**
- Precision $$Pr\left(C_j\right) = \frac{|C_j \cap RR(C_j)|}{|C_j|}$$
- Recall $$R\left(C_j\right) = \frac{|RR(C_j) \cap C_j|}{|RR(C_j)|}$$
- F-Measure $$F\left(C_j\right) = \frac{2 * Pr\left(C_j\right) * R\left(C_j\right)}{Pr\left(C_j\right) + R\left(C_j\right)}$$

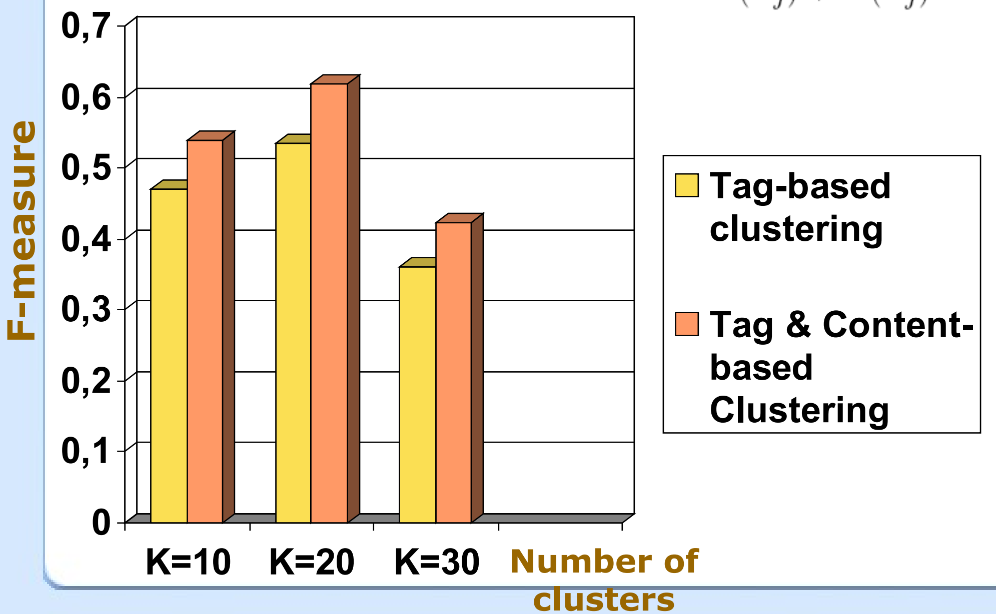


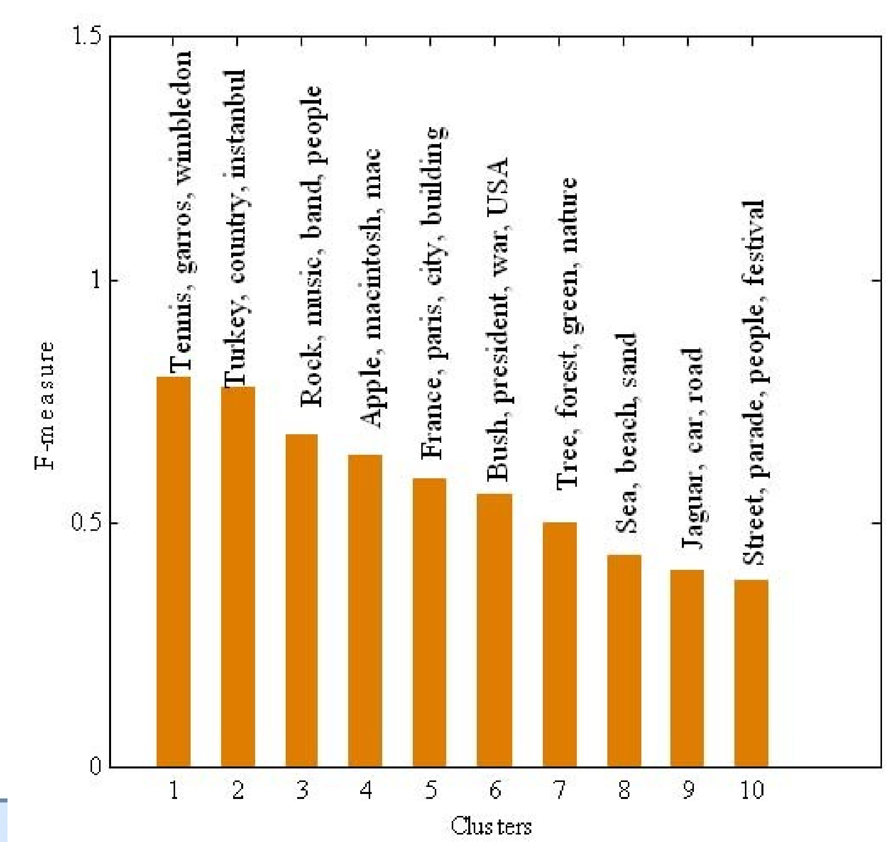
Tag & Content-based Clustering – Experimental Results

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

Evaluation: Manual annotation and use of F-Measure.

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





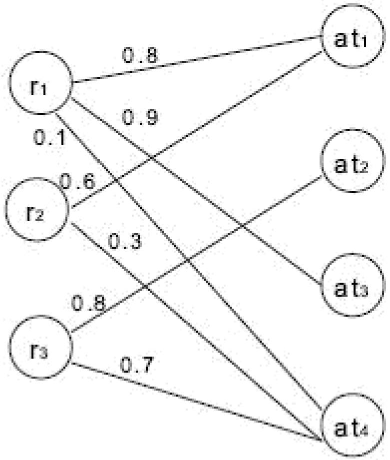
weknowit

SEVENTH FRAMEWORK PROGRAMME

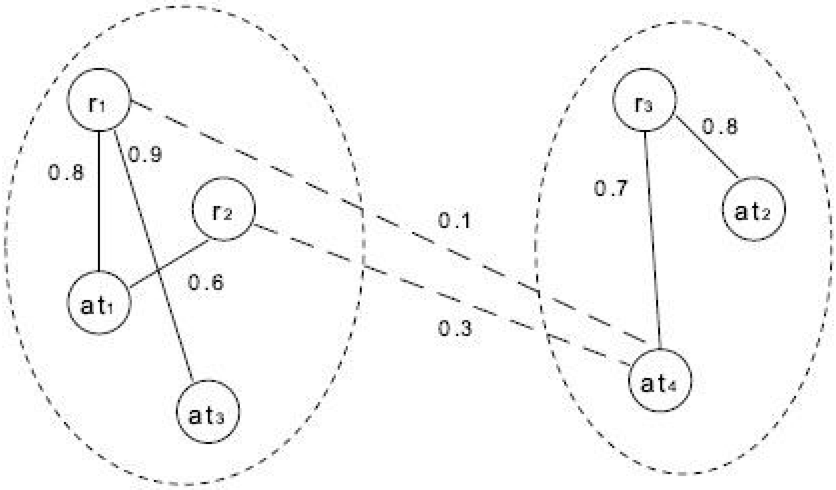
co-funded by the European Union

Co-clustering

- **Graph data model**
- A graph structure $\mathbf{G} = \{\mathbf{V_1}, \mathbf{V_2}; \mathbf{E}\}$ is used for the representation of the dataset, where $\mathbf{V_1}$ and $\mathbf{V_2}$ can be sets of resources, users, tags or time intervals and **E** denotes the relations between the nodes of $\mathbf{V_1}$ and $\mathbf{V_2}$.



Graph representation



Graph-partitioning problem

Co-clustering Tags & Resources

Problem: Find k clusters of both resources and tags, such that:

$$\sum_{x=1}^{k} \sum_{r_i, a_j \in C_x} Similarity\,(r_i, a_j), \forall\, r_i\, R, a_j\, AS$$

is maximized ■

R: Resources Set

AS: Tag-attributes Set

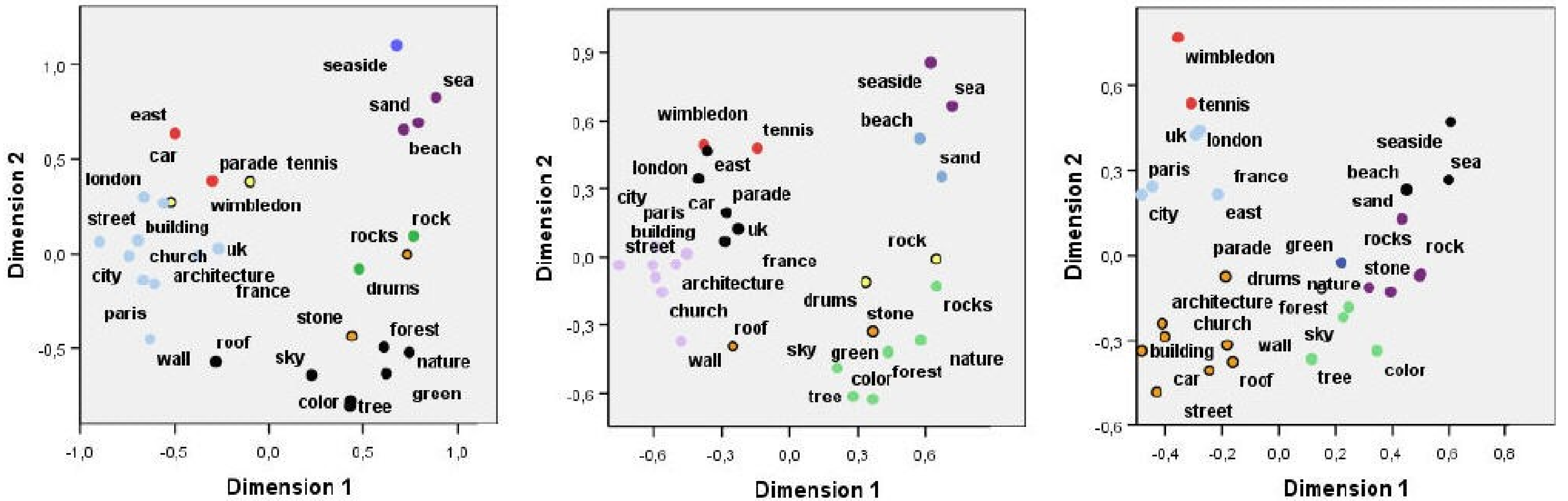
Algorithm 1 The CO-CLUSTERING algorithm.

Input: The set R of n resources, the set T of l tags and two integers k and w where $w \in [0..1]$

Output: A set $C = \{C_1, \ldots, C_k\}$ of k subsets consisting of elements from both R and T, such that the sum of inter-clusters similarities defined by (6) is minimized.

```
/*Preprocessing*/
T* = Preprocess(T)
AS = ExtractAttributes(T*)
/*capturing similarities*/
SoS = CalculateSocialSimilarity(R,AS)
SeS = CalculateSemanticSimilarity(R,AS)
SS = w * SoS + (1 - w) * SeS
RA = Similarity(SS)
/*Co-clustering process*/
(D_r, D_at) = ComputeDegreeTables(RA)
NRA = D_r^{-1/2} RA D_at^{-1/2}
(L_r, R_at) = SVD(NRA)
SV = CreateIntegratedTable(D_r, D_at, L_r, R_at)
C = k - means(SV, k)
```

Co-clustering Tags & Resources - Experimental Results (I)



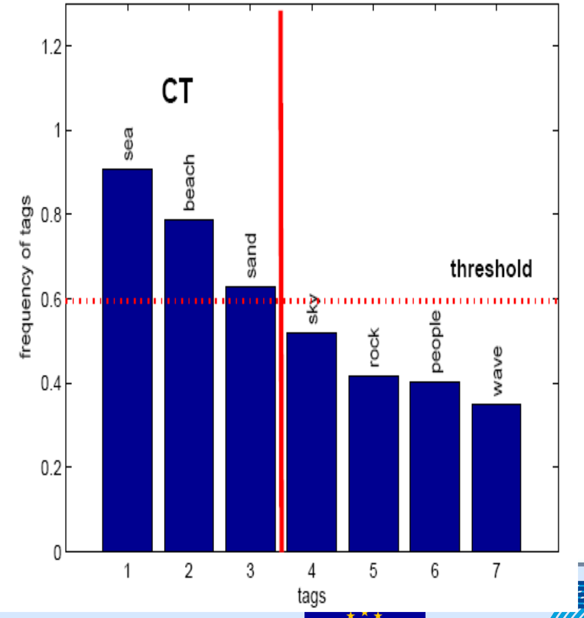
(a) $w = 0.2$ (b) $w = 0.5$ (c) $w = 0.8$

Attributes Assignment to k=8 clusters,

$\mathcal{W}$: weighting factor of semantic similarity against similarity derived from tag co-occurrence

Co-clustering Tags & Resources - Evaluation Method

- **<u>Definition:</u>** Cluster Topic, CT, are the tags that have frequency in cluster's resources annotation over a threshold τ.
- A resource is considered correctly assigned to a cluster C, if it contains **all** the tags of the CT of C.
- **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 $$F(C_j) = \frac{2 * Pr(C_j) * R(C_j)}{Pr(C_j) + R(C_j)}$$

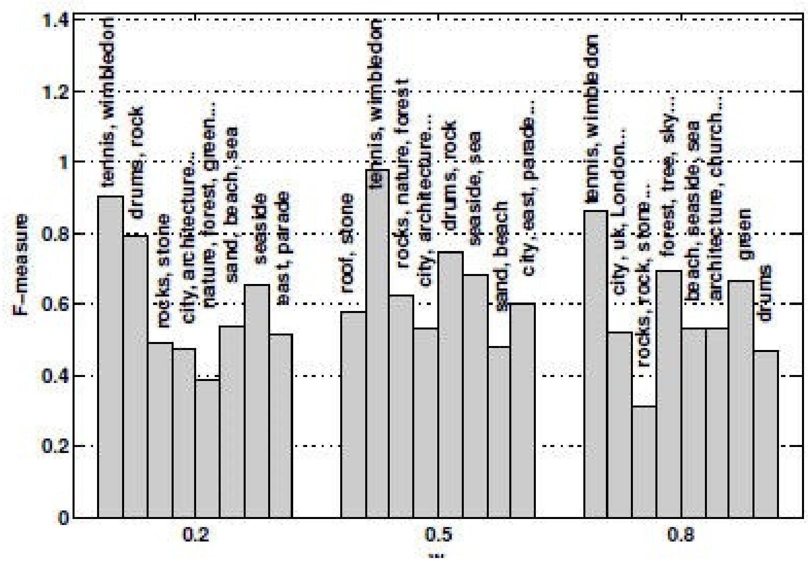


weknowit

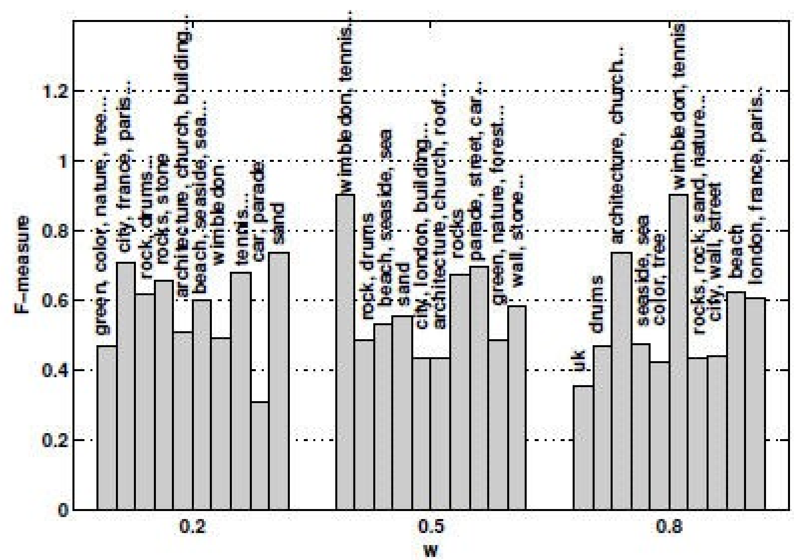
SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Co-clustering Tags & Resources - Experimental Results (II)



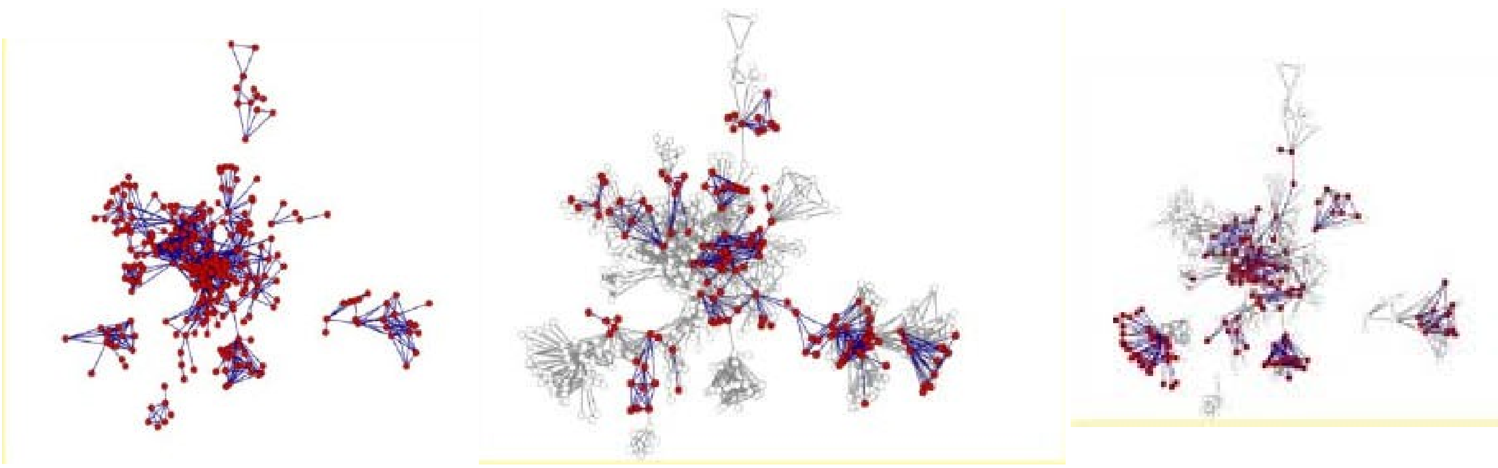
k = 8



k = 10

Users-Tags Co-clustering over time

- **Problem:**
- Compute similarities over time between users and tags
- Find Dominating topics per time slot



Why consider time?

- Motivation

Events, Trends, Changing of user interests

⇓

Users Tagging Behavior changes over time

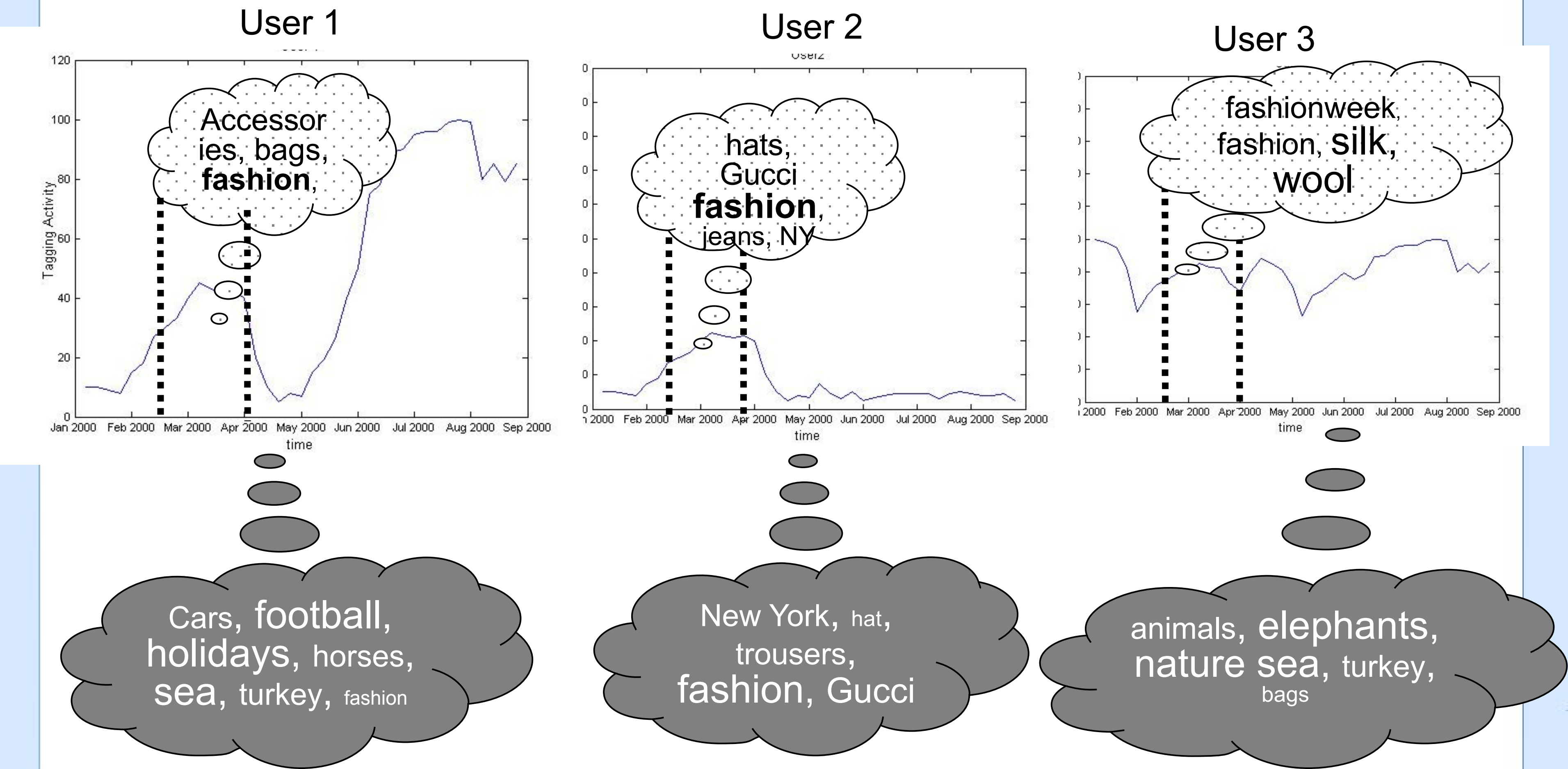
⇓

Time is a fundamental dimension in analysis of users and tags in a social tagging system

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 at specific time periods (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)

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



weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

The basic idea

$$UTF = \begin{array}{cc} & \begin{array}{cccc} i_1 & i_2 & \cdots & i_D \end{array} \\ \begin{array}{c} u_1 \\ u_2 \\ \vdots \\ u_N \end{array} & \begin{bmatrix} ut_{11} & ut_{12} & \cdots & ut_{1D} \\ ut_{21} & ut_{22} & \cdots & ut_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ ut_{N1} & ut_{N2} & \cdots & ut_{ND} \end{bmatrix} \end{array},$$

$$TTF = \begin{array}{cc} & \begin{array}{cccc} i_1 & i_2 & \cdots & i_D \end{array} \\ \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_S \end{array} & \begin{bmatrix} tt_{11} & tt_{12} & \cdots & tt_{1D} \\ tt_{21} & tt_{22} & \cdots & tt_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ tt_{N1} & tt_{N2} & \cdots & tt_{SD} \end{bmatrix} \end{array},$$

Step 1: Representation

Step 3: Focus on time locality

$$< u_j, t_k > = \frac{u_j * t_k}{\sqrt{\sum_{j=1}^{N} u_i^2 * \sum_{k=1}^{S} t_k^2}}$$

Step 4: Combination of semantic and time information

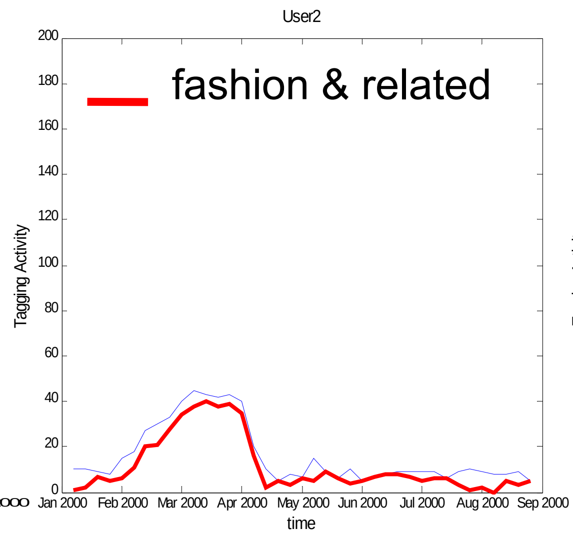
*(u_i,t_j)= SemSim(u_i,t_j) * InnerProduct(u_i,t_j)*

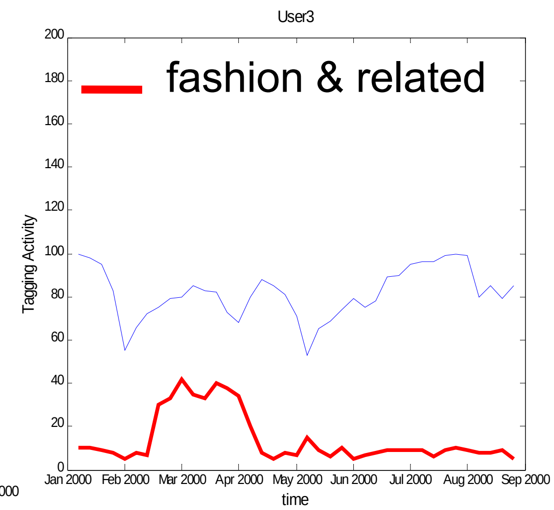
WordNet

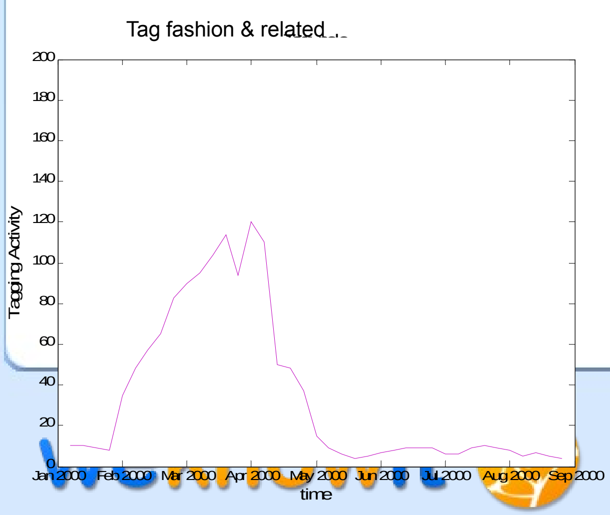
Step 2: Focus on contents (tags semantics)

$$\begin{array}{cc} & \begin{array}{cccc} t_1 & t_2 & \cdots & t_D \end{array} \\ \begin{array}{c} u_1 \\ u_2 \\ \vdots \\ u_S \end{array} & \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1D} \\ w_{21} & w_{22} & \cdots & w_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{SD} \end{bmatrix} \end{array}$$

An example



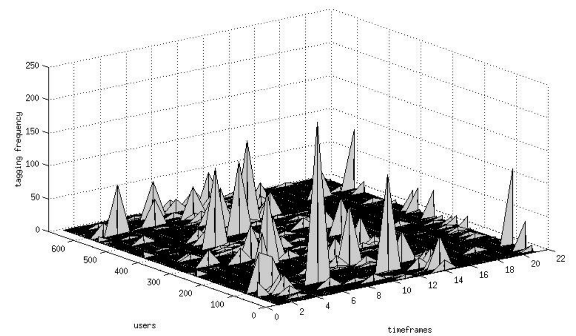


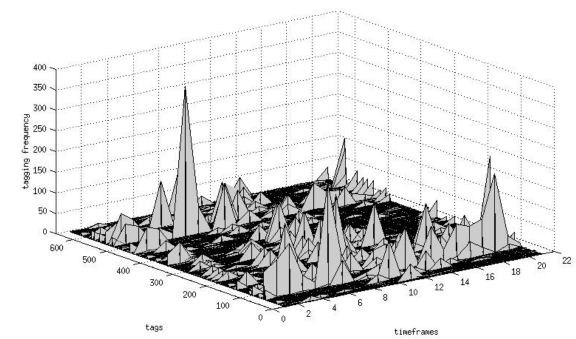


$$(u_i,t_j)= SemSim(u_i,t_j) * InnerProduct(u_i,t_j)$$

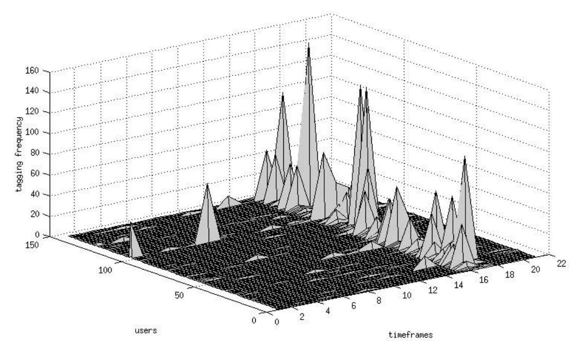
Time-aware user/tags clusters on Flickr (I)

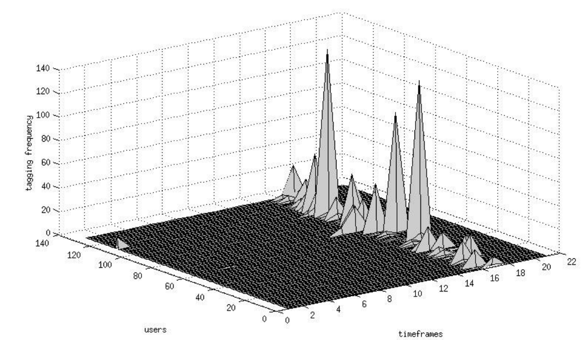
Cluster of users interested regularly in weddings and related tags





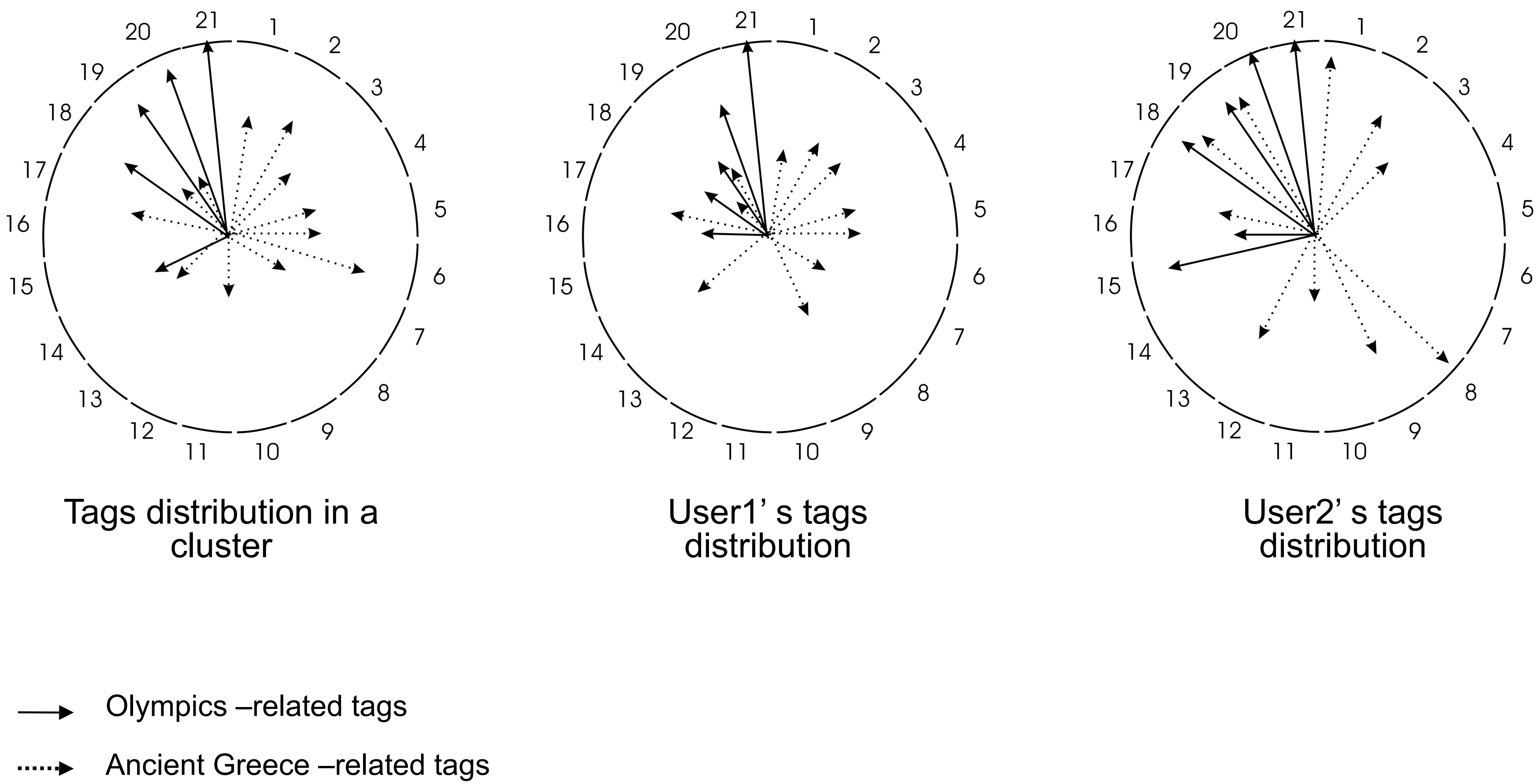
Cluster of users interested in Olympics and related tags





SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Time-aware user/tags clusters on Flickr (II)



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)

Clustering Applications

weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Applications of Clustering

- Generating summaries
  - Identification of representative images (summaries) of areas, objects, events, etc
- Generating recommendations
  - Auto-annotation of unlabelled content
- Enhancing retrieval of multimedia content
  - Training classifiers

Generating summaries issues

- Build the appropriate representation (e.g. usage of tags, resources (features), time, location info etc)
  - Depending on the application
  - Improve results
- Select the appropriate clustering algorithm
  - Usually simple approaches (e.g. k-means)
- Efficiently exploit the results and build application
  - Emphasis on this part

Benefits of summaries

- Quick finding of useful information
- Efficient browsing of large collections of images

Tag summaries in particular:

- Enable applying weights in tags

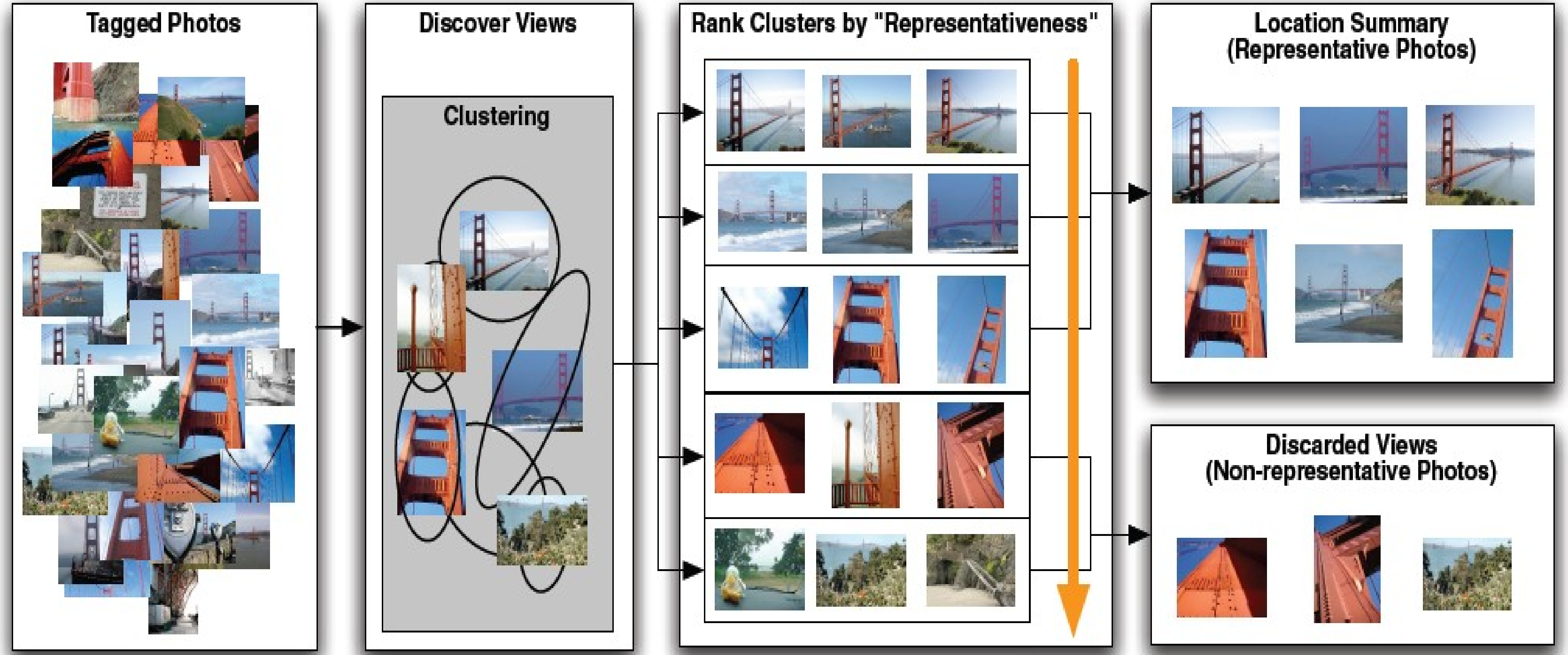
  important tag → larger weight

- Summaries value enhances, as the dataset grows

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

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]

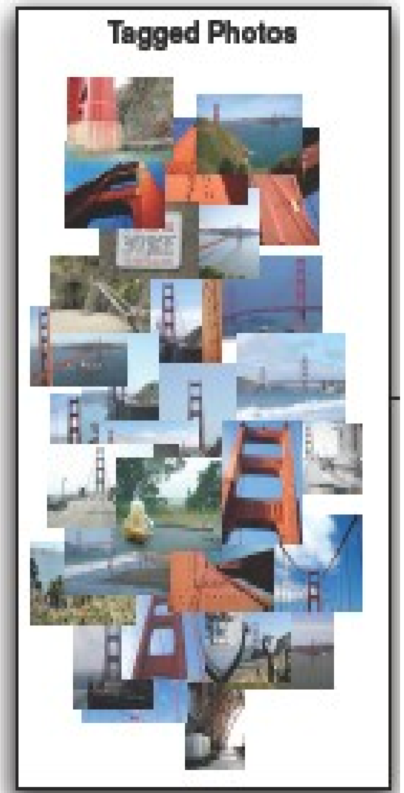
weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Step 1: Get relevant content…

- Using a clustering method, obtain a list of (x, C) – tag x is representative in cluster C – $C_{x,1}, C_{x,2}, \ldots$
- Let $\mathbb{C}_x \triangleq C_{x,1}, C_{x,2}, \ldots$ the set of photo clusters in which x was identified as representative.
- Get the photo set $\mathbb{P}_{x,\mathbb{C}_x} \triangleq \mathbb{P}_x \cap \mathbb{P}_{\mathbb{C}_x}$
  - $\mathbb{P}_{\mathbb{C}_x}$: The photos of the $\mathbb{C}_x$ associated with tag x
  - $\mathbb{P}_x$: The photos of the dataset associated with tag x

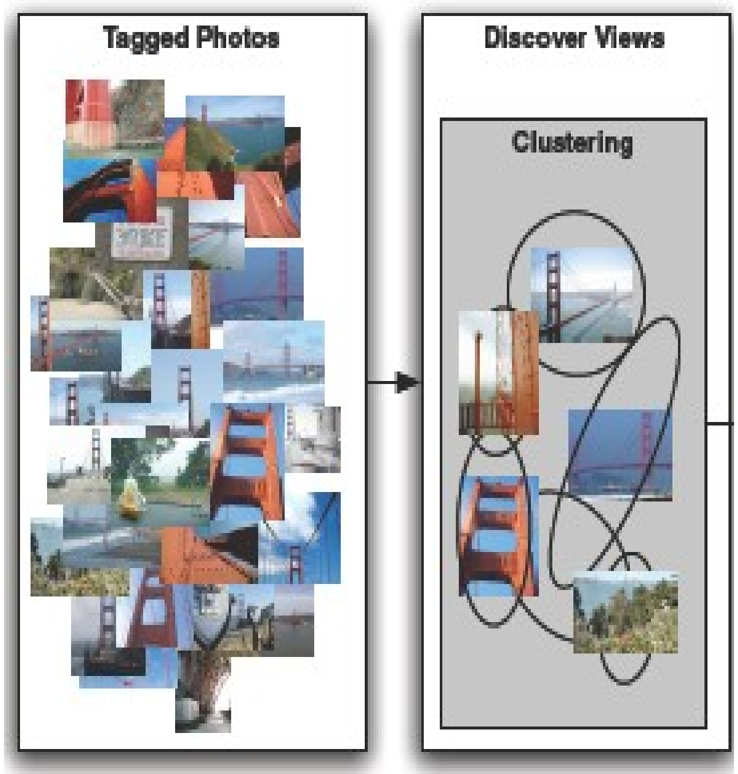


[Kennedy07]

Step 2: Visual clustering…

… or “*discover different views of the object of interest*”

- Extract visual features of the photos
  - Color (grid moment features)
  - texture (Gabor textures)
  - interest points (SIFT)
- Clustering based on visual features



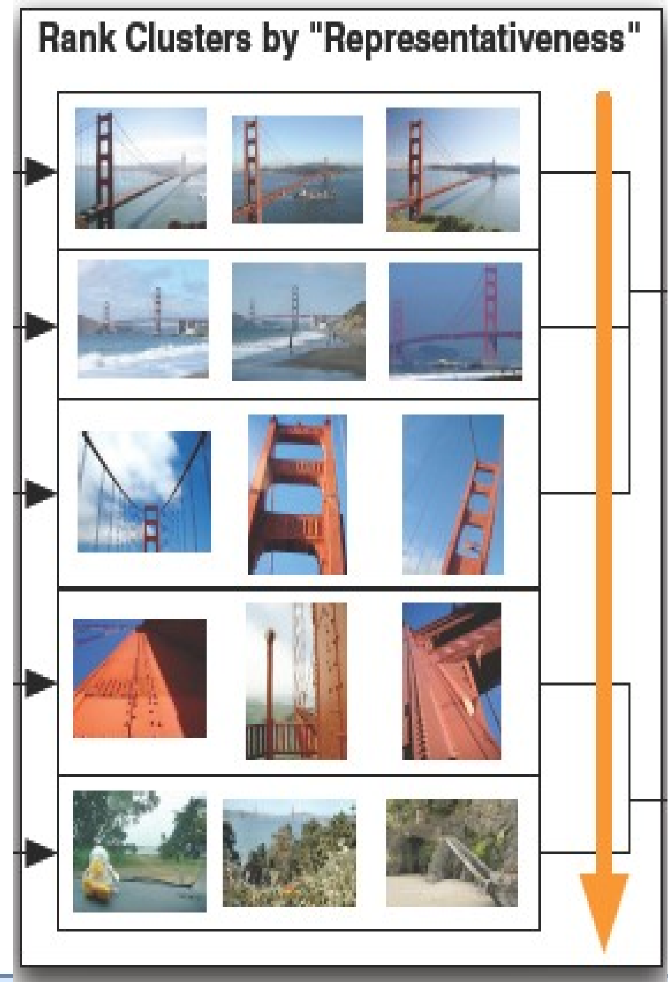
[Kennedy07]

Step 3: Ranking clusters of representative images

- Number of users
- Visual coherence
- Cluster connectivity
- Variability in dates

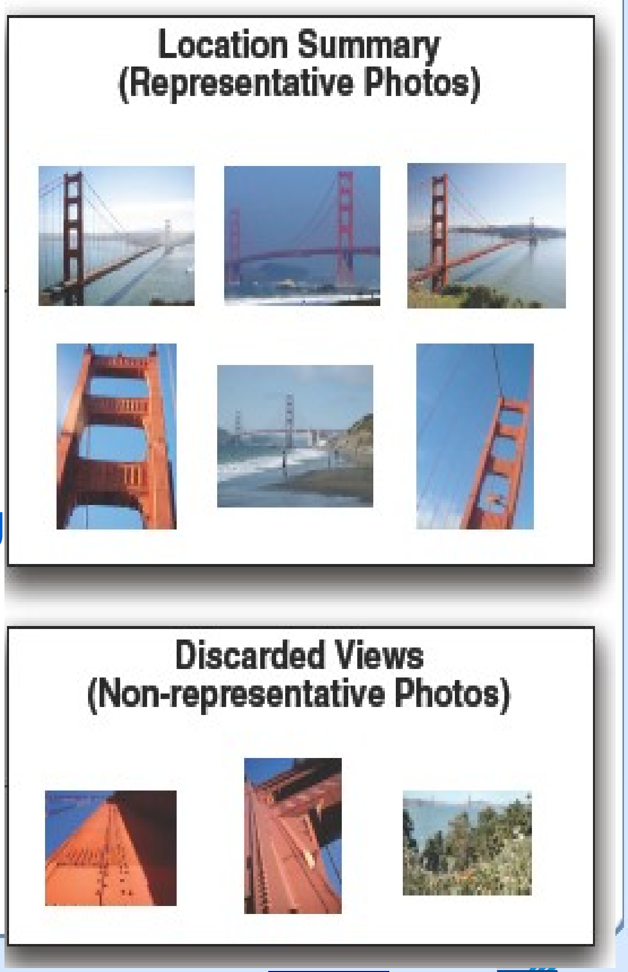
Geographic object
CASE STUDY

[Kennedy07]



Step 4: Ranking images of highly-ranked clusters

- Ranking images of highly-ranked clusters
  - **Low-Level Self-Similarity**
  - **Low-Level Discriminative Modeling**
  - **Point-wise Linking**
  - **Fusion of Ranking Methods**
- Sample the highest-ranking images (The lowest-ranking clusters have no images sampled from them, and the higher-ranking clusters have images sampled proportionally to the score of the cluster)



Geographic object
CASE STUDY

[Kennedy07]

Sample photo summaries of geographic objects

Photo summary for "Golden Gate Bridge"

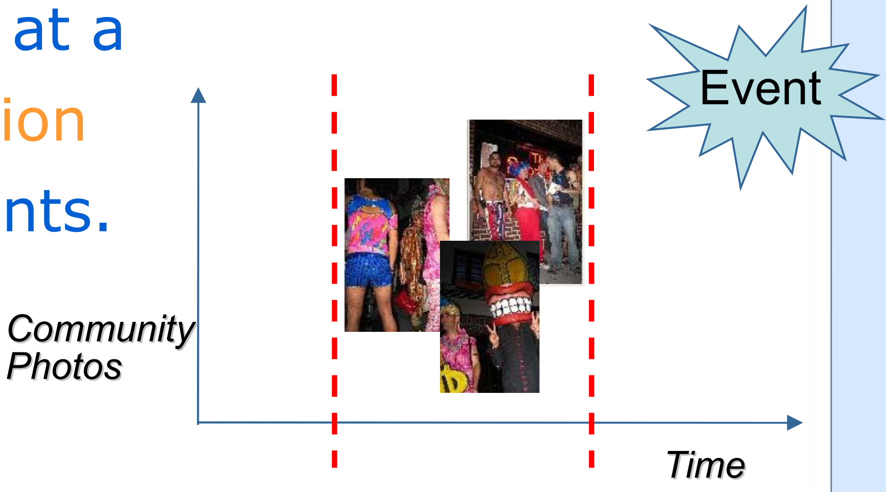
Photo summary for "Palace of Fine Arts"

DATASET: 110000 geo-referenced photos from the San Francisco area (downloaded from Flickr), 700 clusters

[Kennedy07]

Generating event photo summaries

- Idea: Photos that took place at a specific time and specific location are very likely to illustrate events.



- Suggestions:
  - Perform clustering and examine location, user and time metadata distributions on each cluster
  - Train an **ID3 decision tree,** to automatically **classify** clusters of photos into objects, events or none [Quack08]

Generating event photo summaries

[Quack08]

- Gather geotagged photos
- Hierarchical Clustering -- Dissimilarity matrix for various modalities
  - Visual features: SURF features, Euclidean distance between feature vectors, homography mappings
  - Text features: Combine tags, title, description into a single vector: (term weighting of term i into photo j)
- Classification to objects and events
  - Features: $f_1 = |D|$ $f_2 = \frac{|U|}{|N|}$, D = days, U = users, N = #photos in the cluster
  - Manually labeling of 700 clusters and training an ID3 decision tree

Sample photo summaries of events [Quack08]

DATASET: Divide the earth's surface into square tiles of 200m2
70000 geographic tiles
220000 geotagged photos from Flickr
After preprocessing, 73000 photos were assigned to clusters
Manually labeling of 700 clusters

"Oxford Geek nights"
"Movie premiere Italy"
"Exhibition gallery paris"

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

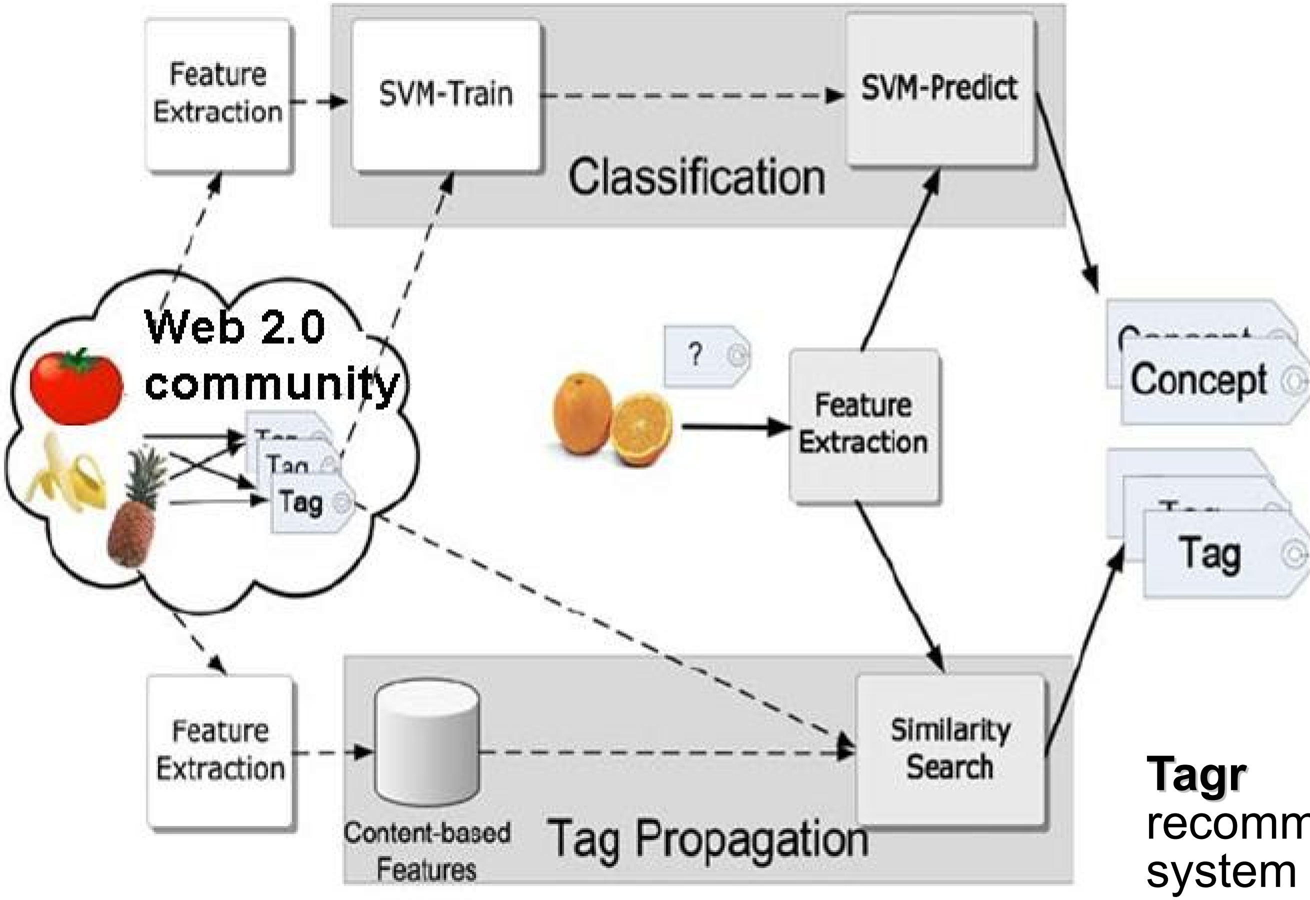
Generating recommendations

- **Idea:** Exploit clustering of multimedia content + content-based analysis algorithms and suggest annotations for unlabelled multimedia content

Benefits

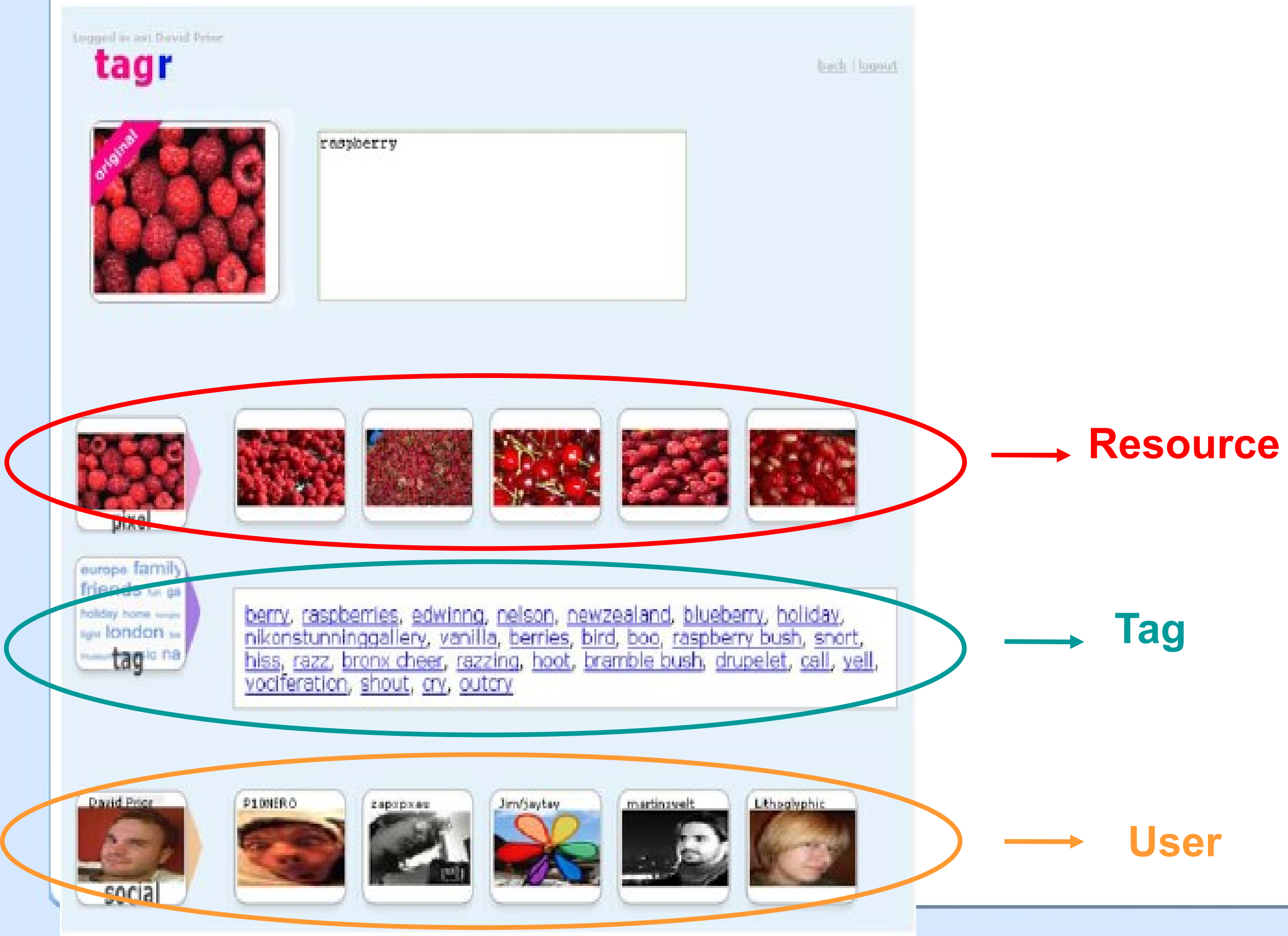
- Increase the chances a resource getting annotated, facilitating, thus, its retrieval
- Consolidate the vocabulary across various users

A tag recommendation system



Tagr recommendation system approach [Lindstaedt09]

Tagr interface [Lindstaedt09]



SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Using Tagr (sample annotation results)

Input	Similar images result	Propagated tags	Classified concept	Flickr tags
		fruit, food, fruits, vegetables, market, yellow, macro, 2006	banana	FoodTrails, homestyle, hom,bananas, fruits, wet, market, Punggol 21, a favourite
		red, fruits, fruit, food, market, strawberries, macro, tomato, summer, berries	strawberry	pike place market, seattle, june, 2007, fruits, produce, red, raspberries, perfect looking, fruit, sosio's, stand, food, m-p-g
		fruits, food, fruit, apple, red, macro, black and white, vegetable, delicious, blueberries	blueberry	friendship, flickrfriends, comments, photos, holidays, happydays, summer, fruits, more, blackberries, fairytale, allthebest, Godblessyou, loveyouall, [+ 16 more]

source [Lindstaedt09]

A tag recommendation and geo-location system

[Quack08]

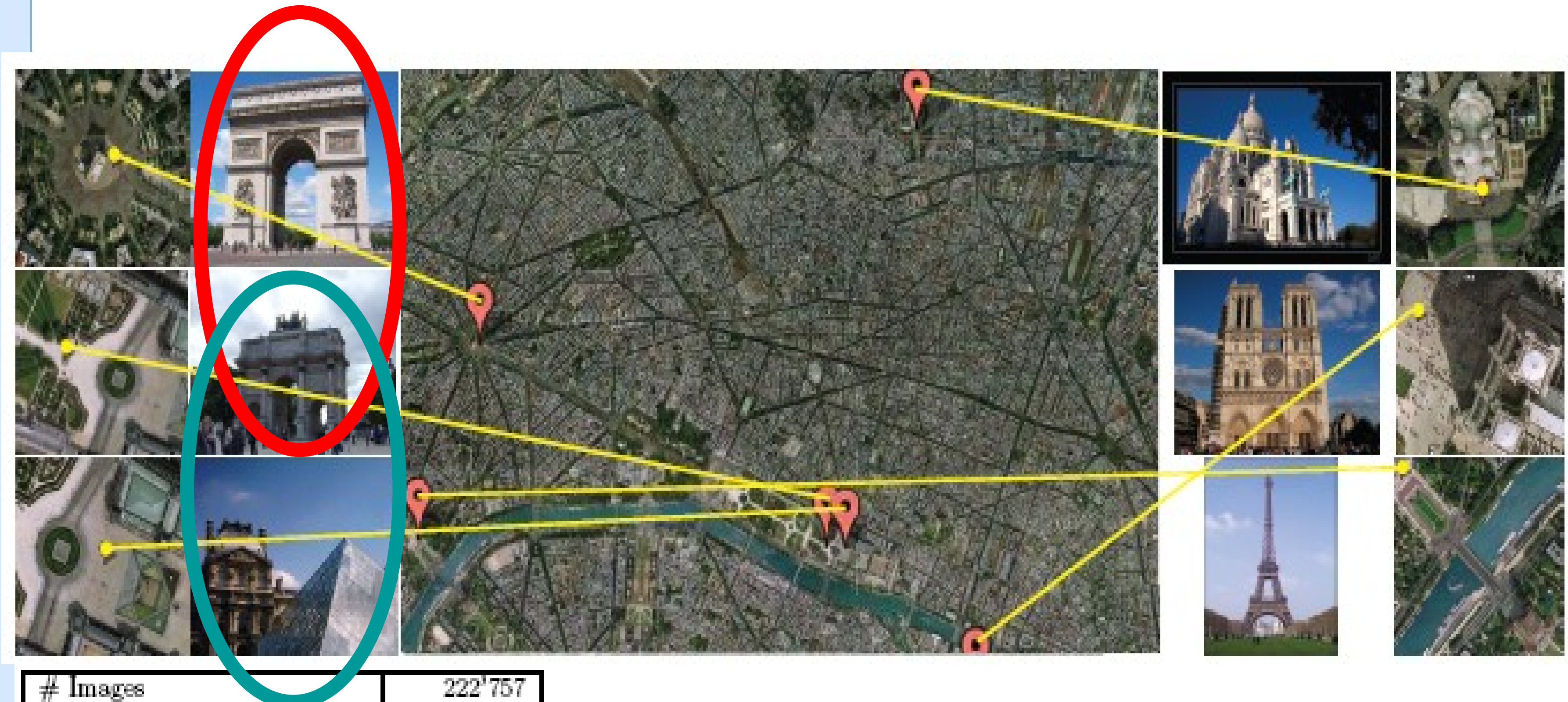
- Gather geotagged photos
- Hierarchical Clustering -- Dissimilarity matrix for various modalities
  - Visual features: SURF features, Euclidean distance between feature vectors, homography mappings
  - Text features: Combine tags, title, description into a single vector: (term weighting of term i into photo j)

Clustering

- Draw a rough bounding box (area) where the query image belong
- Match the query image to the clusters and get the best matching

Recommendation & geo-location

Experimental results (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]

Exploiting clustering for machine learning

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

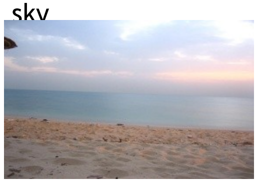
Problems:

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

Tagged images

sand, wave, rock, sky

sea, sand

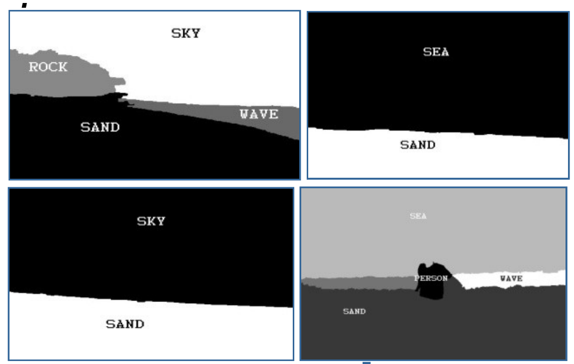


sand, sky

person, sand, wave, see

Social information + Computer Vision

Region-detail annotated images



[Chatzilari09]

Machine Learning

Object Detectors

Solutions:

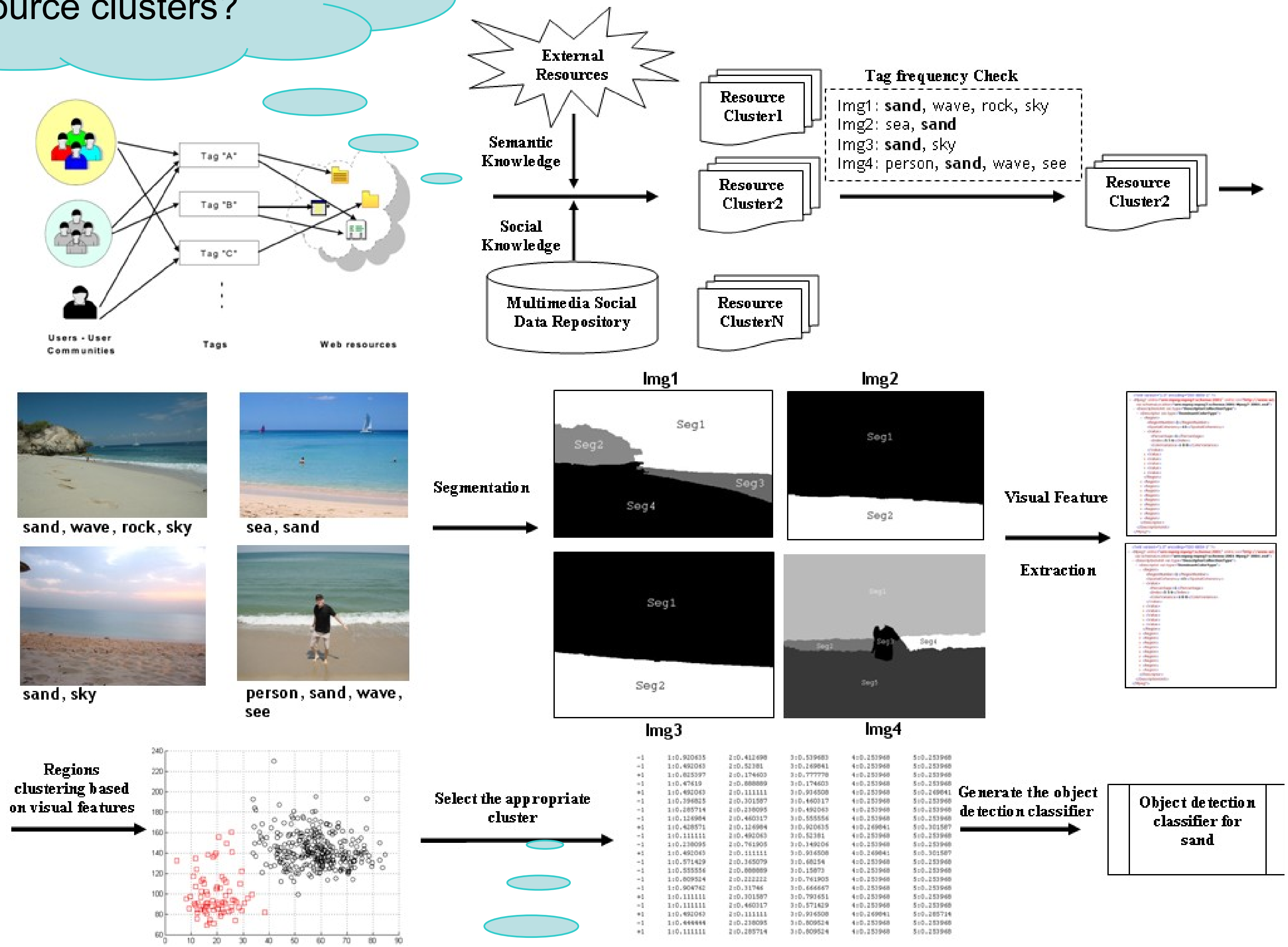
- Exploit user tagged images from social sites like flickr
- Combine techniques operating on tag and visual information space

weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Semantically coherent resource clusters?



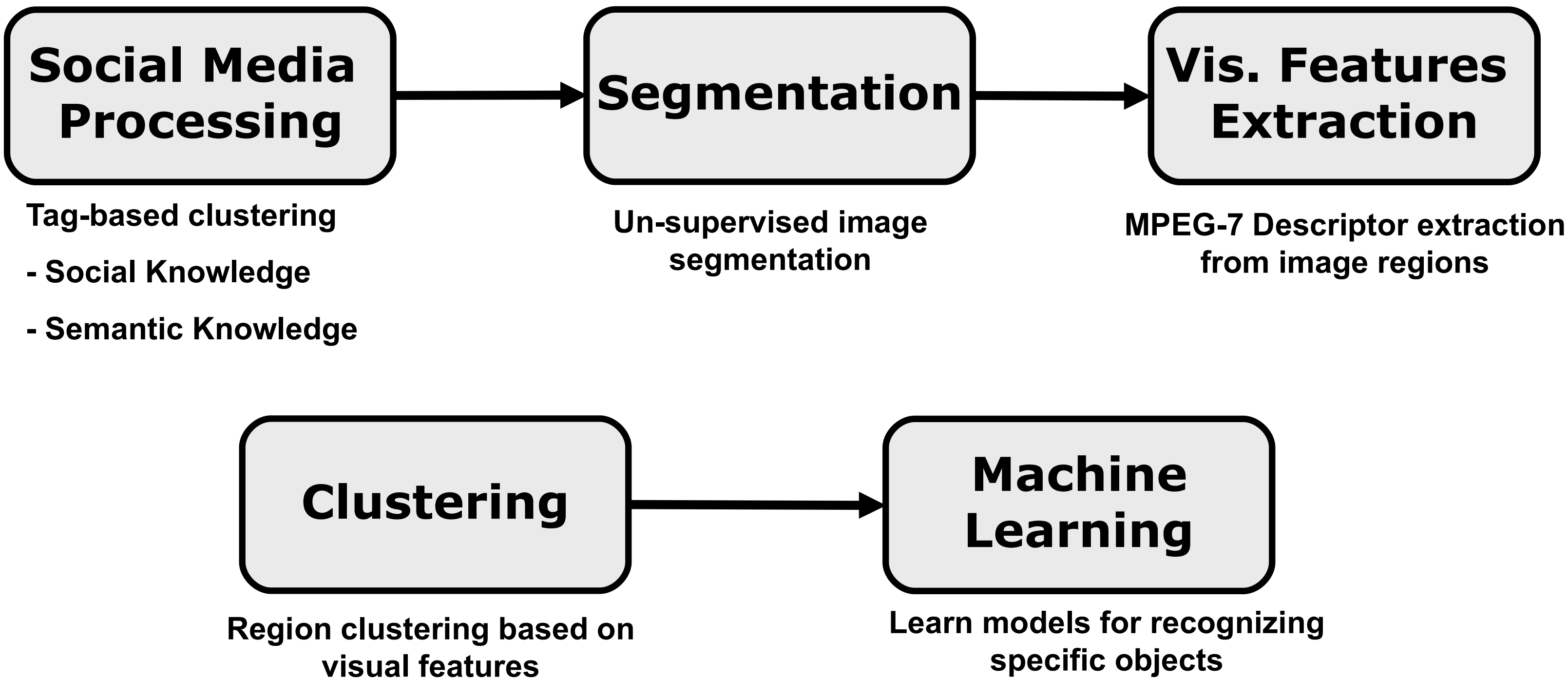
Appropriate Cluster Selection?

weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Framework Components



Semantically Coherent Resource Clusters - SEMSOC

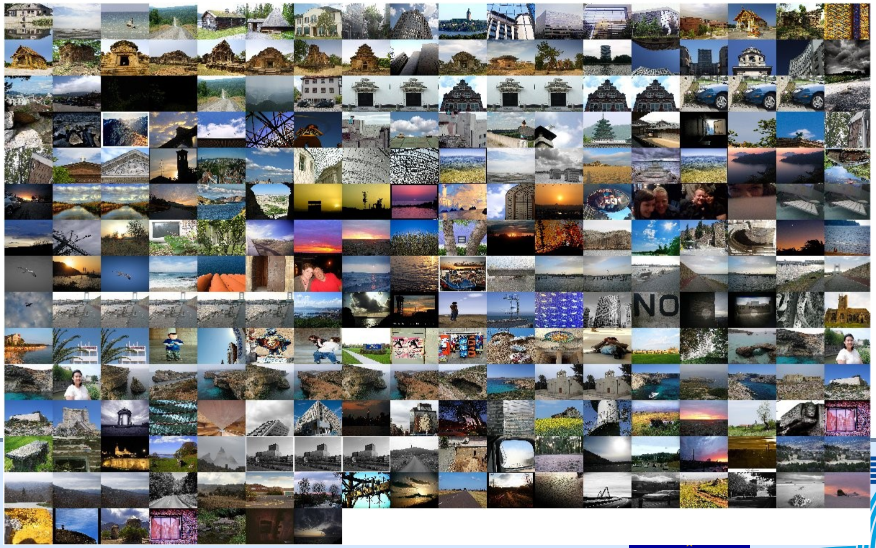
Vegetation

Tag-based processing:

- SEMSOC (Semantics Mining on Multimedia Social Data Sources)
- Jointly considers social and semantic features to cluster images

[Giannakidou08]

Sky



Output:

- Groups of images each one emphasizing on a particular topic

Semantically Coherent Resource Clusters - SEMSOC

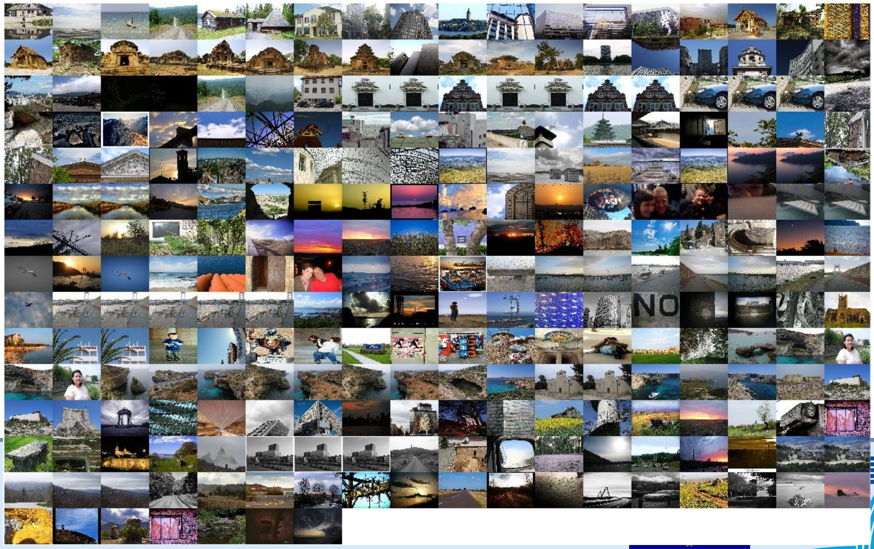
Vegetation

Tag-based processing:

- SEMSOC (Semantics Mining on Multimedia Social Data Sources)
- Jointly considers social and semantic features to cluster images

[Giannakidou08]

Sky



Output:

- Groups of images each one emphasizing on a particular topic

Segmentation & Visual Descriptors

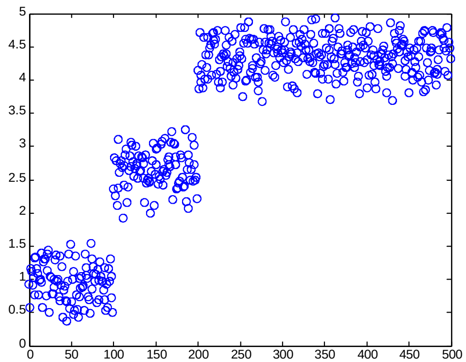
- Segmentation
  - K-means with connectivity constraint (KMCC)

[Mezaris et al., 2004

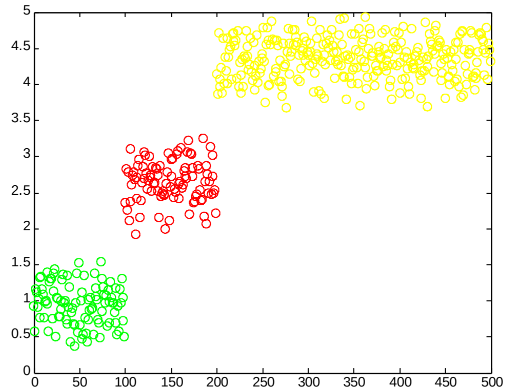
- Visual Descriptors
  - MPEG-7 standard
    - *Dominant Color , Color Layout, Color Structure, Scalable Color, Edge Histogram, Homogeneous Texture, Region Shape.*

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

Region-based Clustering & Cluster Selection

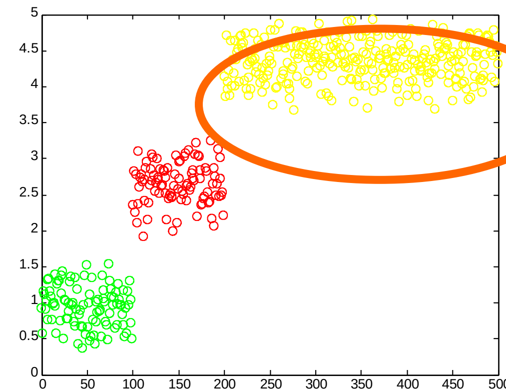


- Perform segmentation and visual feature extraction from all images in an image group (Identified by SEMSOC)



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

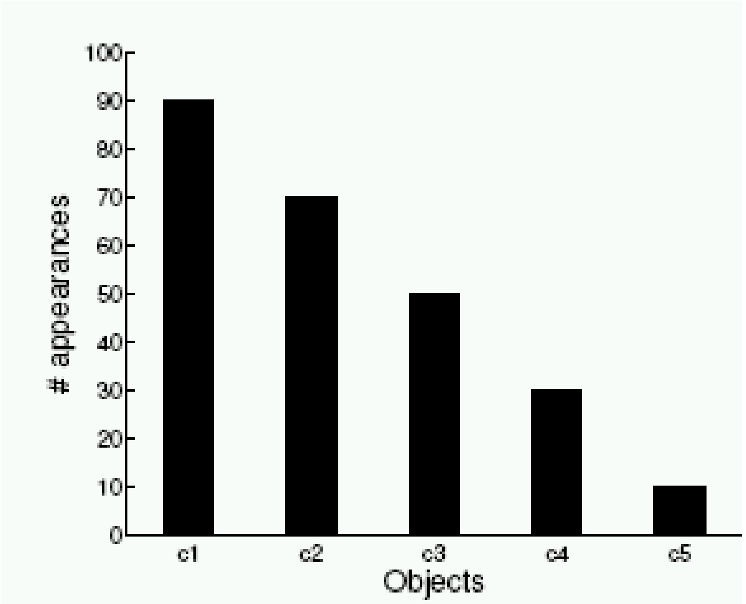
- Pick the most populated cluster as the one representing the most frequently appearing tag of the group



co-funded by the European Union

Frequency distribution of objects

Distribution of (actual) objects in image cluster based on their frequency rank

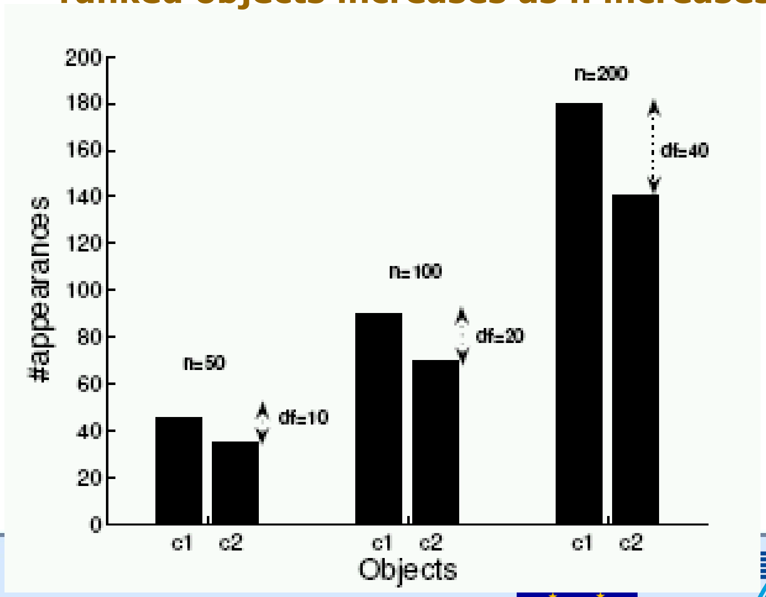


- c_1 is drawn with probability p_{c1} higher than p_{c2}, which is the probability that c_2 is drawn, and so for the remaining objects

- Assuming that the #appearances of each object follows a binomial distribution
- The #appearances of each object increases as a linear function of the dataset size n

$$E(K) = np$$

Absolute difference between 1st and 2nd most highly ranked objects increases as n increases



As Clustering Error Increases …

Perfect Case

error = 0

Most populated still the correct cluster

error > 0

Situation is reversed

error >> 0

The correct cluster is by far missed

error >>> 0

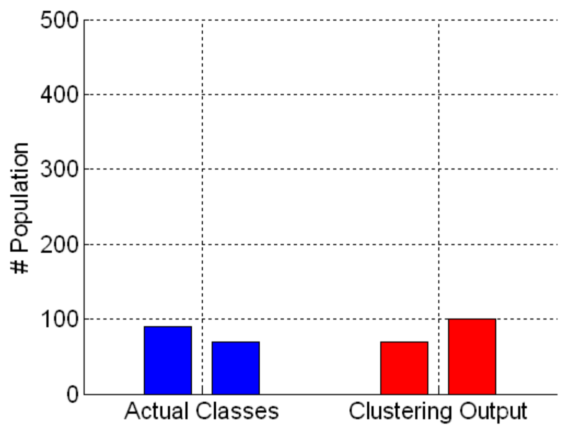
F I X E D D A T A S E T S I Z E

As Dataset Size Increases …

FIXED CLUSTERING ERROR

N = α

Problematic Case

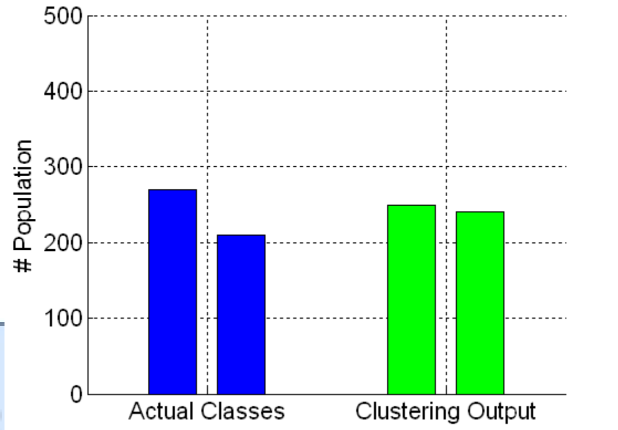


N > α

The gap is shortened

N >> α

Situation is reversed



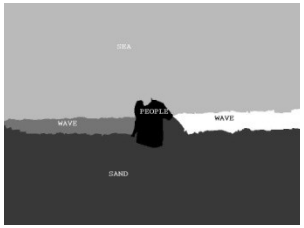
N >>> α

The correct cluster is easily identified

Experimental Setup

648 vacation Images

Fully annotated at region detail



3000 Flickr Images

Weakly annotated with tags

10000 Flickr Images

Weakly annotated with tags

sand, wave, rock, sky

Examined 4 concepts existing in all three datasets

- Sky
- Vegetation
- Sea
- Person

Train detectors using manually provided region detail annotations

Object Detectors

Object Detectors

Train detectors using region-detail annotations obtained with the proposed approach

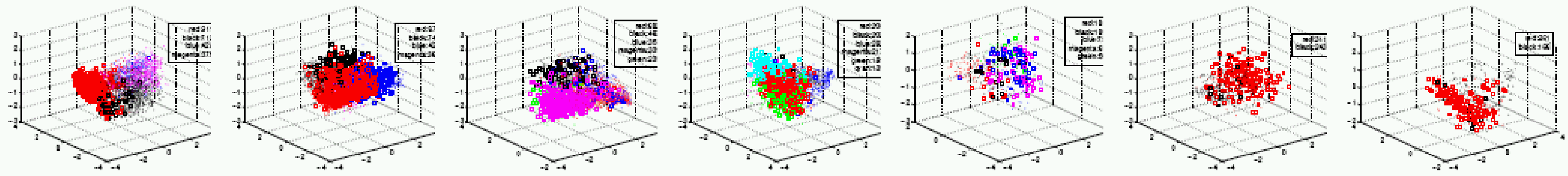
Compare Performance

weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Experimental Results – Cluster Selection



(a) Sky (b) Sea (c) sand (d) Person (e) Boat (f) Vegetation (g) Rock

Setting:

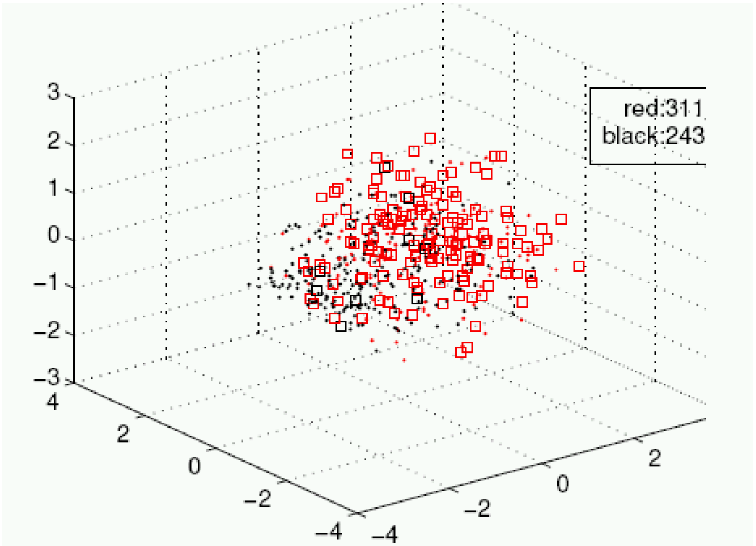
- Visual the way regions are distributed among clusters
- Use shape-code (squares) to indicate the regions of interest and color-code to indicate a cluster's rank (largest cluster: red)
- Ideally all squares should be painted red and all dots should be painted differently

Goal:

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

Conclusions:

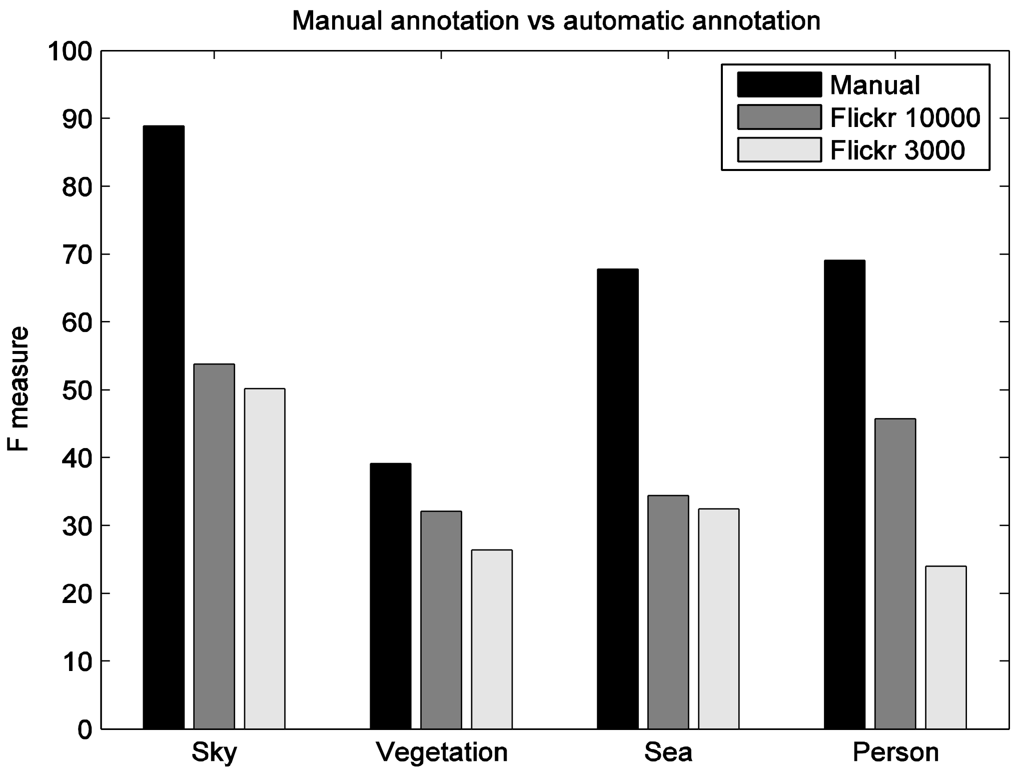
- Our claim is valid in 5 (i.e., sky, sea, person, vegetation, rock) and not valid in 2 (i.e., boat, sand) cases



Vegetation in magnification

weknowit
SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Experimental Results - Man. vs Autom. trained object detectors



Observations:

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

Future work:

- Test the framework for more concepts
- Exploit more of the user contributed information (e.g., Flickr groups)

Clustering Conclusions

- Tag co-occurrence, semantic similarity of tags and content-based similarity of resources are useful indicators of IR in a social tagging system.
- Tag ambiguity, lack of structure and tag spamming can be sufficiently tackled.

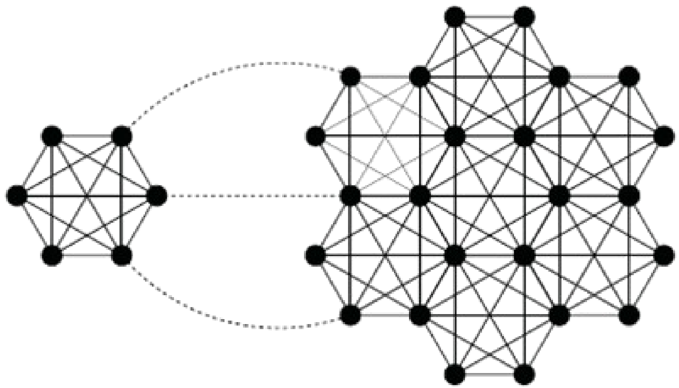
- **Other Use Cases**
- Inducing ontology from Flickr tags (crawling, clustering, relationship extraction)
- Domain Ontology enrichment
- Social assisted analysis
- User profiles
- Recommendations
- Trend detection

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Community Detection

Community Detection in Complex Networks

- Community Detection: The Problem
- Global vs. Local Community Detection
- Bridge Bounding
- Conclusions - Future Work



Communities in formal terms…

Communities on the Web are groups of individuals who share a common interest, together with the Web pages most popular amongst them *[Kumar99]* … explicit or implicit ?

Web community is a set of Web-based objects (documents and users) that has its own logical and semantic structures, such that information retrieval and Web-data management is facilitated *[Zhang06]* . e.g. a Web page set with clusters in it is a community; web pages in a set that are related to a given Web page also form a community;

Communities are groups of vertices which probably share common properties and/or play similar roles within the graph, e.g. groups of Web pages dealing with related topics *[Fortunato07a]*

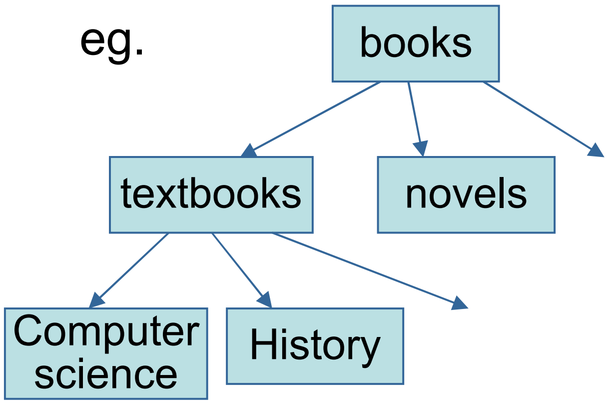
… two different types of communities

easily identified

- explicitly-defined communities

well-known group of web pages sharing a common interest e.g. Yahoo

eg.



(graph) analysis required

- implicitly-defined communities

non-obvious; hidden or unexpected;larger and outnumber the explicit ones; may appear as an emerging Web community for some specific topic or event;

eg. group of web pages for mediterranean cooking

focus on implicitly-defined communities

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

… in other terms

- *managerially coordinated communities:* they have a central authority or process (e.g. a creator) that governs the formation mechanism.
  - e.g. Google groups, LinkedIn, Drupal groups, Facebook, etc
- *self-organized communities:* they emerge from the interaction patterns between the members, i.e. highly related members are identified as a community.
  - e.g. Flickr clusters

communities context …

- typically … communities are defined with reference to some graph (network) which represents a set of entities / objects (nodes) and their relations (edges).

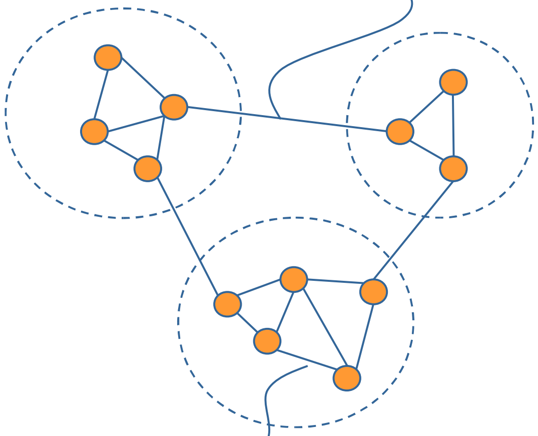
… even when there is no explicit graph, one can infer it, e.g.:

feature vectors → distances →threshold application → graph

- Given a graph, a community is loosely defined as a set of nodes that are more densely connected to each other than to the rest of the graph vertices.

a simple example …

- extremely profound community structure.
- key-concepts : within-community nodes, intra-community edges, inter-community edges.
- rarely appearing in real systems.



Definition of communities is heavily dependent on graph properties and subgraphs discovery

graph theory / graph analysis basics

- vertex/node & edge-level
- graph groups level (clique, k-cores, motifs,…)
- community level

co-funded by the European Union

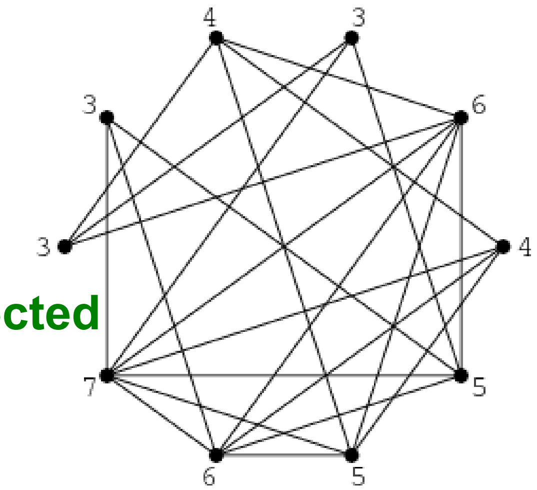
vertex & edges indicate degrees

- **Degree of a graph vertex v**: the number of graph edges which touch v.

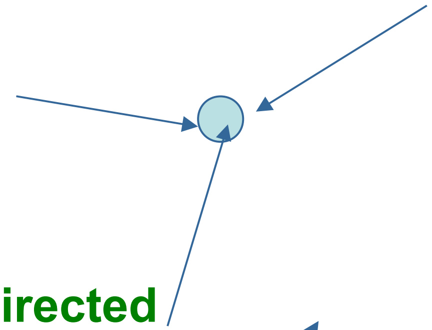
- **Indegree of a graph vertex v**: the number of inward directed graph edges from a given graph vertex in a directed graph

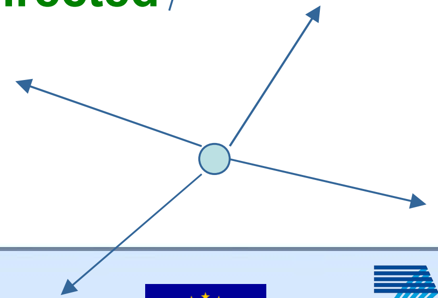
- **Outdegree of a graph vertex v**: The number of outward directed graph edges from a given graph vertex in a directed graph.

undirected



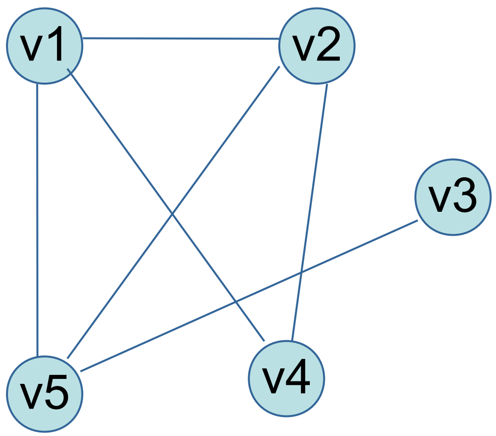
directed





Degrees & adjancencies (I)

Adjacency matrix on an undirected graph : A(i,j), i,j <= n



	v1	v2	v3	v4	v5
v1	0	1	0	1	1
v2	1	0	0	1	1
v3	0	0	0	0	1
v4	1	1	0	0	0
v5	1	1	1	0	0

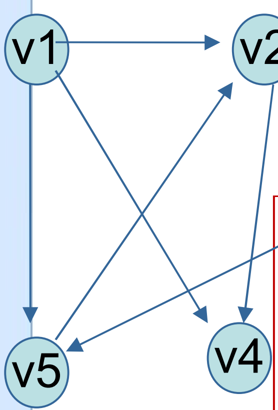
degree of a vertex v
(number of edges incident upon it:

$$k_v = \sum_w A(v,w)$$

The probability of an edge existing between vertices v & w if connections are made at random but with respecting vertex degrees is $\frac{k_v k_w}{2m}$ since $m = \frac{1}{2} \sum_i k_i$

Degrees & adjancencies (II)

Adjacency matrix directed graph

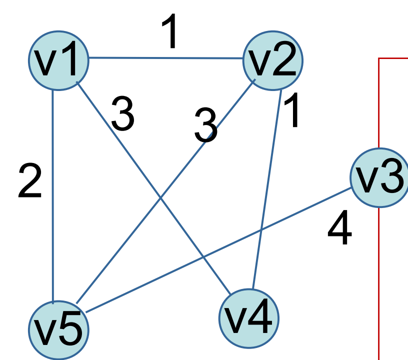


	v1	v2	v3	v4	v5
v1	0	1	0	1	1
v2	0	0	0	1	0
v3	0	0	0	0	1
v4	0	0	0	0	0
v5	0	1	0	0	0

degree of a vertex v
number of edges incident upon it:

$$k_v = \sum_w A(v, w)$$

Adjacency matrix weighted graph



	v1	v2	v3	v4	v5
v1	0	1	0	3	2
v2	1	0	0	1	3
v3	0	0	0	0	4
v4	3	1	0	0	0
v5	2	3	4	0	0

degree of a vertex v
(sum of edges' weights incident upon it):

$$k_v = \sum_w A(v, w)$$

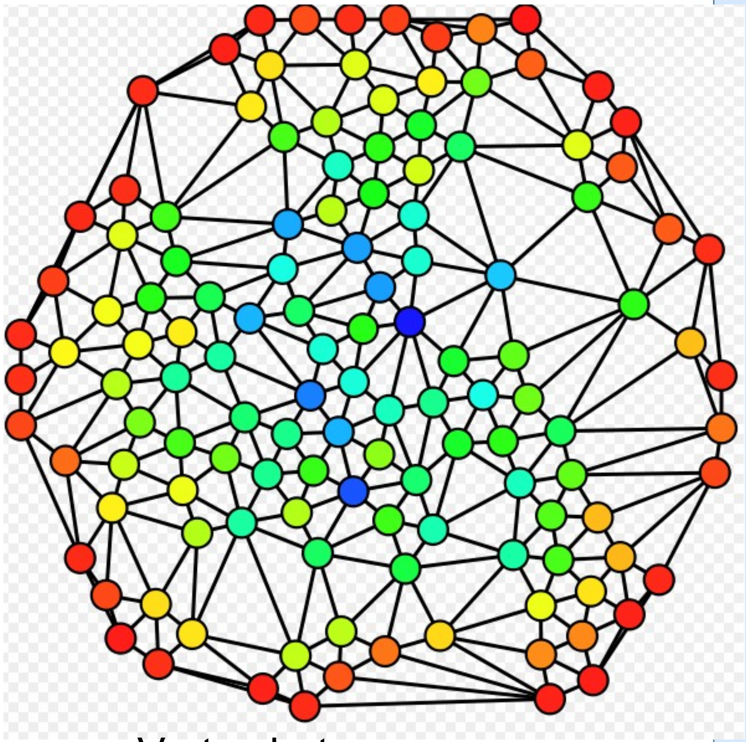
weknowit

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Vertex centrality & betweeness

Vertex betweenness has been studied in the past as a measure of the centrality and influence of nodes in networks.

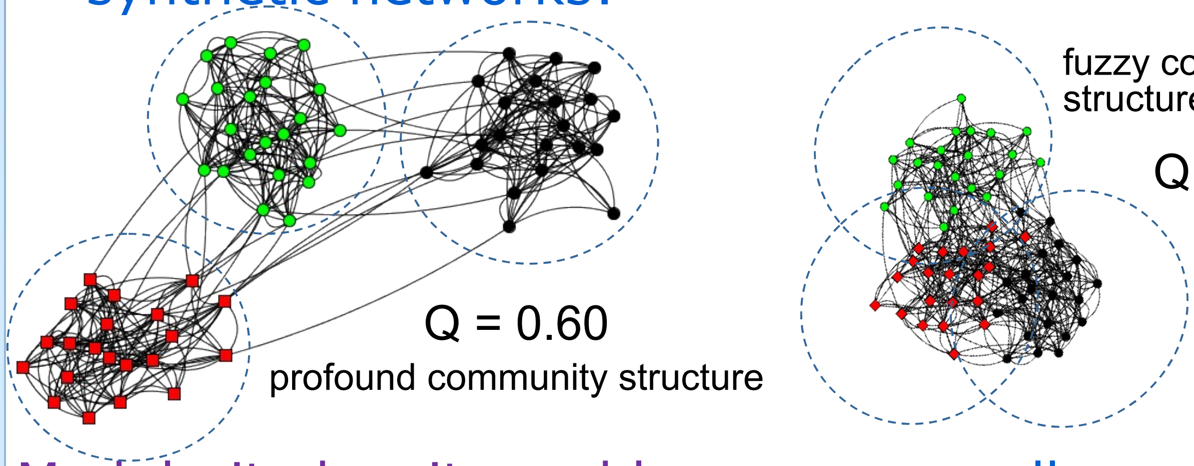
Vertex betweenness is a **centrality** measure of a vertex within a graph based on the idea that vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not [Freeman77].



Vertex betweeness
red=0 to **blue=max**
source wikipedia
File:Graph betweenness.svg

modularity and values

- Modularity is the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random
- Examples of modularity values for synthetic networks:



Modularity has its problems, e.g. small communities may be missed by modularity maximization [Fortunato07a].

- Q=0 number of within-community edges no better than random;
- Q=1 is the maximum and it indicates strong community structure;
- typical values between Q=0.3 to Q=0.7

Problem Statement

- No common definition of community.
- Some definitions:

  A community is a group of vertices with:

  - more edges among them than $\sum_{v \in C} w_{uv} \geq \sum_{v \in V-C} w_{uv}$ *for all* $u \in C$.

between them and the rest of the graph,

  - high *modularity*, $Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$
  - high *conductance*. $\phi(S) = \frac{\sum_{i \in S, j \notin S} A_{ij}}{\min\{A(S), A(\overline{S})\}}$

- In any case, the output of a community detection process on a graph is a set of vertex sets.

weknowit

co-funded by the European Union

SEVENTH FRAMEWORK PROGRAMME

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

Bridge Bounding

Algorithm

- Start a community with a seed node
- Add neighbouring nodes as long as they are connected by edges that are not inter-community (“bridges”).
- Stop when it is not possible to add any more nodes.

Algorithm 1 LocalCommunityDetection

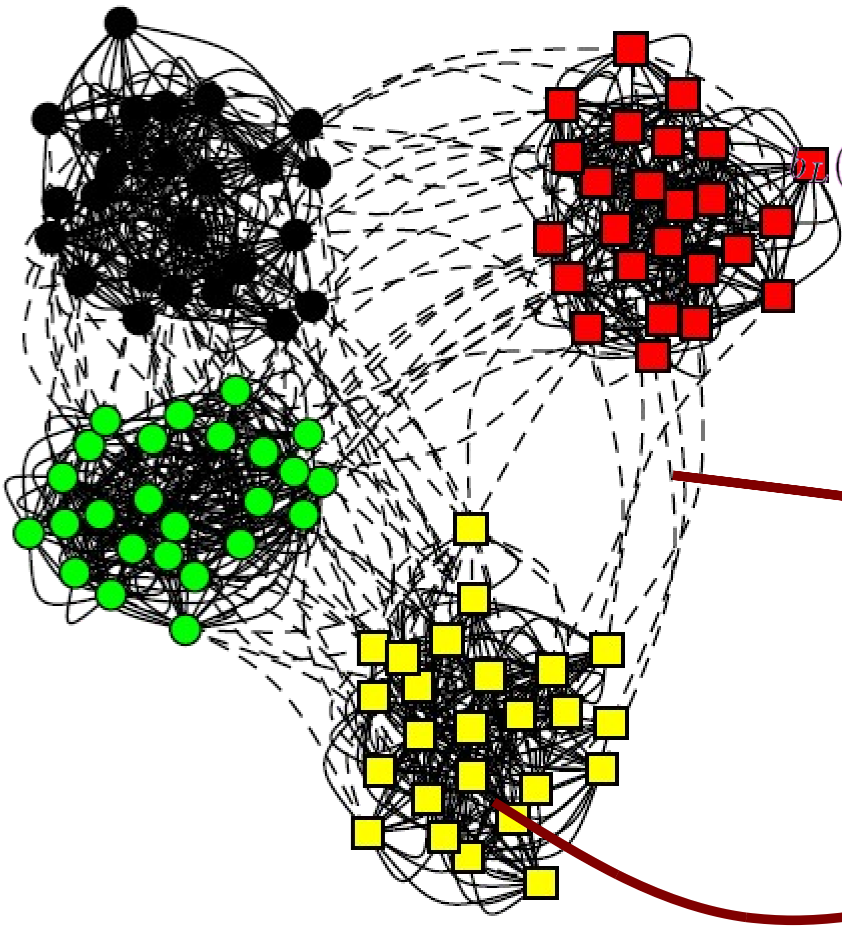
Require: Seed node $s \in G = (V, E)$
Require: Community mapping $g_C : V \rightarrow \mathbf{P}$
Require: Bridge function $b : E \rightarrow [0.0, 1.0]$

$C_s = \oslash$
Frontier set $F = \{s\}$
while $|F| > 0$ **do** $\{F$ is non-empty$\}$
  $c \leftarrow F.\mathrm{pop}()$
  $C_s \leftarrow C_s \bigcup \{c\}$
  $C_U \leftarrow C_U \backslash \{c\}$
  for all $n \in N(c)$ such that $e_{cn} = (c, n) \in E$ **do**
    if $g_C(n) = C_U$ and $b(e_{cn}) \leq B_L$ **then**
      $F.\mathrm{push}(n)$
    end if
  end for
end while
$\mathbf{P} \leftarrow \mathbf{P} \bigcup C_s$

Basic success factor:

Edge Bridge-ness: The property of an edge to lie between two communities.

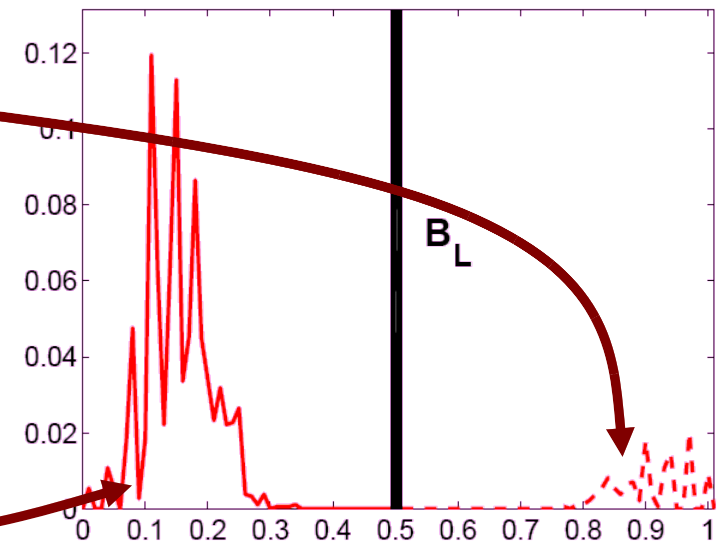
Bridge Bounding – Toy Example



Local bridging of an edge

$$B_L(e_{st}) = 1 - C_{st}^{(3)} = 1 - \frac{|N(s) \bigcap N(t)|}{min[(d(s)-1),(d(t)-1)]}$$

s, t: endpoints of edge
N(s), N(t): neighbourhoods of s, t
d(s), d(t): degrees of s, t



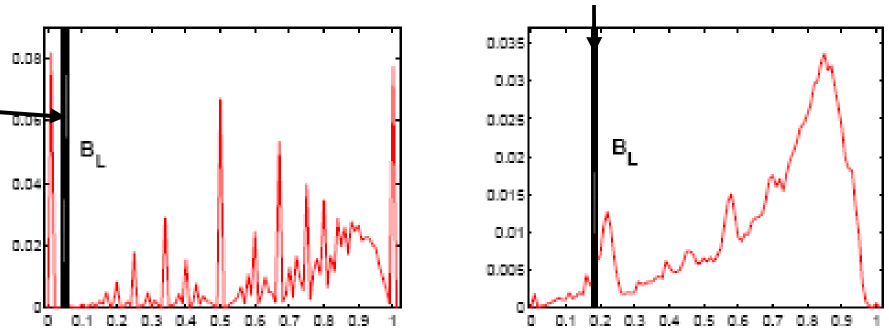
Bridge Bounding - Problems

- Local bridging not suitable for scale-free networks
- Solution (partial) 2[nd] order local bridging.

$$b'_L(e_{st}) = \alpha \cdot b_L(e_{st}) + (1-\alpha)\frac{1}{|N(e_{st})|}\sum_{e \in N(e_{st})} b_L(e)$$

B_L = 0.17 leaves just 1% of edges as non-bridges.

B_L as low as 0.05 leaves 8% of edges as non-bridges.

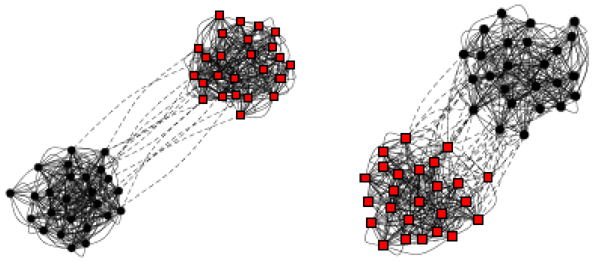


(a) b_L distribution (b) b'_L distribution, $\alpha = 0.7$

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) $p_{out} = 0.08$

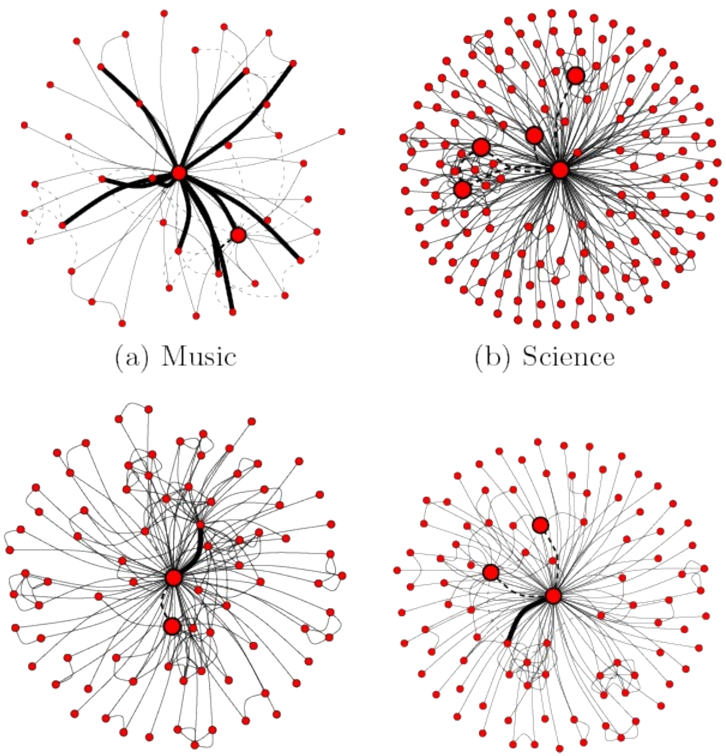
Change complexity of underlying communities.

	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.

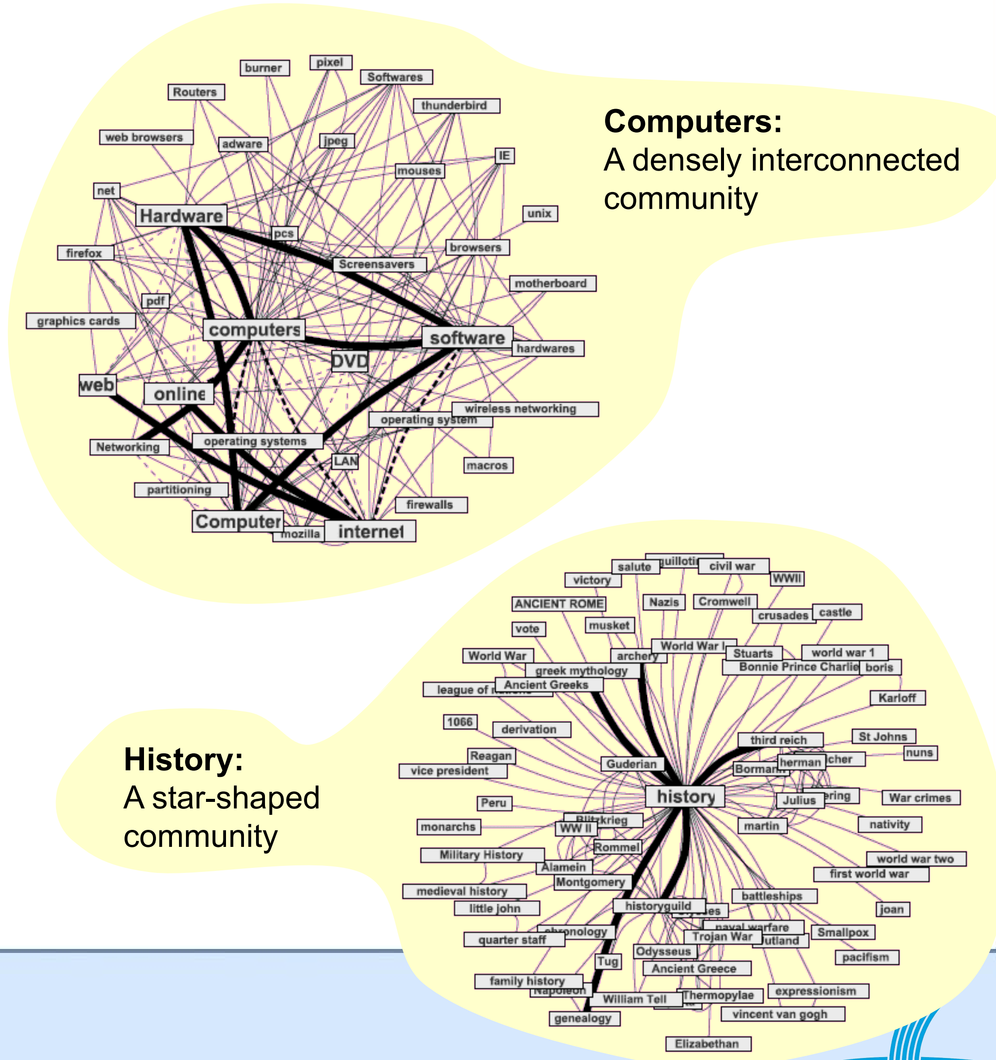
	F_C			NMI		
s_{var}	BB	BB'	GN	BB	BB'	GN
1.1	100	100	100	1.0	1.0	1.0
1.5	100	100	100	1.0	1.0	1.0
1.6	99.5	100	100	0.99	1.0	1.0
1.7	88	98	100	0.82	0.96	1.0
1.8	85.5	97	100	0.79	0.95	1.0
1.9	58.5	87	90	0.68	0.82	0.88
2.0	12.5	80	82	0.45	0.73	0.81
2.5	0	62	75	0.45	0.63	0.72

LYCOS iQ Tag Network



(a) Music (b) Science

(c) Film (d) Animals



weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

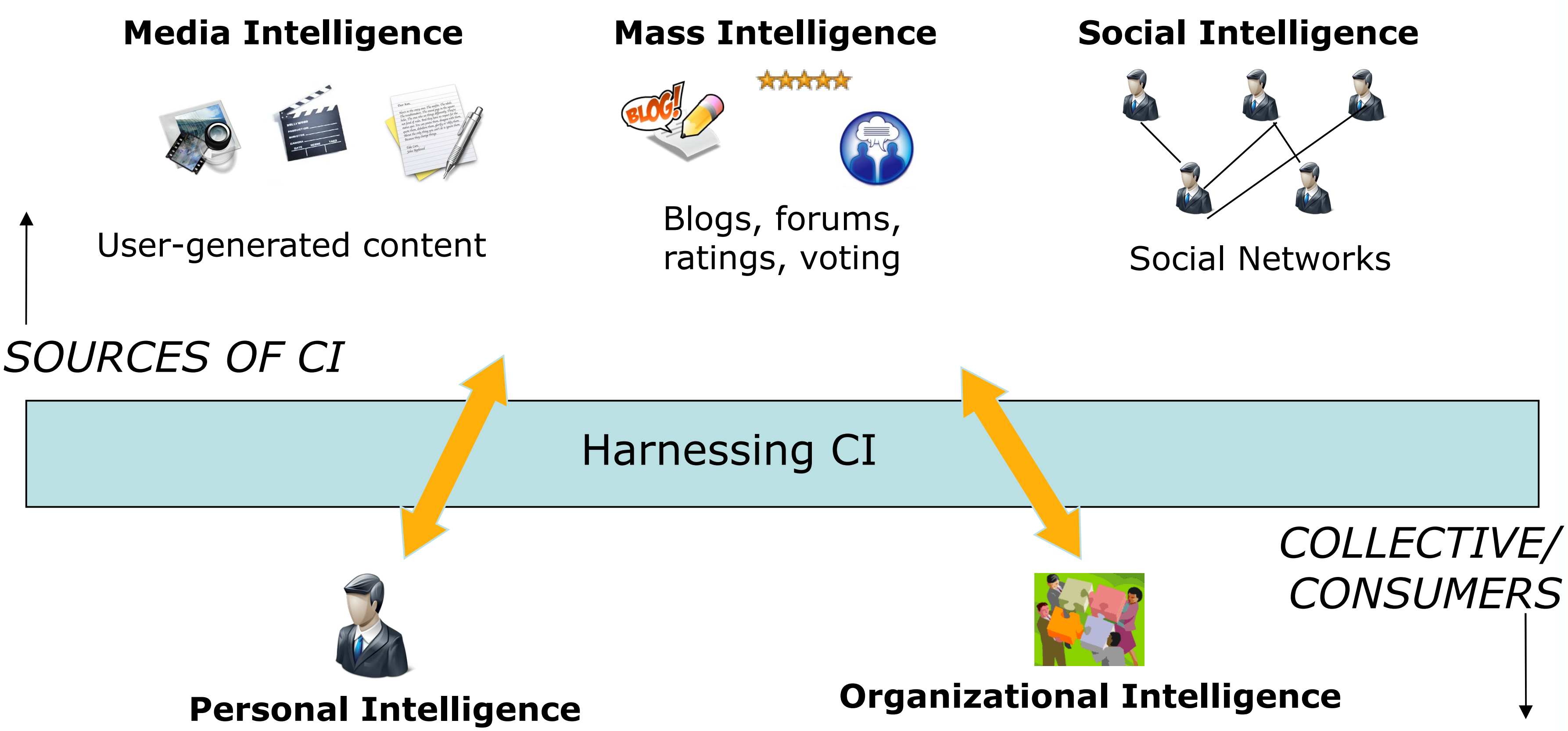
Future Work for Community Detection

- Remove ad-hoc parts of the algorithm:
  - Selection of B_L threshold.
  - Heuristics for artificially stopping community building process (e.g. co-occurrence frequency)
- Compare with other methods.
- Evaluate on real networks.
- Other applications

WeKnowIt and CI

WeKnowIt and CI

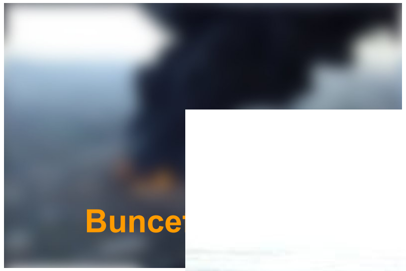
Decomposition of Collective Intelligence



weknowit

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Personal Intelligence



Profile of contributor

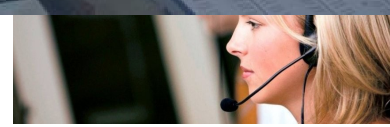
>> What to send where,
e.g. location, age

Media Intelligence

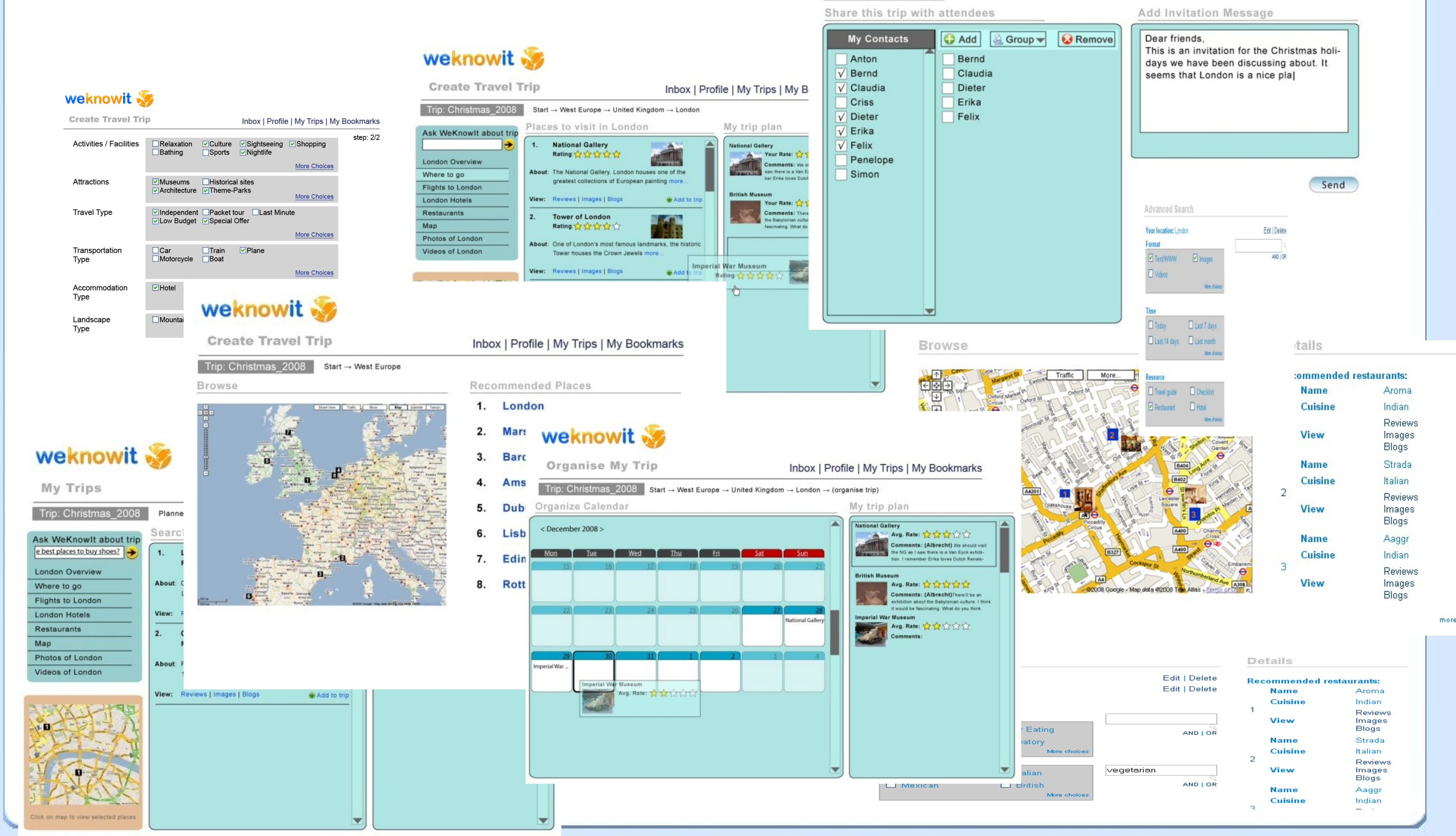
Organizati

Trust and feedback

>> Determine trustworthiness and hub-structures by SNA



Access: mock-ups



co-funded by the European Union
SEVENTH FRAMEWORK PROGRAMME

Harnessing CI @ WeKnowIt

Media Intelligence

"Knowledge and information extraction from raw content in conjunction with contextual information, personal and social context"

Intelligent Content Analysis: fusing information from

diverse modalities (video/image, audio, text)

contextual information (location, time)

personal Context (user profile)

+ **social Context** (friends, communities, tags, related items)

= fusion task, semantic analysis of content

A Source of Collective Intelligence

Harnessing CI @ WeKnowIt

Mass Intelligence

"is recognition and understanding of facts and trends by exploitation of massive user contributions"

Sources of Collective Intelligence

Blogs (comments)

Forums (threads/discussions)

Ratings

Questions & Answers

weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Social Network Analysis

- Individuals (actors) are not isolates regarding their actions. They always act within the possibilities and constraints given by their social environment
- Examples
  - Smoking in groups of high school kids
  - Fashion
  - Trading at the stock market
- Interactions are modelled as networks
- Methods from such fields as graph theory, mathematics, physics, sociology, social psychology are used to analyze these networks

CI Issues

- Trust, security, privacy, wrong data
- Applications and commercialization
- Integration with services - organizations
- Efficiency of semantics and analysis
- Real integration
  - not just sum of different analysis
  - formal framework and approach
  - representation
- User interaction – Interfaces, functionalities
- Performance, scalability
  - speed, storage, power

Thank you!

weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

CERTH-ITI

http://mklab.iti.gr

Symeon Papadopoulos, PhD student
Eirini Yiannakidou, PhD student
Spyros Nikolopoulos, PhD student
Elisavet Chatzilari, Researcher

Prof. Athena Vakali, AUTH

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.