SOCIAL NETWORKS : EVOLVING DATA MINING and SENTIMENT ANALYTICS

Prof. Athena Vakali, Informatics Dept, Aristotle University

WIMS 2014, Thessaloniki, June 3rd 2014

Presentation Outline

- Web 2.0 facts and social data
- Social associations and all kinds of graphs
- Evolving social data mining
- Emotion-aware social data analytics
- Frameworks and Applications

- Web 2.0 facts and social data
  - evolution & characteristics
  - is there hidden information there?
  - motivation for social (evolving) data mining
- Social associations and all kinds of graphs
- Emotion-aware social data analysis
- Frameworks and Applications

Web 2.0 facts & social data

- Web 2.0 has become a source of vast amounts of social data evolving at fast rates
- Users participate massively in Web 2.0 applications such as:
  - social networking sites (e.g. facebook)
  - blogs, microblogs (e.g. Mashable All That's New on the Web twitter)
  - social bookmarking/tagging systems (e.g. del.icio.us flickr)
- Social data refer to different types of ***entities***
  - users
  - content
  - metadata

  handled separately or jointly

associated via various types of ***interactions*** / ***relationships***

SOCIAL DATA SOURCES: is there hidden information here?

- Web 2.0 users act **explicitly** by declaring their associations

user-user	user-resource	user-metadata	user-group
“is friend with”	*“likes”*	*“creates tag”*	*“belongs to”*

- Relationships may be **multipartite** (e.g. user u_1 making a comment on the post p of user u_2) **but** are usually simplified in **bipartite** associations between some involved entities
- However, explicit associations generate **implicit** threads of relationships which:
  - are triggered by users’ common activities
  - may hold among the non-user entities (resources, metadata)

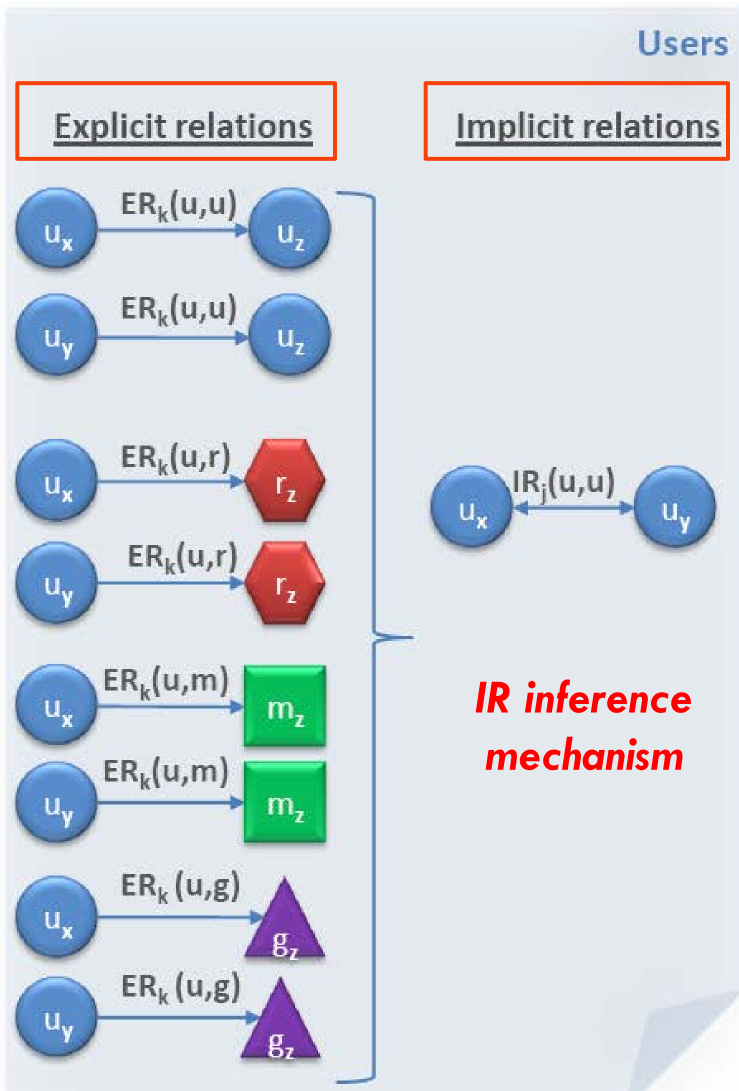
Web 2.0 relations initiated by users' online behavior

i	$ER_i(u,u)$
1	follows
2	is followed by
3	is friend with
4	comments on post of

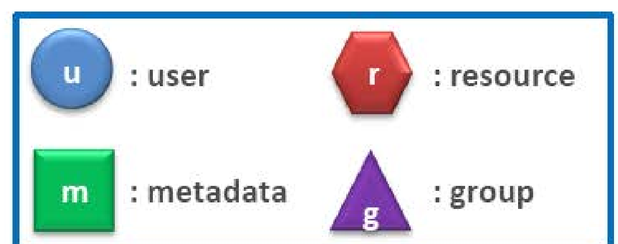
i	$ER_i(u,r)$
1	likes
2	uploads
3	leaves comment on
4	rates
5	downloads
6	assigns tag to

i	$ER_i(u,m)$
1	creates (tag)
2	assigns to group (tag)
3	assigns on resource (tag)
4	makes on resource (comment)

i	$ER_i(u,g)$
1	creates
2	belongs to
3	assigns (tag) to

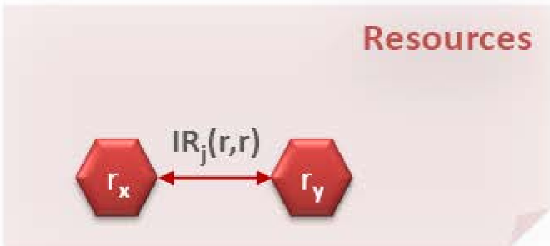


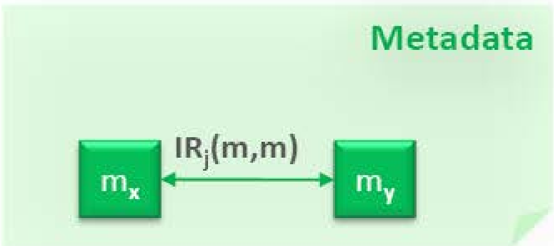
j	pivot	$IR_j(u,u)$
1	u	followers of same user
2	u	are followed by same user
3	u	friends of same user
4	u	comment on post of same user
5	r	like same resource
6	r	upload same resource
7	r	download same resource
8	r	comment on same post
9	r	like same post
10	r	assign tag on same post
11	m	use same tag
12	m	assign same tag to group
13	m	leave semantically related comments
14	g	belong to same group
15	g	assign tags to same group

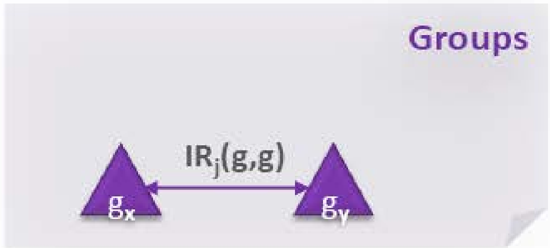


Implicit relations uncovered for various types of entities

Implicit relations … *generation mechanisms*







j	pivot	$IR_j(r,r)$
1	u	are liked by same user
2	u	are rated by same user
3	u	are uploaded by same user
4	u	are downloaded by same user
5	u	are assigned tag by same user
6	m	are assigned the same tag

j	pivot	$IR_j(m,m)$
1	u	are assigned by same user (tags)
2	u	are left by same user (comments)
3	r	are applied on same resource
4	g	are assigned to same category

j	pivot	$IR_j(g,g)$
1	u	are created by same user (tags)
2	u	have same member (user)
3	r	have same member (tag)

- **Numerous relations might unfold …** however only some of them are selected based on analysis' focus or application's context
  - user-user IRs such as: "*like same resource*", "*use same tag*" can be leveraged for studying behavioral patterns (e.g. in applications like Flickr)
  - tag-tag IRs such as: "*are assigned on same resource*" can be leveraged for tag clustering

Motivation for Social Data Mining

- The availability of massive sizes of data gave new impetus to data mining.
  - by the end of 2013, Facebook boasted **1.23bn monthly active users** worldwide, adding 170m in just one year; **300 million photo uploads** daily ! [Facebook Statistics 2013]
- Mining social web data can act as a barometer of the users' opinion. Non-obvious results may emerge.
  - Collaboration and contribution of many individuals formation of *collective intelligence*
  - *Wisdom of the crowd*: more accurate, unbiased source of information.
- Social data mining results can be useful for applications such as recommender systems, automatic event detectors, etc
- Various mining techniques are/can be used: **community detection, clustering,** statistical analysis, classification, association rules mining, …

Static vs evolving data mining

- Social data interactions are constantly updating in fast rates
  - however, a data mining approach could be **static** or **evolving**

static mining approaches: aggregate all social interactions over a specific period and deal with them as a unique dataset

evolving mining approaches:
track and exploit more fine-grained & "richer" information

dynamic data mining:
- ✓ emphasis placed on data evolution and not on aggregation
- ✓ data modeling with a given time granularity which affects the amount of details contained in the dataset

streaming data mining:
- ✓ new user activity data received in a streaming fashion
- ✓ time-aware data approximating model incrementally created and maintained, subject to time and space constraints
- ✓ model readapted on arrival of either single update or batch of updates

Motivation for Evolving Social Data Mining

- Identifying over time **the events** that affect social interactions
  - ✓ tracking posts in a micro-blogging website to identify floods, fires, riots, or other events and inform the public
- Highlighting **trends** in users' opinions, preferences, etc
  - ✓ companies can track customers' opinions and complaints in a timely fashion to make strategic decisions
- Tracking the **evolution** of groups (*communities*) of users or resources, finding changes in time and correlations
  - ✓ develop better personalized recommender systems to improve user experience
  - ✓ scientists can more easily identify and relate social phenomena

- Social data in the Web 2.0
- Social associations and all kinds of graphs
  - structures for static social data
  - evolving data representation structures
- Evolving social data mining
- Emotion-aware social data analysis
- Frameworks and Applications

Social associations and all kinds of graphs

The network model as an obvious choice…

- Social data are interconnected through associations forming a **network** or **graph** *G(V, E)*, where *V* is the set of nodes and *E* is the set of edges.
  - ***nodes*** represent entities/objects and ***edges*** represent relations
  - different types of nodes and edges
  - weighted/unweighted
  - directed/undirected

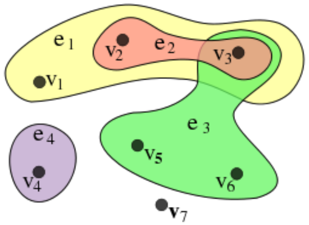
… the multi-graph structures

A hypergraph

example

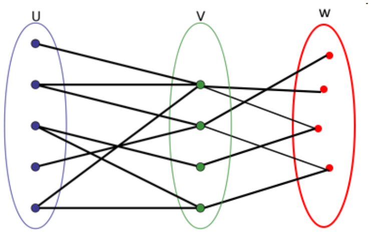
Structures for static social data

- **Hypergraph:** generalization of a graph where an edge (*hyperedge*) connects more than two nodes [Brinkmeier07]



- **Folksonomy:** lightweight knowledge representation emerging from the use of a shared vocabulary to characterize resources – *tripartite hypergraph* [Hotho06, Mika05]

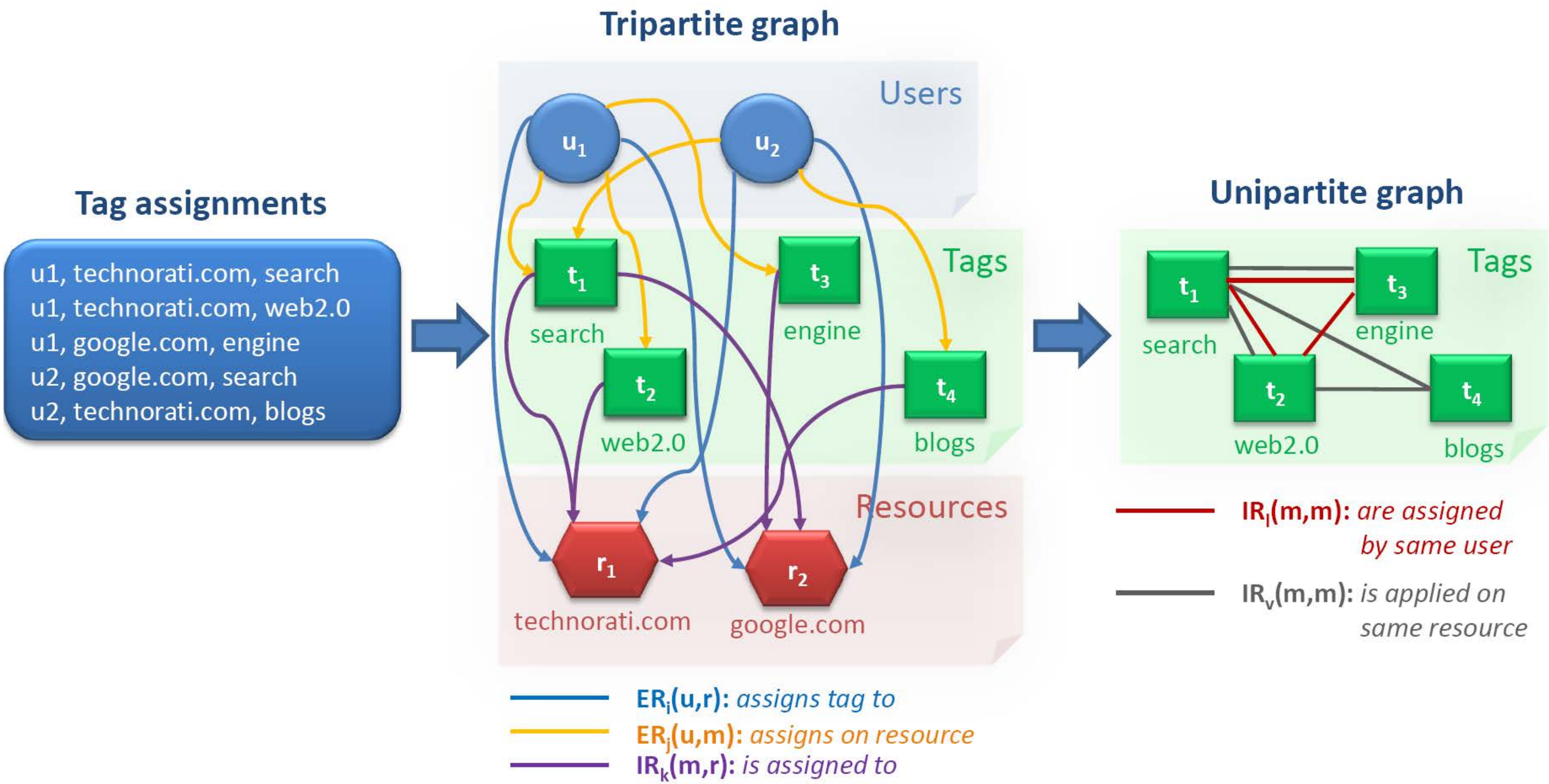
- Projection on **simple graphs** to lower complexity [Au Yeung09]
  - further simplifications in *bipartite* & *unipartite graphs*
  - e.g. tag-tag network where two tags are connected if assigned to the same resource



tripartite graph

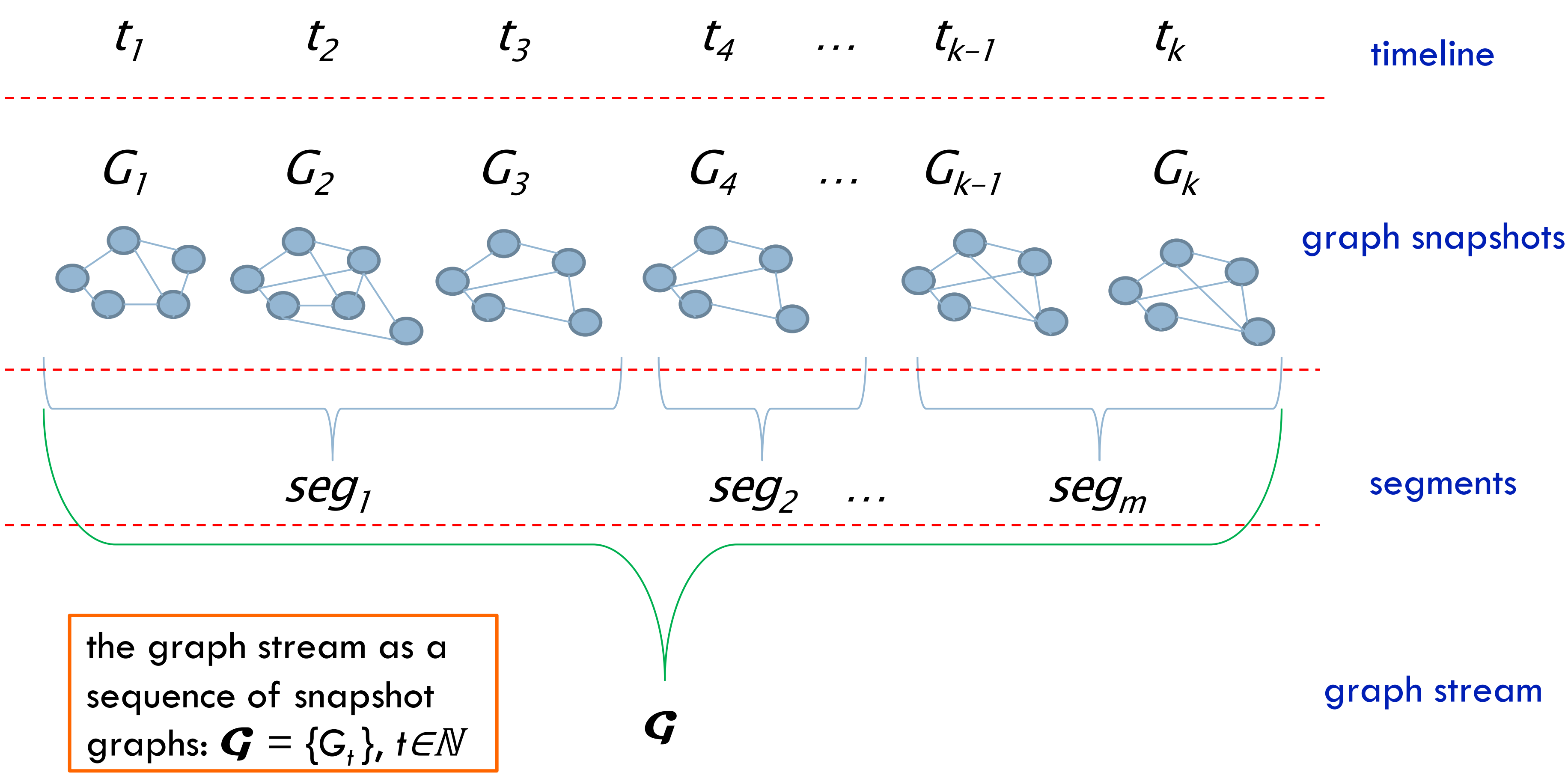
- Simple graphs' structure can be encoded in an *adjacency matrices* if **G**: unweighted or *similarity matrices* if **G**: weighted

Folksonomy projections on simple graphs

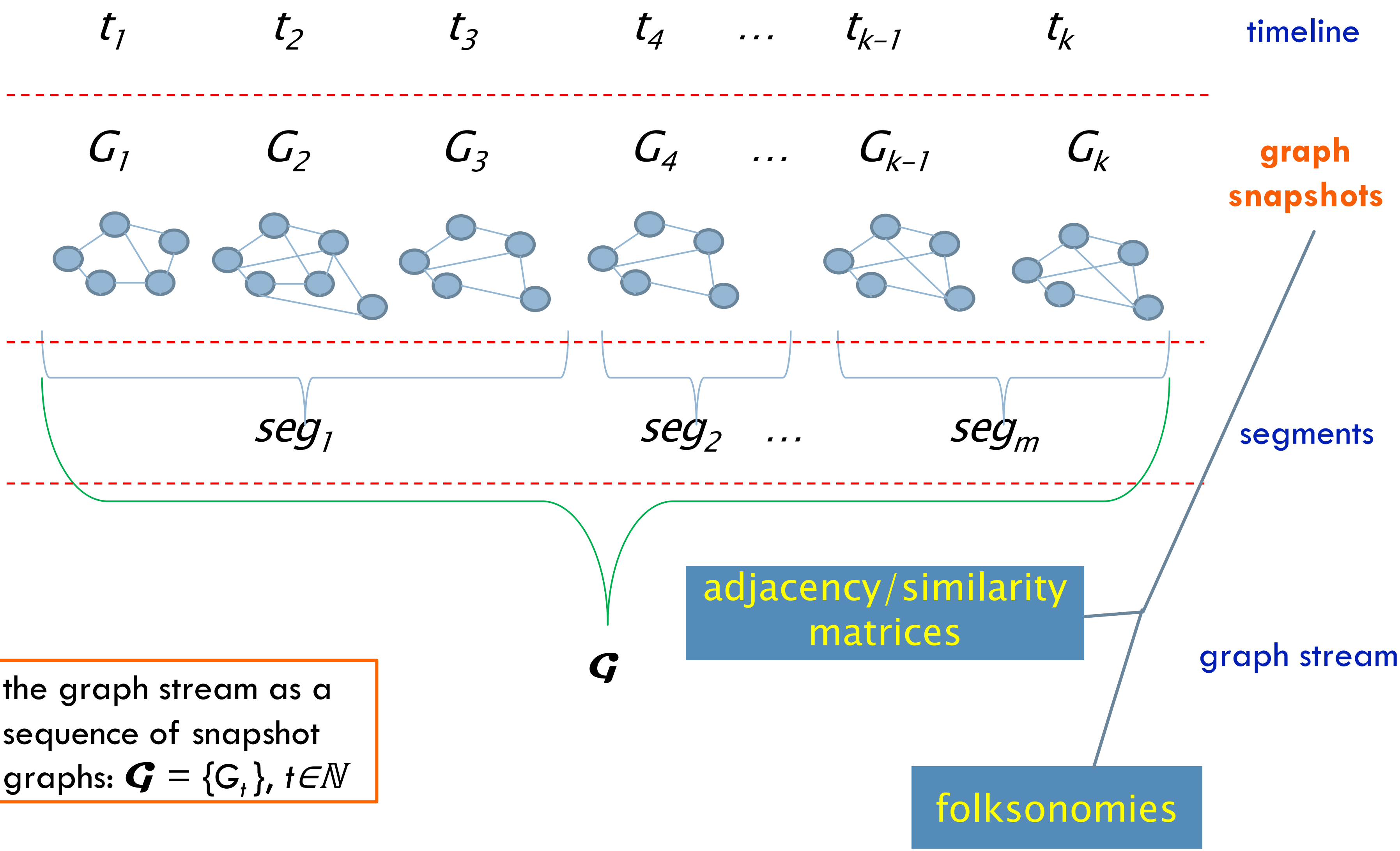


Evolving data representation structures

- Need for modeling the different data states in successive time-steps, often determined by the data's sampling rate



The snapshot layer

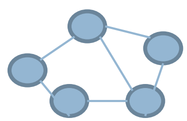


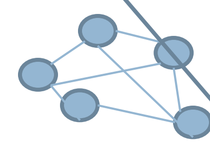
the graph stream as a sequence of snapshot graphs: $\mathcal{G} = \{G_t\}, t \in \mathbb{N}$

The segment layer

t_1 t_k timeline

G_1 G_k graph snapshots

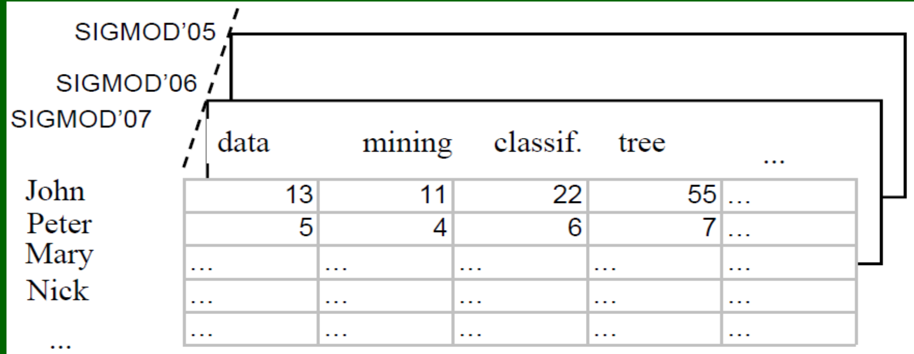




seg_m segments

graph stream

tensors: generalization of matrices (> 2 dimensions) [Sun06]



SIGMOD'05
SIGMOD'06
SIGMOD'07

	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

adjacency/similarity matrices

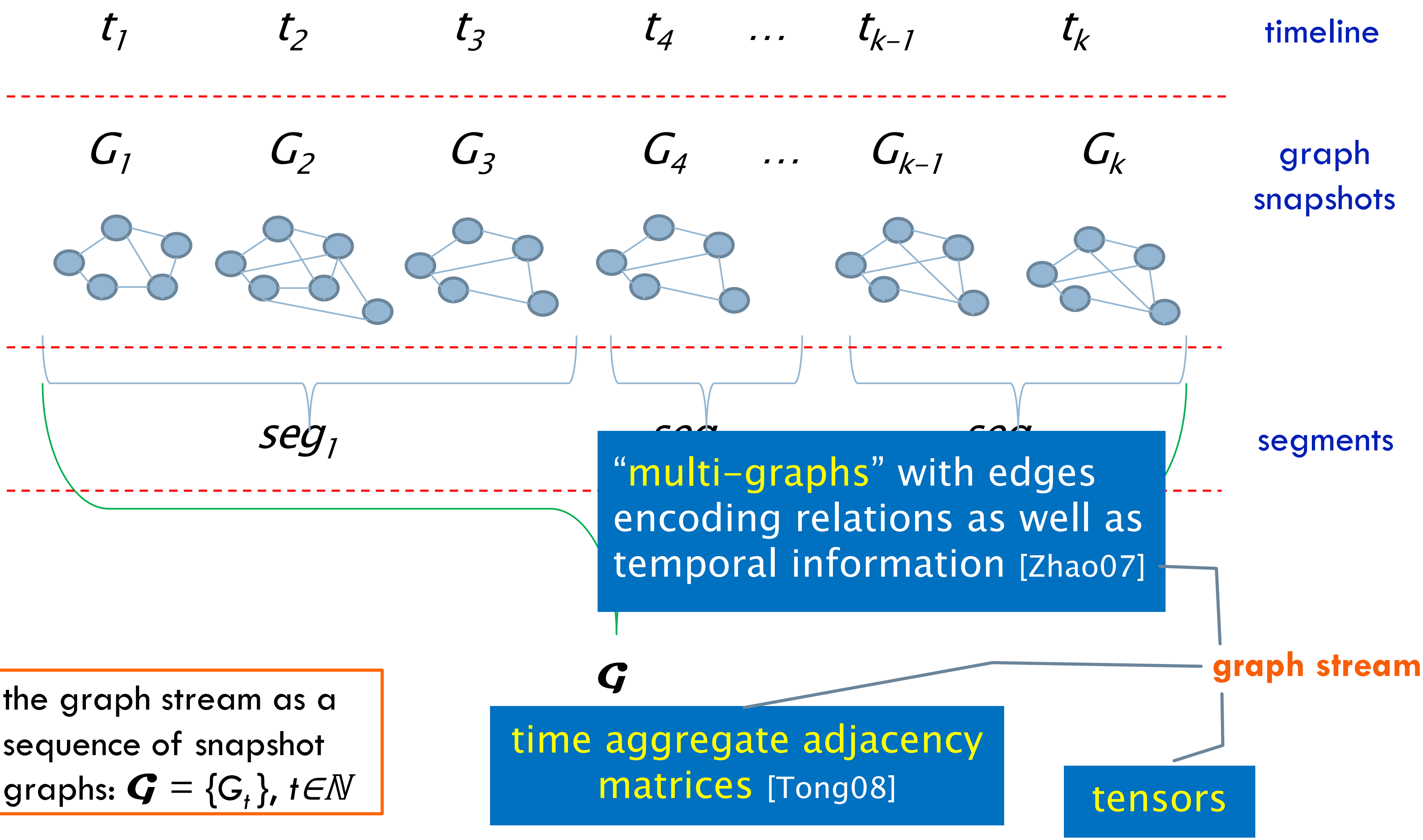
- ✓ more coarse-grained
- ✓ time-aggregated matrices are used for emphasizing on most recent edges [Tong08]

<u>Pre-processing technique</u>

Identify graph segments consisting of similar snapshots and compute a smooth graph approximation for each segment [Yang09]

the graph stream as a sequence of snapshot graphs: $\mathcal{G} = \{G_t\}, t \in \mathbb{N}$

The stream layer



the graph stream as a sequence of snapshot graphs: $\mathcal{G} = \{G_t\}, t \in \mathbb{N}$

References

[Giatsoglou12] M. Giatsoglou, A. Vakali. Capturing Social Data Evolution via Graph Clustering. In IEEE Internet Computing, IEEE computer Society Digital Library. IEEE Computer Society. DOI: 10.1109/MIC.2012.24, Feb. 2012.

Au Yeung09] Au Yeung, C.M., Gibbins, N., and Shadbolt, N. 2009. Contextualising Tags in Collaborative Tagging Systems. In Proceedings of 20th ACM Conference on Hypertext and Hypermedia, pp. 251-260.

[Brinkmeier07] Brinkmeier, M., Werner, J., and Recknagel, S. 2007. Communities in graphs and hypergraphs. CIKM'07. ACM, 869-872.

[Hotho06] Hotho, A., Robert, J., Christoph, S., and Gerd, S. 2006. Emergent Semantics in BibSonomy. GI Jahrestagung Vol. P-94, 305–312. Gesellschaft fr Informatik.

[Mika05] Mika, P. 2005. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Proceedings of the 4th international SemanticWeb Conference. ISWC'05. Springer Berlin /Heidelberg, pp. 522-536.

[Sun07] Sun, J., Faloutsos, C., Papadimitriou, S., and Yu, P. S. 2007. GraphScope: parameter-free mining of large time-evolving graphs. In Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining . KDD '07. ACM, 687-696.

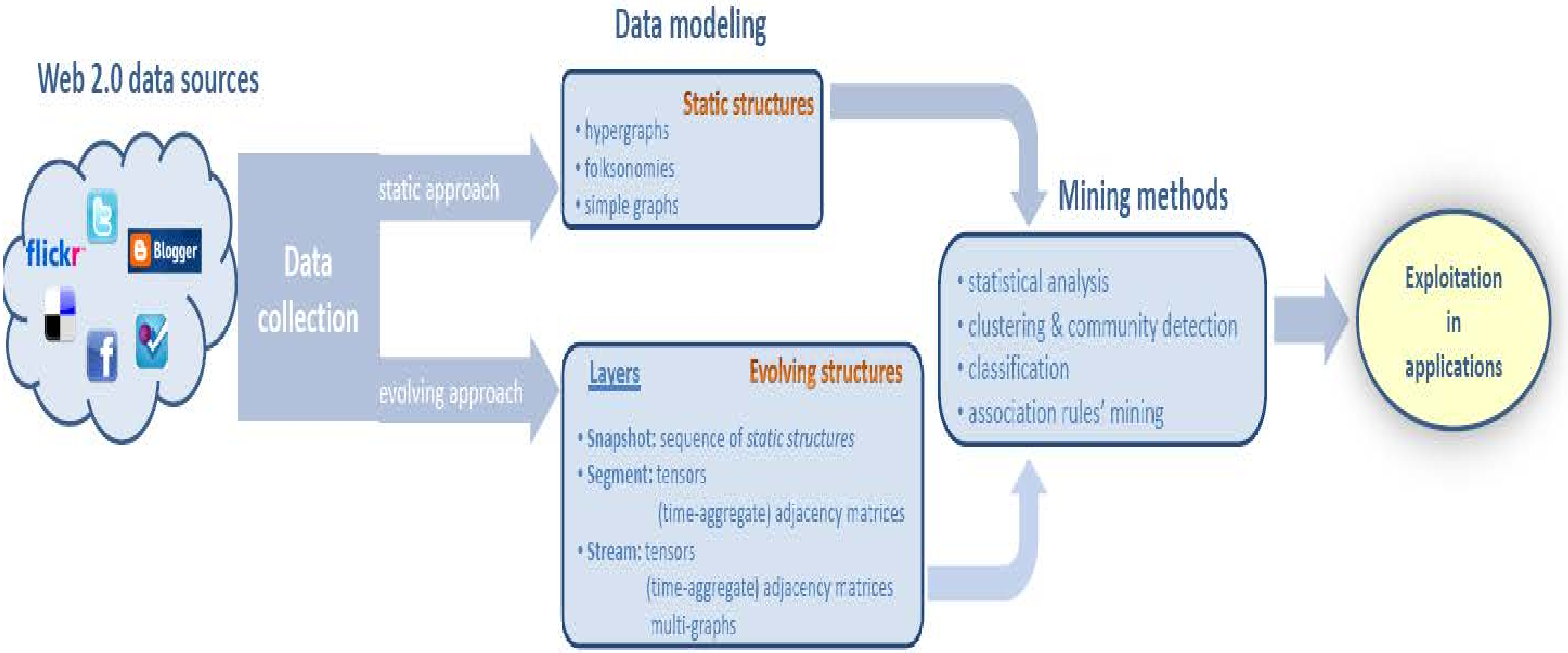
[Tong08] Tong, H., Papadimitriou, S., Yu, P. S., and Faloutsos, C. 2008. Proximity tracking on time-evolving bipartite graphs. In Proceedings of the 9th SIAM international Conference on Data Mining . SDM '08. 704–715.

[Yang09] Yang, S., Wu B., Wang, B. 2009. Tracking the Evolution in Social Network: Methods and Results. Complex (1): 693-706.

Presentation Outline

- Social data in the Web 2.0
- Social associations and all kinds of graphs
- Evolving social data mining
  - the clustering approach
  - Applications
- Emotion-aware social data analysis
- Frameworks and applications

Web 2.0 social data mining : The generic workflow



next, focus on clustering and community detection

Evolving social data clustering/community detection approaches

Preliminary efforts followed the **Community Mapping -CM- approach**:

1. application of traditional community detection algorithms on individual static graph snapshots
2. identification of correlations and mapping between successive snapshots' communities using special similarity measures & temporal smoothing techniques

Limitation: tend to find community structures with high temporal variation (due to real world ambiguous & noisy data)

Evolving social data clustering/community detection approaches

Evolutionary community identification approaches: utilize community structure's history to maximize **temporal smoothness** and lead to smoother community evolution

- **traditional clustering revisited for an evolutionary setting (TCR approach):** modification of existing static clustering algorithms to have memory of previous states of the data;
- **spectral clustering (SC approach):** uses the spectrum of the graph's similarity matrix to perform dimensionality reduction for clustering in fewer dimensions;
- **non-negative matrix/tensor factorization (NFC approach):** apply NFC for discovering communities jointly maximizing the fit to the observed data & the temporal evolution;
- **graph stream segment identification & community structure detection (SGC approach):** finds characteristic change-timepoints for *segmenting* a graph stream and identifies communities within each segment

CM approach

method	structure	dataset: source	dataset: time-period	dataset: entities / relations	outcome/scope
Palla 2008	undirected graph snapshots	I. co-authorship network II. mobile phone call network	I. 142 months II. 52 weeks (26 *timeslots*)	I. {30K authors} {"*are co-authors*" } II. {4M users} {"*calls*"}	community detection based on *clique percolation;* successive communities mapping by relative nodes' overlap;
Lin 2007	directed graph snapshots	data crawled from 407 blogs	63 weeks	{275K bloggers} {149K "*replies to post of*"}	**hypothesis:** *users'* ***mutual awareness*** *(e.g. bloggers commenting on each other's blog) drives community formation;* mutual awareness' expansion as a random walk process leads to community detection *community mapping using interaction correlation*

TCR approach

Simultaneous optimization of two potentially conflicting criteria:

(i) snapshot quality ***sq***, and (ii) history quality ***hq***

At each time-step the framework finds a clustering based on the new similarity matrix M_t and the so far history

method	structure	dataset: source	dataset: time-period	dataset: entities / relations	outcome/scope
Chakrabarti 2006	(time-aware) similarity matrix	photo sharing service	68 weeks	{5K tags} {*"are applied on same resource"*}	joint optimization of 2 criteria: ▪ **snapshot quality** ▪ **history quality;** generic framework with 2 instantiations proposed: ▪ agglomerative hierarchical algorithm ▪ K-means

Evolutionary clustering in an online setting

SC approach

Spectral clustering uses the spectrum of the graph's similarity matrix to perform dimensionality reduction for clustering in fewer dimensions

method	structure or model	dataset: source	dataset: period	dataset: entities / relations	outcome/scope
Tang 2008	series of network snapshots; interaction matrix for each snapshot	I. mail exchange network II. co-authorship network	I. 12 months II. 25 years	I. {2.4K users} {emails} {36.7K words} *{"sends email"}* *{"receives email"}* *{"contains term"}* II. {492K papers} {347K authors} {2.8K venues} {9.5K title terms} *{"writes"}* *{"participates in"}* *{"is published in"}* *{"contains term"}*	community evolution in dynamic networks of multiple social entities; iterative approximation of community evolution using **eigenvector calculation** and **K-means clustering**

Evolutionary spectral clustering applied in multi-mode networks

NFC approach

method	structure or model	dataset: source	dataset: period	dataset: entities / relations	outcome/scope
Lin 2009	***metagraph***: hypergraph with nodes representing *facets* and edges multipartite interactions; **tensors**	social bookmarking service with voting capabilities	27 days (9 *timeslots*)	{users} {posts} {keywords} {topics} {152K "*post-keyword-topic*"} {56.4K "*is friend with*"} {44K"*uploads*"} {1.2M "*votes on*"} {242K "*user-post-comment*"} {94.6K "*replies with*"}	community extraction via time-stamped **tensor factorization;** **on-line method** handling time-varying relations through incremental metagraph factorization; communities derived by jointly leveraging all types of multipartite relations

SGC approach

method	structure	dataset: source	dataset: period	dataset: entities / relations	outcome/scope
Sun 2007	unweighted undirected **bipartite** graphs; graph segments	I. mail exchange network II. mobile phone call network III. mobile device proximity records	I. 165 weeks II. 46 weeks III. 46 weeks	I. {34.3K senders} {34.3K recipients} {15K "*send mail*" / week} II. {96 callers} {3.8K callees} { 430 "*calls*"/ week} III. {96 users } {96 users} {689 "*is located near*"/week}	unparametric method based on Minimum Description Length; each segment's source and destination nodes are partitioned separately to decrease cost; compressed graph in < 4% than original space

References

[Chakrabarti06] Chakrabarti, D., Kumar, R., and Tomkins, A. 2006. Evolutionary clustering. In Proc. of KDD '06. ACM, New York, NY, 554-560.

[Tang08] Tang, L., Liu, H., Zhang, J., and Nazeri, Z. 2008. Community evolution in dynamic multi-mode networks. In Proc.. KDD '08. ACM, New York, NY, 677-685.

[Palla05] Palla, G., Derény, I., Farkas, I., and Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818.

[Palla07] Palla, G., Barabási, A.-L., and Vicsek, T. 2007. Quantifying social group evolution. Nature 446, 664-667.

[Lin07] Lin, Y., Sundaram, H., Chi, Y., Tatemura, J., and Tseng, B. L. 2007. Blog Community Discovery and Evolution Based on Mutual Awareness Expansion. In Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence. Web Intelligence. IEEE Computer Society, Washington, DC, 48-56.

[Lin09] Lin, Y., Sun, J., Castro, P., Konuru, R., Sundaram, H., and Kelliher, A. 2009. MetaFac: community discovery via relational hypergraph factorization. KDD'09. ACM, 527-536.

[Sun07] Sun, J., Faloutsos, C., Papadimitriou, S., and Yu, P. S. 2007. GraphScope: parameter-free mining of large time-evolving graphs. In Proc. of KDD '07. ACM,, 687-696.

[Quack08] Quack, T., Leibe, B., and Van Gool, L. 2008. World-scale mining of objects and events from community photo collections. In Proc. of CIVR '08. ACM, 47-56.

Zhao07] Zhao, Q., Mitra, P., and Chen, B. 2007. Temporal and information flow based event detection from social text streams. In Proc. of the 22nd National Conference on Artificial intelligence - Volume 2. Aaai Conference On Artificial Intelligence. AAAI Press, 1501-1506.

[Duan09] Duan, D., Li, Y., Jin, Y., and Lu, Z. 2009. Community mining on weighted directed graphs. In Proc. of CNIKM '09. ACM, New York, NY, 11-18.

[Chi06] Chi, Y., Tseng, B. L., and Tatemura, J. 2006. Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In Proc of CIKM '06. ACM, 68-77.

[Papadopoulos11] Papadopoulos, S., Zigkolis, C., Kompatsiaris, Y., and Vakali A. 2011. Cluster-Based Landmark and Event Detection for Tagged Photo Collections. IEEE Multimedia, pp. 52-63.

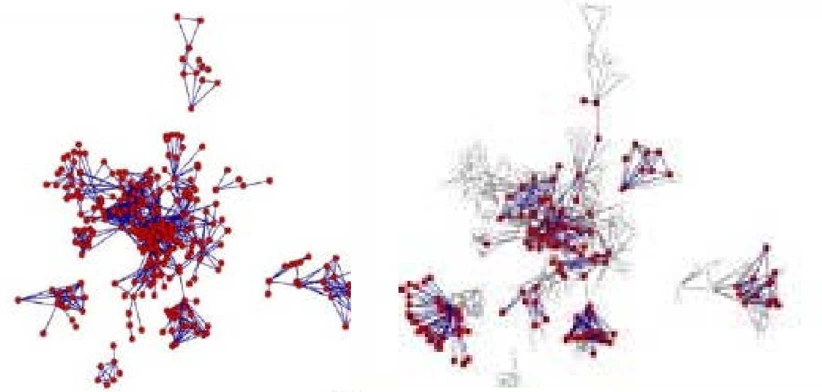
[Java07] Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. WebKDD/SNA-KDD'07. ACM, 56-65.

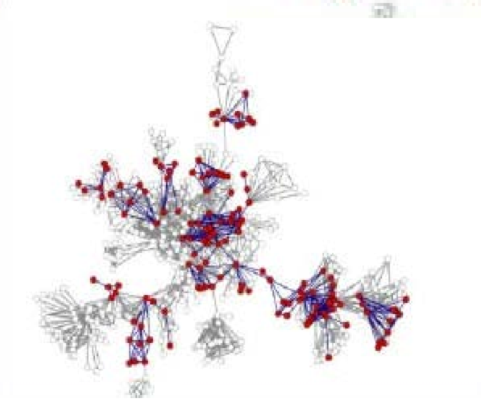
[Jansen,09] Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. Technol. 60, 11, 2169-2188.

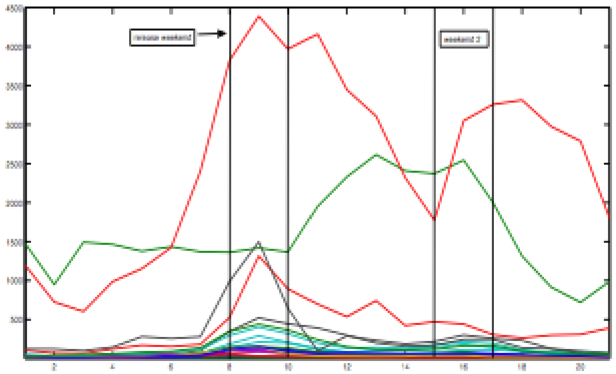
[Sankaranarayanan 09] Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., and Sperling, J. 2009. TwitterStand: news in tweets. GIS'09. ACM, 42-51.

Why focusing on time as a criterion ?

- typical analysis involves "static" views (users-tags)
- events, trends affect user interests
- users Tagging Behavior changes over time
- Time is a fundamental dimension in analysis of users and tags in a social tagging system

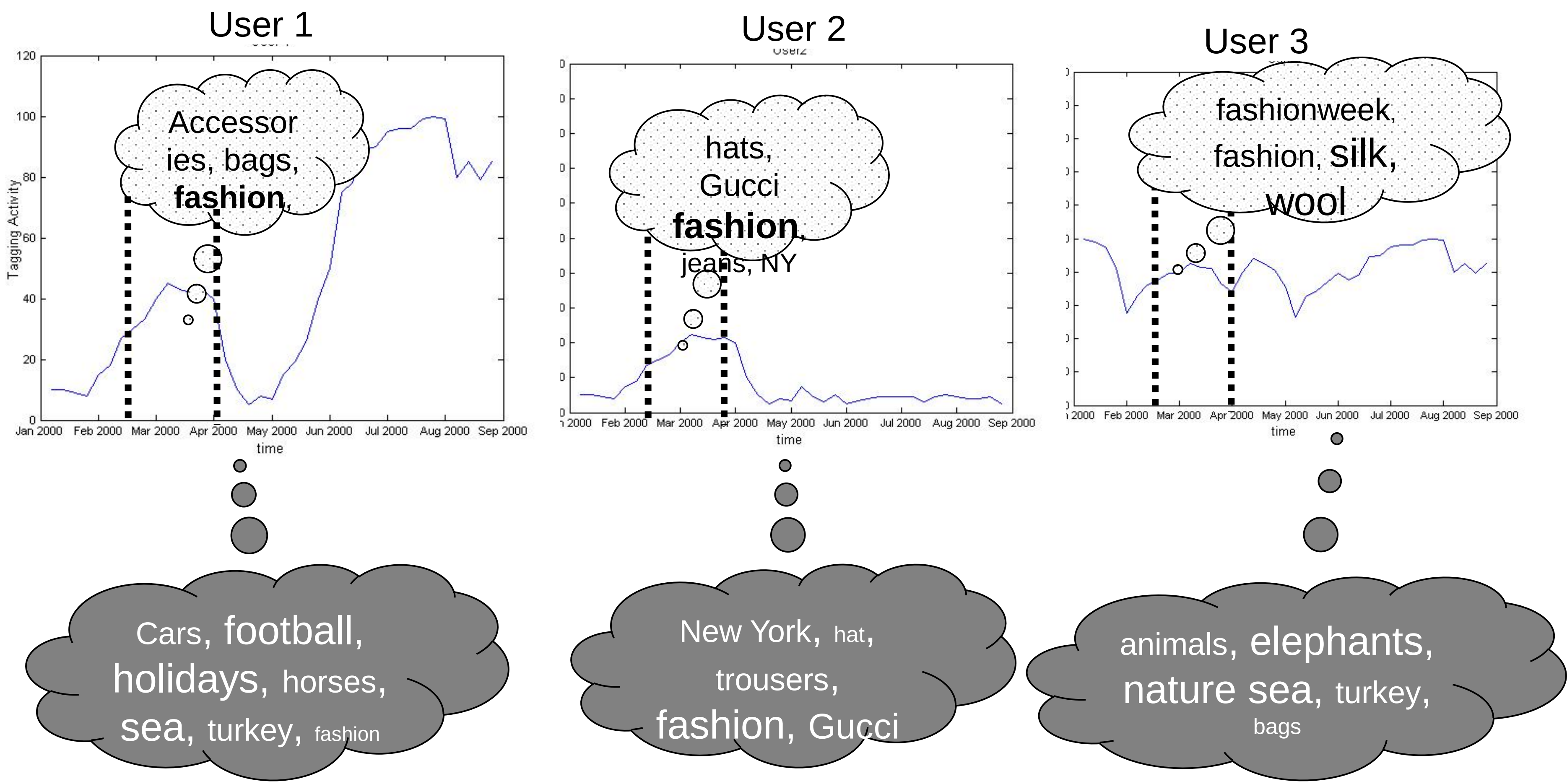






e.g. : prediction of first weekend box-office revenues using tweets

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



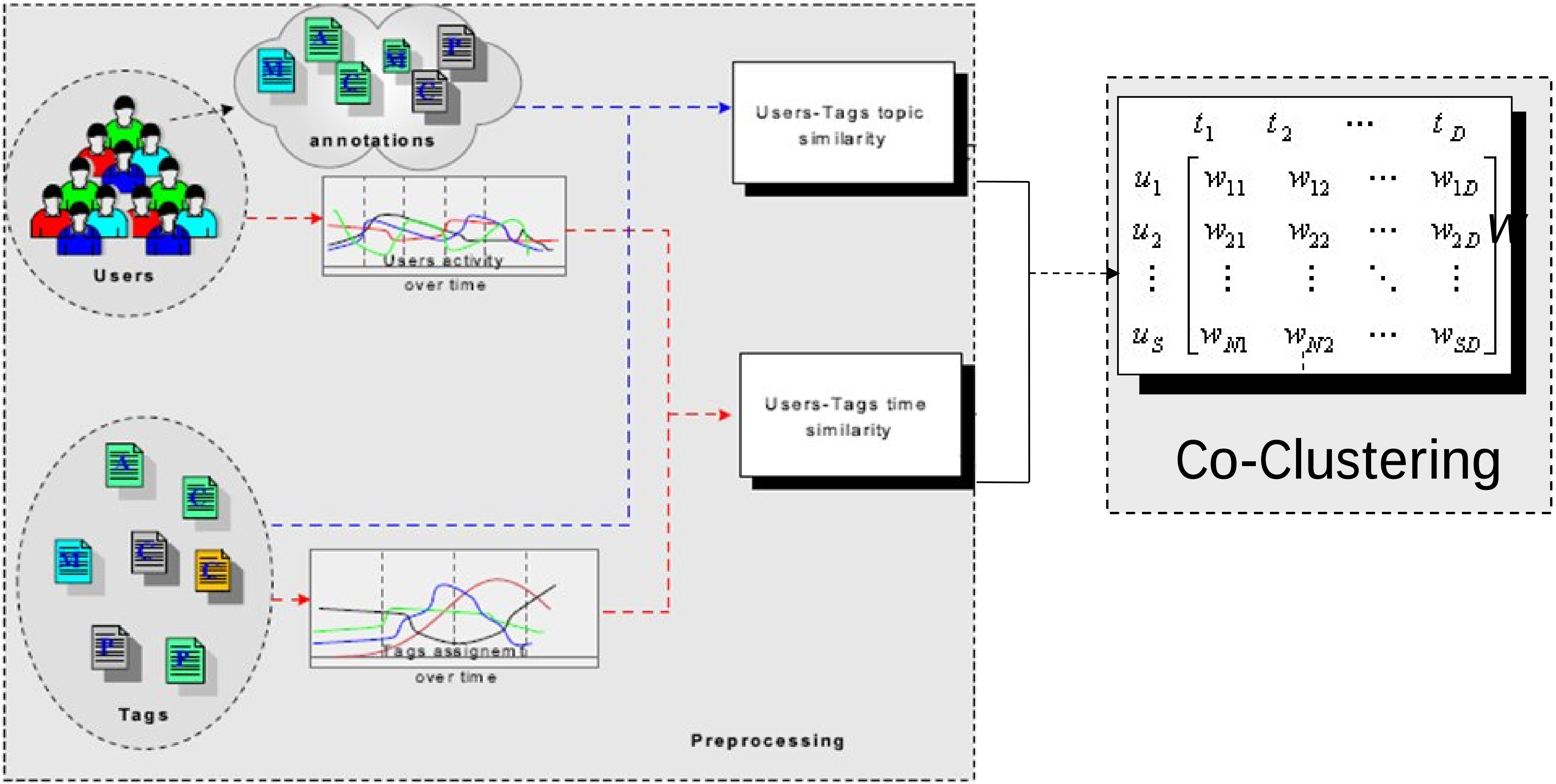
Time-aware user/tag clustering

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)

Related Approaches

- Sun and colleagues [Sun08] use the χ^2 *statistical* model, to determine whether the appearance of *tag t in a time frame i is significant and, thus, to discover tags that constitute* "topics of interest" at particular time frames.
- Wetzker and colleagues [Wetzker08] claim that a tag signifies a trend, if it attracts significantly more new users in a currently monitored time frame than in past time frames.
- A trend detection measure is introduced in [Hotho06], which captures topic-specific trends at each time frame and is based on the weight-spreading ranking of the PageRank algorithm

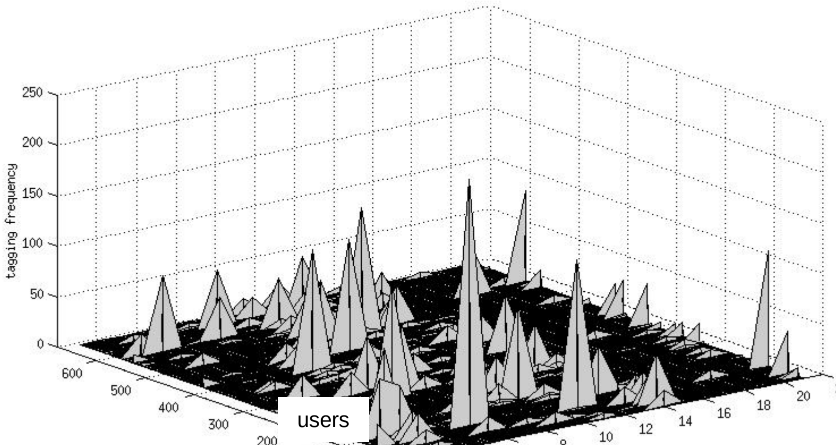
a co-clustering approach

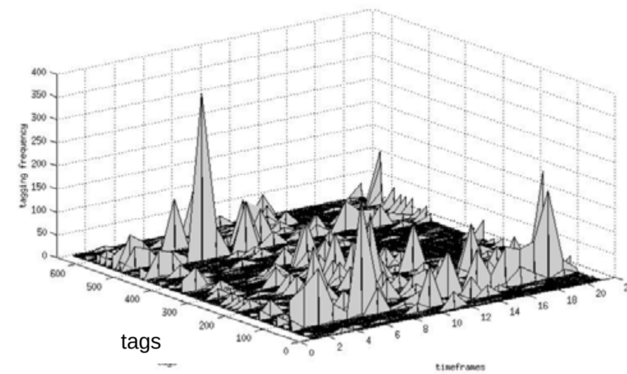


E. Giannakidou, V. Koutsonikola, A. Vakali and I. Kompatsiaris, "Exploring Temporal Aspects in User-Tag Co-Clustering", In Proc. 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010), 12-14 April 2010, Italy

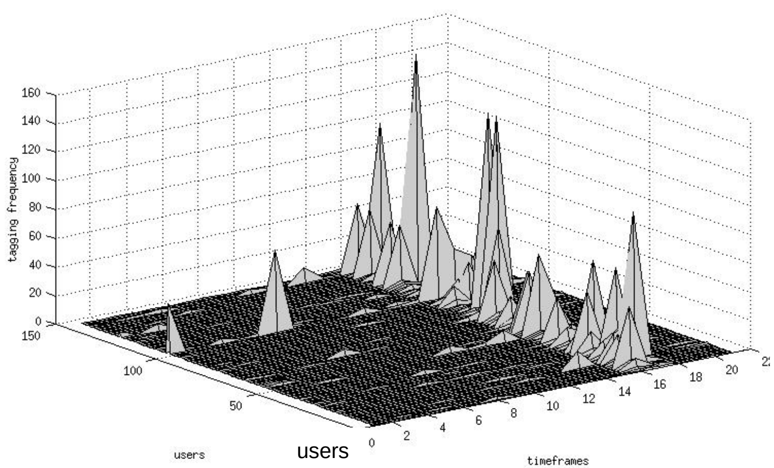
Clusters structures formation

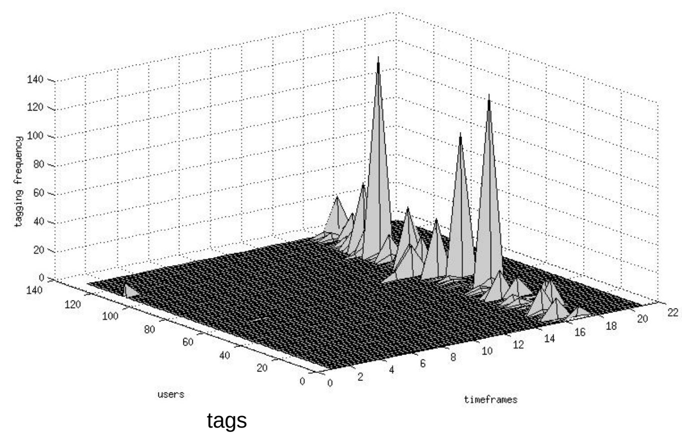
conventional clustering





co-clustering





[E. Giannakidou, V. Koutsonikola, A. Vakali and I. Kompatsiaris, "Exploring Temporal Aspects in User-Tag Co-Clustering", In Proc. 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010), 12-14 April 2010, Italy]

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 data in the Web 2.0
- Social associations and all kinds of graphs
- Evolving social data mining
- In & out zooming on time aware user/tag clusters
- Emotion-aware social data analysis
  - User generated data
  - The idea for emotion aware clustering
  - Application on micro-blogging sources
- Frameworks and applications

The idea for emotion-aware analysis

Clustering methods embed various criteria such as:

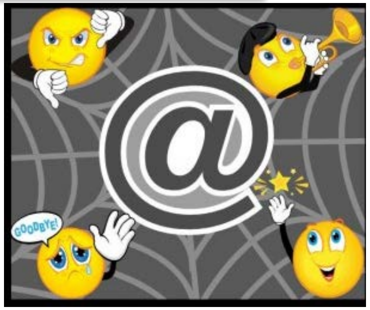
semantics, tags, time, geo-information … etc

but

since social sources are driven and managed by humans

➲ **emotion/sentiment**

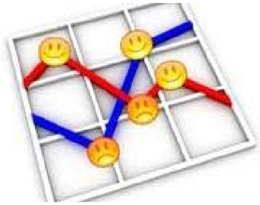
MUST also be considered …



The nature of evolving social data

- Two main types of textual information.
  - Facts and Opinions
    - Note: factual statements can imply opinions too.

- Most current text information processing methods (e.g., web search, text mining) work with factual information.



- Sentiment analysis or opinion mining
  - computational study of opinions, sentiments and emotions expressed in text.
- Why sentiment analysis now? Mainly because of the Web; huge volumes of opinionated text.

User-generated data (I)

Opinions are important because whenever we need to make a decision, we're influenced by others' opinions.

- According to [Horrigan09] of more than 2000 American adults :
  - 81% of Internet users (or 60% of Americans) have done online research on a product at least once;
  - among readers of online reviews of restaurants, hotels, and various services (e.g., travel agencies or doctors), between 73% and 87% report that reviews had a significant influence on their purchase;
  - 32% have provided a rating on a product, service, or person via an online ratings system, and 30% have posted an online comment or review regarding a product or service

user-generated data (II)

- Word-of-mouth on the Web
  - User-generated media: One can express opinions on anything in reviews, forums, discussion groups, blogs ...
  - Opinions of global scale: No longer limited to:
    - Individuals: one's circle of friends
    - Businesses: Small scale surveys, tiny focus groups, etc.

- Why affect/sentiment analysis?
  - Customers: need peer opinions to make purchase decisions
  - Business providers:
    - need customers' opinions to improve product
    - need to track opinions to make marketing decisions
  - Social researchers: want to know people's reactions about social events
  - Government: wants to know people's reactions to a new policy
  - Psychology, education, etc.

More Applications

- **Product review mining:** What features of the iPhone do customers like and which do they dislike?
- **Review classification:** Is a review positive or negative toward the iPhone?
- **Tracking sentiments toward topics over time:** Is anger ratcheting up or cooling down?
- **Prediction (election outcomes, market trends):** Will Clinton or Obama win?

Why Extracting sentiments from Web 2.0 data sources?

- **Web 2.0 data features**
  - Easy to collect: huge amount, clean format
  - Broadly distributed: demographics
  - Topic diversified: free discussion about any topic/product/event
  - Opinion rich: highly personalized
  - Distributed over time, user generated content
- **Motivation**
  - Sentiment is a very natural expression of a human being.
  - Sentiment Analysis aims at getting sentiment-related knowledge especially from the huge amount of information on the internet
  - Can be generally used to understand opinion in a set of documents or user generated content

Challenges

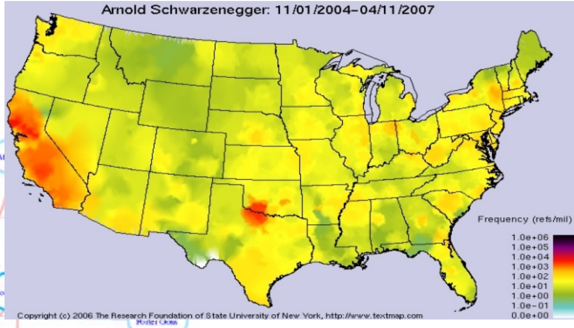
- **Contrasts with Standard Fact-Based Textual Analysis**
  - typically, text categorization seeks to classify documents by topic
  - BUT nature, strength of feelings, degree of positivity, etc imposes a tailored sentiment categorization
- **Key Factors that Make Sentiment Analysis challenging**
  - choosing the right set of keywords might be less trivial than one might initially think;
  - Sentiment and subjectivity are quite context-sensitive, and, at a coarser granularity, quite domain dependent
    - e.g., "go read the book" most likely indicates positive sentiment for book reviews, but negative sentiment for movie reviews.
  - Web users postings are of a challenging nature, since there is no code in expressions

so far … Lexical Resources

- *SentiWordnet*
  - Built on the top of WordNet synsets
  - Attaches sentiment-related information with synsets
  - SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity
- *General Inquirer*
  - Included are manually-classified terms labeled with various types of positive or negative semantic orientation, and words having to do with agreement or disagreement.
- *OpinionFinder's Subjectivity Lexicon*
  - OpinionFinder is a system that performs *subjectivity analysis*, automatically identifying when opinions, sentiments, speculations, and other *private states* are present in text.

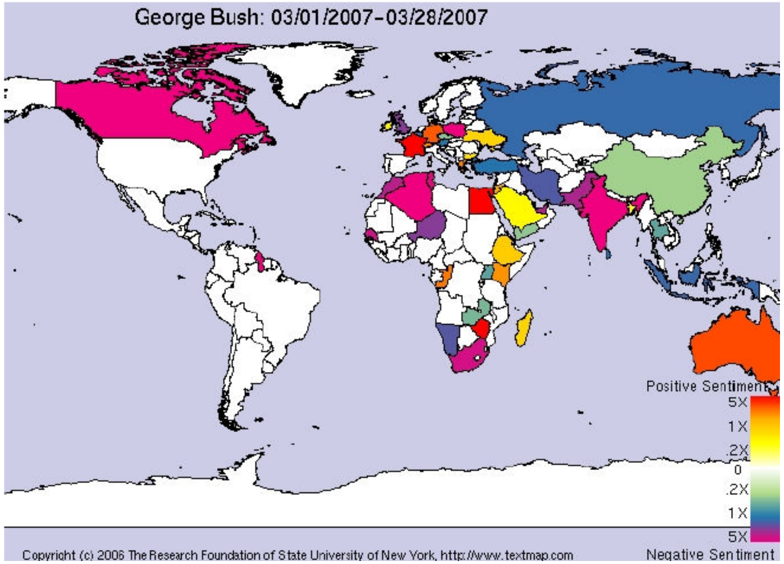
so far … Lydia System

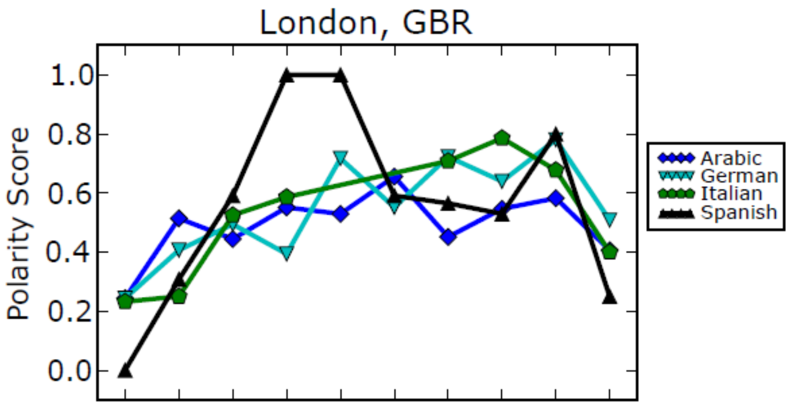
- Lydia [Lioyd05] news analysis system does a daily analysis of over 1000+ online English newspapers, Blogs, RSS feeds, and other news sources.
- It identifies who is being talked about, by whom, when and where?
- Applications of Lydia
  - heatmap generation (pos/neg for a topic);
  - relational networks



so far … integration of news & blogs

- Bautin, Vijayarenu and Skiena [Bautin08] presented an approach for the international analysis for news and blogs … still on the positive/negative side …
  - Cross-language analysis across news streams

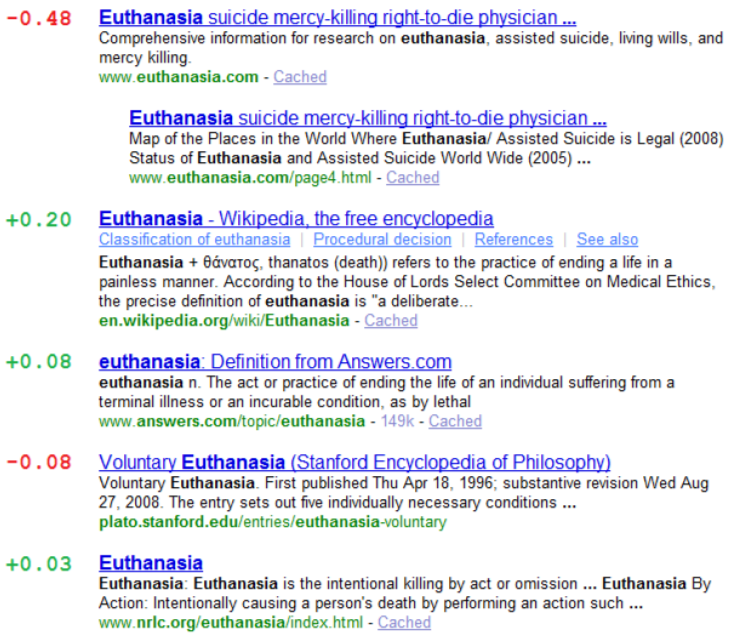


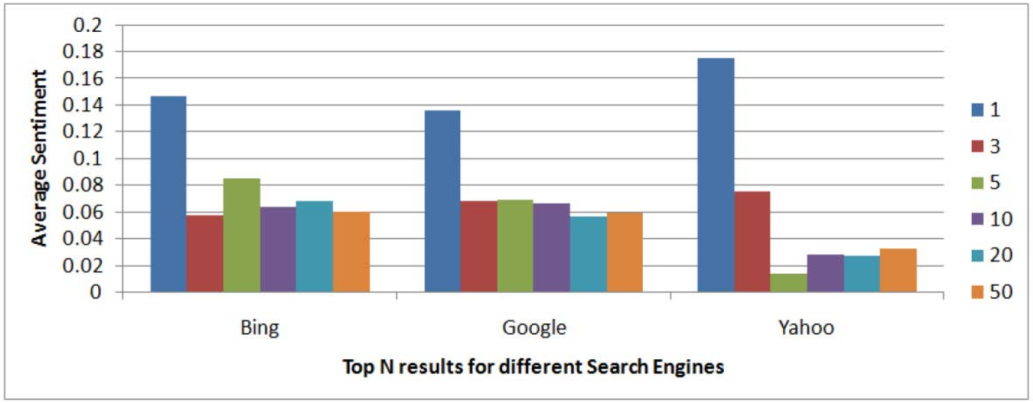


Polarity score of London in Arabic, German, Italian and Spanish over the May 1-10, 2007 period.

so far … sentiment-aware searching

- Sentiment Analysis for Semantic Enrichment of Web Search Results [Demartini10]
- the first few results a representative sample of the entire result set

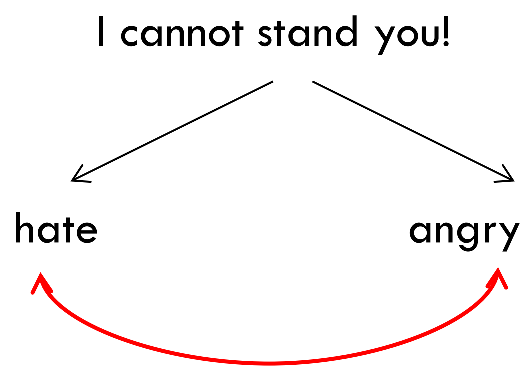




Average Sentiment score in top N results for 3 search engines

so far … beyond pos/neg : the affect analysis

- it involves several affects at the same time.
- affect classes may be correlated or opposed.
- Abbasi, Chen Thoms and Fu [Abbasi08] proposed a support vector regression correlation ensemble (SVRCE) method for text-based affect classification.
  - affect feature and technique comparison.
  - apply to multi-domain.



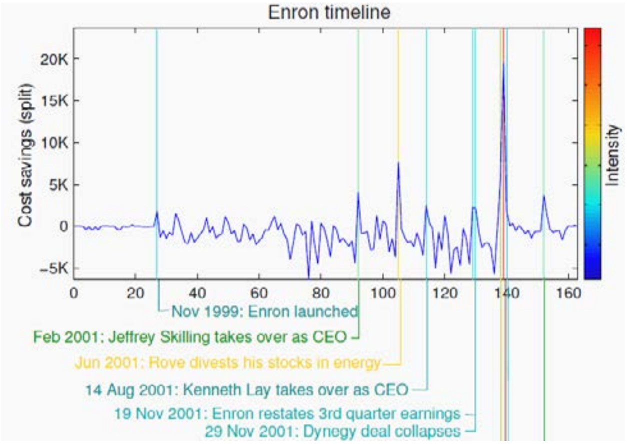
I cannot stand you! →

Affect class	Intensity
Happiness	0.01
Sadness	0.03
Anger	0.6
Hate	0.5

- Social data in the Web 2.0
- Social associations and all kinds of graphs
- Evolving social data mining
- In & out zooming on time aware user/tag clusters
- Emotion-aware social data analysis
- Frameworks and applications
  - Most popular applications
  - A mining and analysis framework
  - Social data analysis on the cloud
  - Emotions capturing in microblogs

Applications of Mining Evolving Social Data

The results of community detection, or different mining techniques, on evolving social data can be exploited in applications:



event detection
diagram from [Sun07]

clustering of users
exploiting the time dimension

social network analysis
image from [Touchgraph]

trend detection
image from [Trendsmap]

[Touchgraph] *http://www.touchgraph.com/TGFacebookBrowser.html*
[Trendsmap] *http://trendsmap.com/*

Event detection

Definition of event

- **information flow** between a group of **social actors** on a specific **topic** over a certain **time period** [Zhao07]
- occasions which take place at a **specific time** and **location** (concerts, festivals, etc.) [Quack08]

event detection from social data streams [Zhao07]

- features exploration in 3 dimensions: textual content, social, temporal
- generation of multiple intermediate clustering structures using content-based similarity & information flow patterns
- events as groups of nodes closely related in time & topic

Graph segmentation community detection methods [Sun07, Duan09] identify events as significant change-timepoints in the stream.

Trend tracking

- Social data fluctuate in <u>structure and frequency as they evolve</u> and over time some topics, images, tags, etc, become most popular amongst users.
  Trends can be identified by a data mining approach *globally* or *locally* (within communities) and they usually indicate what interests users the most at a given time.

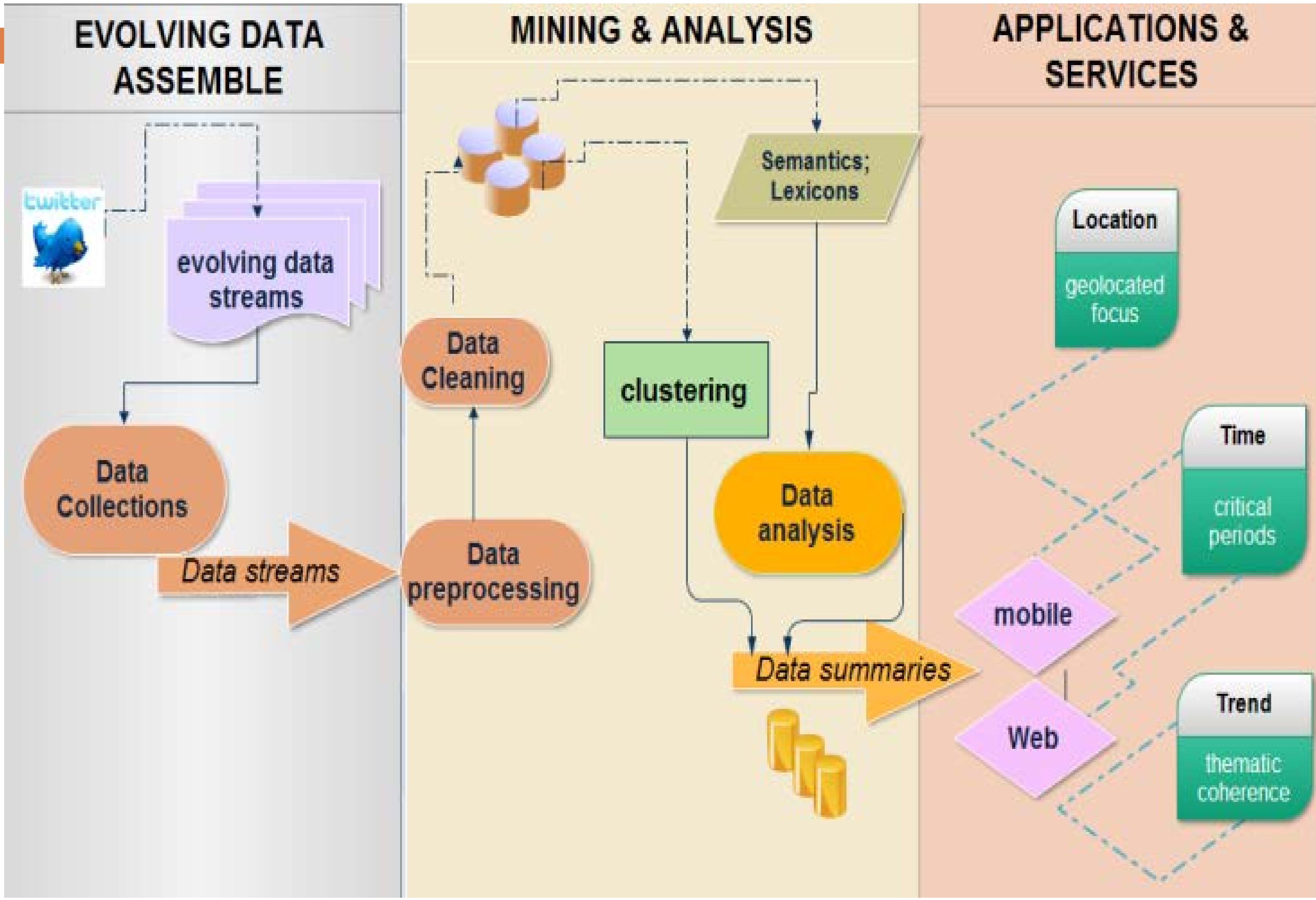
trend detection in blogs [Chi06]

- focused on:
  - ✓ keyword popularity in successive timeframes
  - ✓ detection of different topics relating to a keyword
  - ✓ contribution of individual users to a trend
- uses the results of **Singular Value Decomposition** as trend indicators capturing both temporal data changes & bloggers' characteristics
- exploits textual content and citations between blogs

trend detection in Twitter [Java07], [Jansen09], [Sankaranarayanan09]

- several attempts using statistical analysis methods
  - ✓ analysis of evolving Twitter data to identify trending keywords for different weekdays
  - ✓ sentiment identification on tweets to identify trending sentiments about brands
- online clustering on streaming tweets, combined with classification, to identify breaking news

A mining and analysis framework



2 indicative applications

- **Cloud4Trends**

Leveraging the cloud infrastructure for localized real-time trend detection in social media

- **CapturEmos**

Capturing emotional patterns in micro-blogging data streams

Cloud4Trends - Motivation

- Social media reflect societal concerns exhibiting ‘bursts’ of content generation on the occurrence of ***events***
  - *popular topics / interests fluctuate with time*
- Challenging for both computer scientists & application developers to reach unbiased, meaningful conclusions about ***trending*** users’ opinion and interests

VENUS-C
Virtual multidisciplinary EnviroNments USing Cloud infrastructures

Cloud4Trends is a microblogging & blogging localized **content collection and analysis framework** for detecting currently popular topics of users’ interest

Challenges & Outcome

- **Massive** content sizes and **unpredictable** content generation rates :
  - scalable analysis is needed
- Trending topics should be discovered when they are **"fresh"** :
  - an on-line analysis approach is demanded
- Trends should be **meaningful**
  - need for contextual trends
- Content is **dispersed** in multiple sources :
  - trend detection needs a combined approach

The Cloud deployment of the Cloud4Trends scenario with use of the VENUS-C services verified that Cloud-based architectures are a viable solution for online web data mining applications that are beneficial for both researchers and entrepreneurs.

Welcome **avakali**! [Log Off]

Home | About

Cloud4Trends: Localized Real-Time Trend Detection In Social Media

Index

Create New

Title	Description	Location	Coordinates	Inititation date/time	Status	
Another demo experiment	A London experiment	gblondon	-0.56,51.26,0.28,51.68	6/27/2012 1:39:47 PM	Running	Subscribe
My demo experiment	A demo experiment	usnewyork	-74,40,-73,41	6/22/2012 3:35:28 PM	Running	Subscribe
A london experiment	Experiment	gblondon	-0.56,51.26,0.28,51.68	6/22/2012 8:12:33 PM	Stopped	

OSWINDS

OPERATING SYSTEMS WEB/INTERNET DATA SOURCES MANAGEMENT

Venus-C

Home | About

Cloud4Trends: Localized Real-Time Trend Detection In Social Media

Trends from Blogger

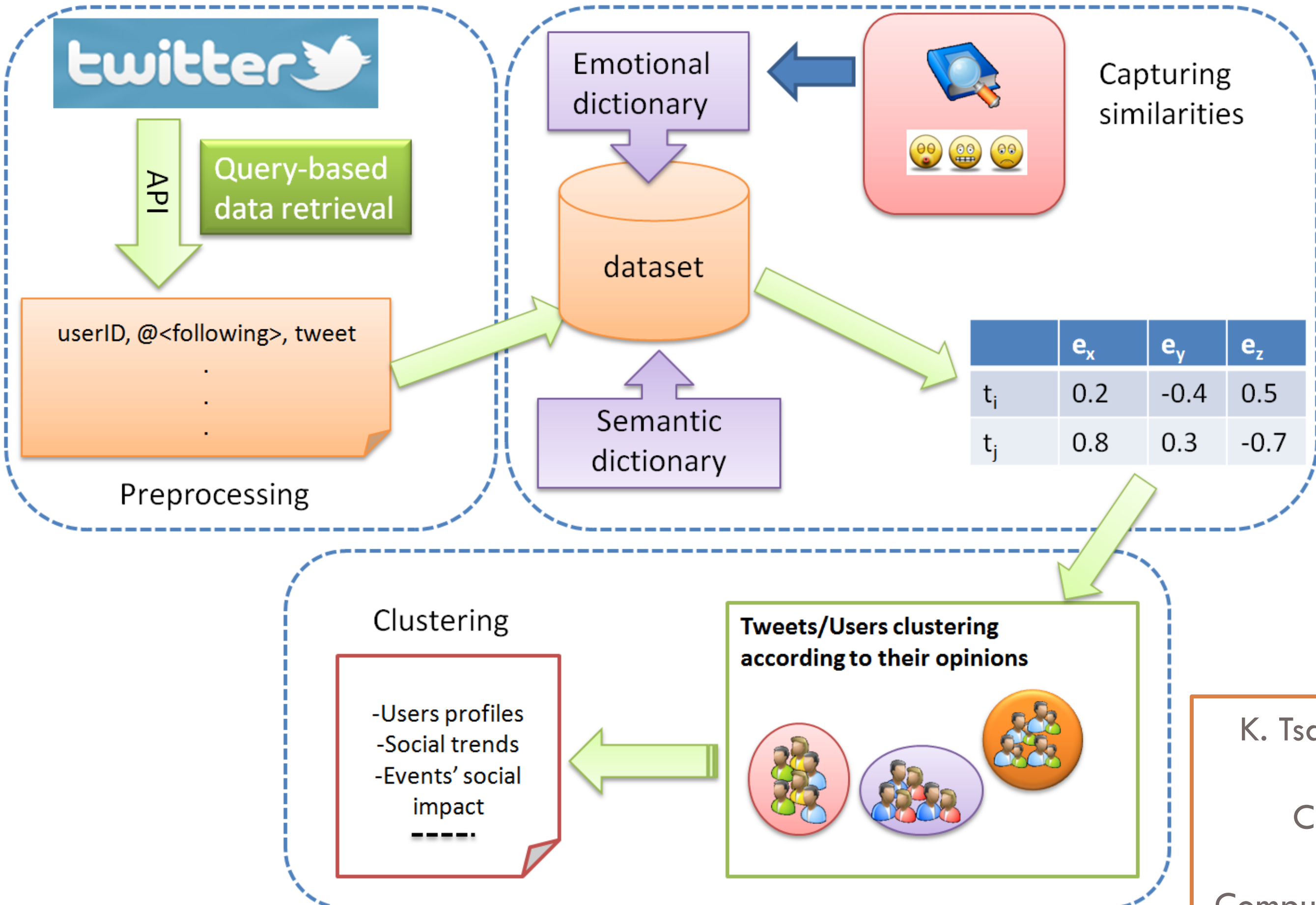
1

2

3

http://cloud4trendsdemo.cloudapp.net

Emotional Aware Clustering on Micro-Blogging Sources (affect analysis)



	e_x	e_y	e_z
t_i	0.2	-0.4	0.5
t_j	0.8	0.3	-0.7

K. Tsagkalidou, V. Koutsonikola, A. Vakali and. K. Kafetsios: Emotional Aware Clustering on Micro-Blogging Sources, accepted for publication, Affective Computing and Intelligent Interaction 2011

our emotional dictionary

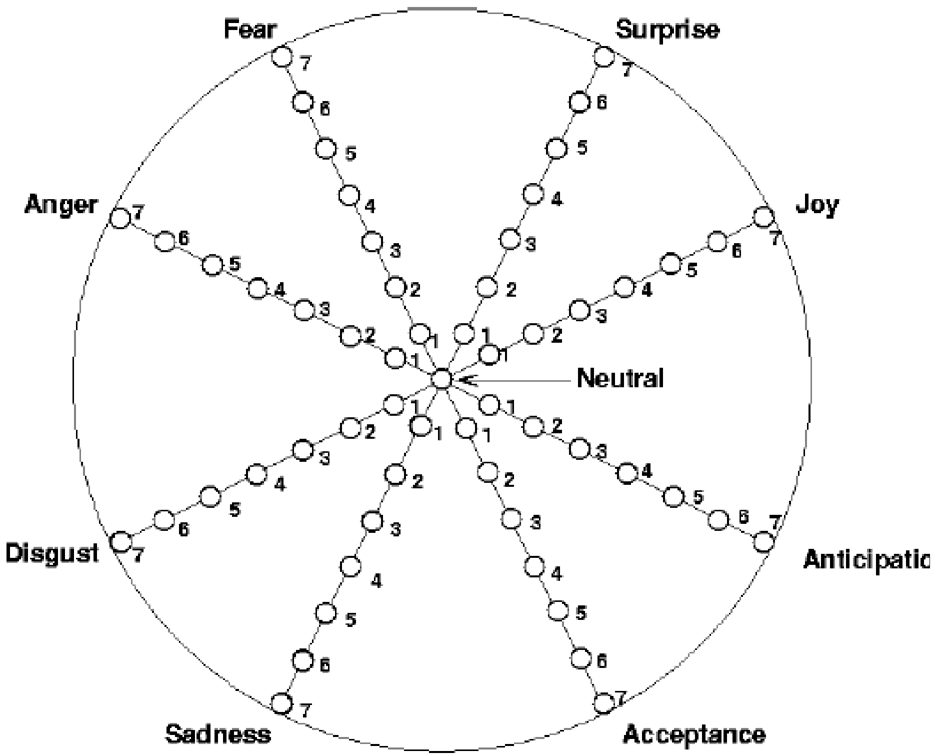
- create an extended emotional dictionary by enriching an opinion lexicon provided by the UMBC university with synonymous words from WordNet

1.0	:-)	
-0.25	:-O	
-1.0	:-(	
-0.5	:-|	
0.85	;-)	
0.5	(H)	

```
'0.95', 'fortunate'
'0.875', 'wonderfulness'
'0.8125', 'ideal'
'0.75', 'tastefulness'
'0.625', 'truthful'
'0.5625', 'delicious'
'0.5', 'wishful'
'-0.375', 'unhumorous'
'-0.520833333333', 'deadly'
'-0.5625', 'unstylish'
'-0.625', 'vanish'
'-0.9375', 'damaging'
```

The used affect space

- representing the extreme ends of four emotional pairs [Gill08]
- emotion exemplar words



acceptance	**fear**	**anger**	**joy**
acceptance	fear	anger	joy
agreement	phobia	rage	delight
affirmation	terror	fury	bliss
admission	fright	outrage	rejoicing
adoption	scare	hatred	elation
approval	dread	tantrum	gaiety
assent	nightmare	animosity	glee
anticipation	**sadness**	**disgust**	**surprise**
anticipation	sadness	disgust	surprise
awaiting	depression	revulsion	unexpected
expectancy	sorrow	distaste	unforeseen
prospect	melancholy	aversion	astonishment
hope	woe	loathing	shock
promise	grief	dislike	amazement
apprehension	mourning	nausea	incredulity

Relevance between tweet and emotion

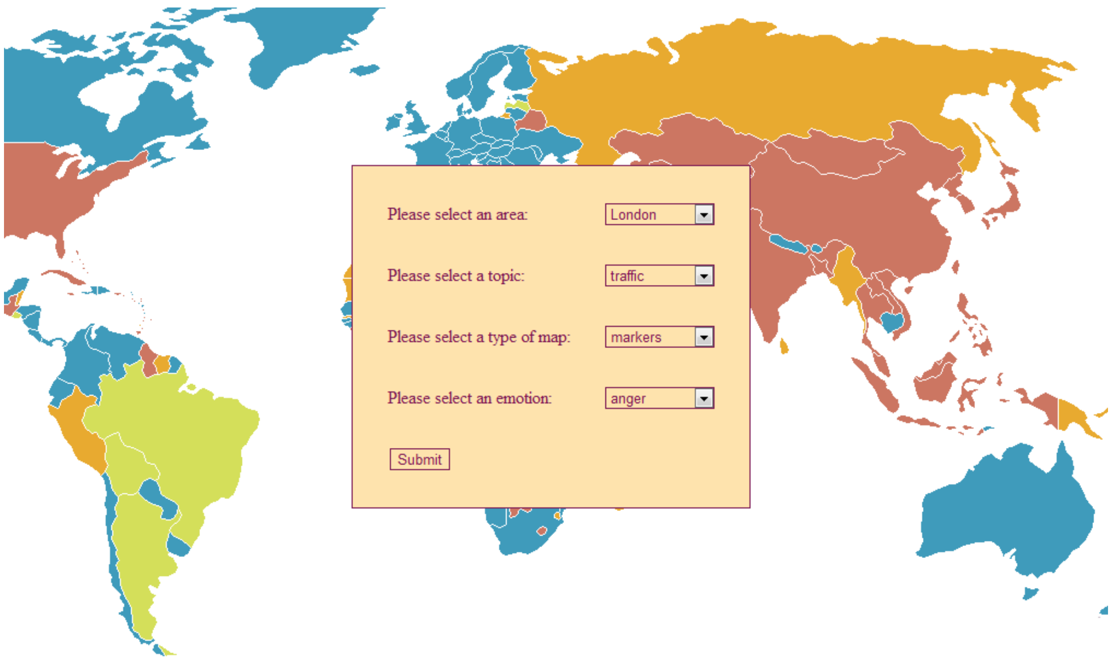
- Semantic similarity
  - the maximum semantic similarity between each of the tweet's *words* and the emotion's representatives, as defined in Wordnet
- Sentiment similarity
  - expresses a word's emotional intensity as defined in the extended dictionary
- Overall similarity between a tweet & an emotion

$$\frac{\Sigma_{i=1\ldots|words|} Sem(w_i,emotion) * Sent(w_i)}{|w_i| : Sent(w_i) \neq 0, \text{ for } 1 <= i <= |words|}$$

CapturEmos- Motivation

- capturing and understanding crowd’s emotions for a particular topic or product in an implicit manner via computational methods.
- *sentiment analysis & microblogging (statistical) processing* : emphasis on affective and opinion mining, lexicon-based processing, knowledge extraction techniques;
- *development of applications:* web application enhanced with of crowds emotions visualization capabilities.

The application/service



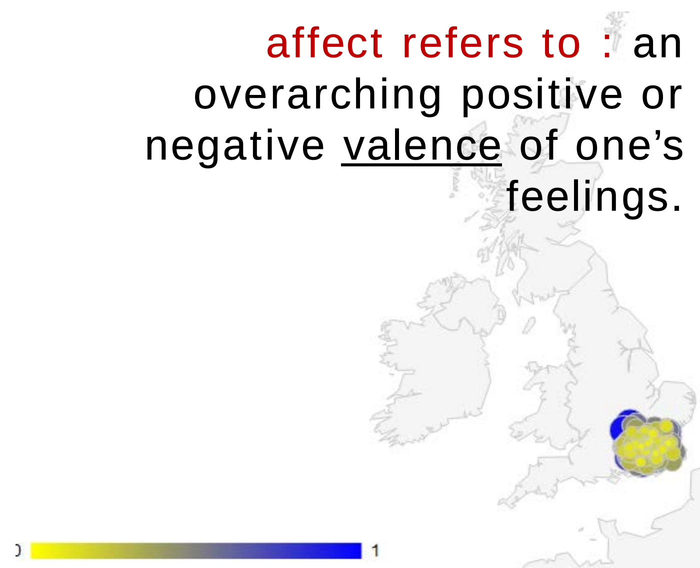
Our innovation principle :

focus on the "affect" which is distinguished from discrete emotions.

discrete emotions : concern affective reactions in relation to one's goals

affect refers to : an overarching positive or negative <u>valence</u> of one's feelings.

useful for capturing branding success & diffusion in the market, as expressed by the crowds emotions



Markets & challenges

Niche targeted Markets

political stakeholders, public authorities (e.g. municipalities), consumer behavior policy makers, chambers of commerce, tourism and infotainment sectors…

markets characteristics : emerging, unpredicted bursts, multi-profiles

Challenges : multi-lingual support; privacy and anonymity preservation; real-time emerging data flows; time and space complexities;

proposed framework and applications :

- address wide stakeholders and markets audiences;
- certain tasks can be realized (e.g. capturing branding success & diffusion in the market) expressed by the crowds emotions;
- can support policy and decision making.

Future work and horizon …

- emerging, and unpredicted bursts detections in evolving social media;
- user multi-profiles patterns;
- support applications with multi-lingual support;
- privacy and anonymity preservation
- development of intelligent and collective information retrieval techniques are required and well expected.

References

[Horrigan09] J. A. Horrigan, "Online shopping," Pew Internet & American Life Project Report, 2008.

[comScore07] comScore/the Kelsey group, "Online consumer-generated reviews have significant impact on offline purchase behavior," Press Release, http://www.comscore.com/press/release.asp?press=1928, November 2007.

[Lioyd05] Lloyd, L., Kechagias, D., Skiena, S.: Lydia: A system for large-scale news analysis. In: String Processing and Information Retrieval (SPIRE 2005). Volume Lecture Notes in Computer Science, 3772. (2005) 161-166

[Bautin08] M. Bautin, L. Vijayarenu, S. Skiena : International sentiment analysis for News and Blogs, In Proceedings of ICWSM (2008)

[Demartini10] G. Demartini and S. Siersdorfer: **Dear Search Engine: What's your opinion about...?** Sentiment Analysis for Semantic Enrichment of Web Search Results, In Proc. Of WWW201 0, April 2630, 2010

[Abbasi08] A. Abbasi, H. Chen, S. Thoms, T. Fu: Affect Analysis of Web Forums and Blogs Using Correlation Ensembles, IEEE Transactions on Knowledge and Data Engineering (2008) Vol. 20, Issue 9

[OM] Opinion Mining and Sentiment Analysis: NLP Meets Social Sciences, Bing Liu Department of Computer Science University Of Illinois at Chicago

[Sun 08] Sun A, Zeng D, Li H, Zheng X (2008) Discovering trends in collaborative tagging systems. In: Proceedings of the IEEE ISI 2008 PAISI, PACCF, and SOCO international workshops on Intelligence and Security Informatics, Springer, pp 377-383

[Wetzker08] Wetzker R, Plumbaum T, Korth A, Bauckhage C, Alpcan T, Metze F (2008a) Detecting trends in social bookmarking systems using a probabilistic generative model and smoothing. In: Proceedings of 19th International Conference on Pattern Recognition (ICPR 2008), IEEE, pp 1-4

[Hotho06] Hotho A, Jaschke R, Schmitz C, Stumme G (2006a) Information retrieval in folksonomies: Search and ranking. In: Proceedings of the 3rd European Semantic Web Conference, Springer, Budva, Montenegro, LNCS, vol 4011, pp 411-426

[Gill08] Gill, A.J., French R.M., Gergle, D., and Oberlander, J.: Identifying Emotional Characteristics from Short Blog Texts. Proc. of the 30th Annual Conf. of the Cognitive Science Society, Washington DC (2008) 2237–2242


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
tapadh leat
спасибо
dziękuję
bedankt
obrigado
sukriya
kop khun krap
go raibh maith agat
mochchakkeram
arigatō
dakujem
grazie
terima kasih
merci
мерси
감사합니다
dank je
hvala
sagolun
takk