SOCIAL NETWORKS:
EVOLVING DATA MINING and SENTIMENT ANALYTICS
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 twitter)
  - social bookmarking/tagging systems (e.g. del.icio.us flickr)

- Social data refer to different types of entities:
  - users
  - content handled separately or jointly
  - metadata

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

- 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

Explicit relations

Implicit relations

IR inference mechanism

**Diagram Details**

**Table 1:**

<table>
<thead>
<tr>
<th>i</th>
<th>$ER_i(u,u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>follows</td>
</tr>
<tr>
<td>2</td>
<td>is followed by</td>
</tr>
<tr>
<td>3</td>
<td>is friend with</td>
</tr>
<tr>
<td>4</td>
<td>comments on post of</td>
</tr>
</tbody>
</table>

**Table 2:**

<table>
<thead>
<tr>
<th>i</th>
<th>$ER_i(u,r)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>likes</td>
</tr>
<tr>
<td>2</td>
<td>uploads</td>
</tr>
<tr>
<td>3</td>
<td>leaves comment on</td>
</tr>
<tr>
<td>4</td>
<td>rates</td>
</tr>
<tr>
<td>5</td>
<td>downloads</td>
</tr>
<tr>
<td>6</td>
<td>assigns tag to</td>
</tr>
</tbody>
</table>

**Table 3:**

<table>
<thead>
<tr>
<th>i</th>
<th>$ER_i(u,m)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>creates (tag)</td>
</tr>
<tr>
<td>2</td>
<td>assigns to group (tag)</td>
</tr>
<tr>
<td>3</td>
<td>assigns on resource (tag)</td>
</tr>
<tr>
<td>4</td>
<td>makes on resource (comment)</td>
</tr>
</tbody>
</table>

**Table 4:**

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

**Symbols**

- $u$: user
- $r$: resource
- $m$: metadata
- $g$: group
Implicit relations uncovered for various types of entities

- 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
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

- Simple graphs’ structure can be encoded in an *adjacency matrices* if $G$: unweighted or *similarity matrices* if $G$: weighted
Folksonomy projections on simple graphs

Tag assignments:
- u1, technorati.com, search
- u1, technorati.com, web2.0
- u1, google.com, engine
- u2, google.com, search
- u2, technorati.com, blogs

Tripartite graph:
- Users: u1, u2
- Tags: t1, t2, t3, t4
- Resources: r1, r2
- Relations: ER(u,r): assigns tag to, ER(u,m): assigns on resource, IR(m,r): is assigned to

Unipartite graph:
- Tags: t1, t2, t3, t4
- Resources: r1, r2
- Relations: IR(m,m): are assigned by same user, IR(v,m): is applied on same resource
Evolving data representation structures

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

\[ G = \{ G_t \}, t \in \mathbb{N} \]
The snapshot layer

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

$\mathcal{G} = \{G_t\}, t \in \mathbb{N}$
The segment layer

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

Pre-processing technique
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: $G = \{G_t\}, t \in \mathbb{N}$

adjacency/similarity matrices
✓ more coarse-grained
✓ time-aggregated matrices are used for emphasizing on most recent edges [Tong08]
The stream layer

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

“multi-graphs” with edges encoding relations as well as temporal information [Zhao07]

time aggregate adjacency matrices [Tong08]

tensors


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

Web 2.0 data sources

Data collection

Data modeling
- Static structures
  - hypergraphs
  - folksonomies
  - simple graphs

Mining methods
- statistical analysis
- clustering & community detection
- classification
- association rules' mining

Exploitation in applications

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)
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

<table>
<thead>
<tr>
<th>method</th>
<th>structure</th>
<th>dataset</th>
<th>outcome/scope</th>
</tr>
</thead>
</table>
| **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.

<table>
<thead>
<tr>
<th>method</th>
<th>structure</th>
<th>dataset</th>
<th>outcome/scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chakrabarti</td>
<td>(time-aware) similarity matrix</td>
<td>photo sharing service</td>
<td>joint optimization of 2 criteria:</td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td>68 weeks</td>
<td>- snapshot quality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- history quality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>{5K tags}</td>
<td>proposed:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>{&quot;are applied on same resource&quot;}</td>
<td>- agglomerative hierarchical algorithm</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- K-means</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>method</th>
<th>structure or model</th>
<th>dataset</th>
<th>outcome/scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tang 2008</td>
<td>series of network snapshots; interaction matrix for each snapshot</td>
<td>I. mail exchange network II. co-authorship network</td>
<td>community evolution in dynamic networks of multiple social entities; iterative approximation of community evolution using <em>eigenvector calculation</em> and K-means clustering</td>
</tr>
</tbody>
</table>

**Evolutionary spectral clustering applied in multi-mode networks**

<table>
<thead>
<tr>
<th>source</th>
<th>period</th>
<th>entities / relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>12 months</td>
<td>{2.4K users} {emails} {36.7K words} {“sends email”} {“receives email”} {“contains term”}</td>
</tr>
<tr>
<td>II.</td>
<td>25 years</td>
<td>{492K papers} {347K authors} {2.8K venues} {9.5K title terms} {“writes”} {“participates in”} {“is published in”} {“contains term”}</td>
</tr>
</tbody>
</table>
## NFC approach

<table>
<thead>
<tr>
<th>method</th>
<th>structure or model</th>
<th>dataset</th>
<th>outcome/scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin 2009</td>
<td><strong>metagraph</strong>: hypergraph with nodes representing facets and edges multipartite interactions; tensors</td>
<td>source: social bookmarking service with voting capabilities, period: 27 days (9 timeslots), entities: {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”}</td>
<td>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</td>
</tr>
</tbody>
</table>
# SGC approach

<table>
<thead>
<tr>
<th>method</th>
<th>structure</th>
<th>dataset</th>
<th>outcome/scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun 2007</td>
<td>unweighted undirected bipartite graphs; graph segments</td>
<td></td>
<td>unparametric method based on Minimum Description Length; each segment’s source and destination nodes are partitioned separately to decrease cost; compressed graph in &lt; 4% than original space</td>
</tr>
<tr>
<td></td>
<td>I. mail exchange network</td>
<td>I. 165 weeks</td>
<td>{34.3K senders} {34.3K recipients} {15K “send mail” / week}</td>
</tr>
<tr>
<td></td>
<td>II. mobile phone call network</td>
<td>II. 46 weeks</td>
<td>{96 callers} {3.8K callees} {430 “calls”/ week}</td>
</tr>
<tr>
<td></td>
<td>III. mobile device proximity records</td>
<td>III. 46 weeks</td>
<td>{96 users} {96 users} {689 “is located near”/week}</td>
</tr>
</tbody>
</table>
References


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

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

- Sun and colleagues [Sun08] use the $\chi^2$ 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

Clusters structures formation

conventional clustering

co-clustering

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.
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 [Lloyd05] 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

![Map of George Bush: 03/01/2007 - 03/28/2007](image)

![Graph of Polarity Score](image)

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.

<table>
<thead>
<tr>
<th>Affect class</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>0.01</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.03</td>
</tr>
<tr>
<td>Anger</td>
<td>0.6</td>
</tr>
<tr>
<td>Hate</td>
<td>0.5</td>
</tr>
</tbody>
</table>
- 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:

- social network analysis
  [image from Touchgraph]

- event detection
  [diagram from Sun07]

- clustering of users
  exploiting the time dimension

- trend detection
  [image from Trendsmap]

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 structure and frequency as they evolve 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

Cloud4Trends is a microblogging & blogging localized content collection and analysis framework for detecting currently popular topics of users’ interest
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.
Cloud4Trends: Localized Real-Time Trend Detection In Social Media

Create New

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Location</th>
<th>Coordinates</th>
<th>Initiation date/time</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Another demo experiment</td>
<td>A London experiment</td>
<td>gb london</td>
<td>-0.56,51.26,0.28,51.68</td>
<td>6/27/2012 1:39:47 PM</td>
<td>Running</td>
</tr>
<tr>
<td>My demo experiment</td>
<td>A demo experiment</td>
<td>us new york</td>
<td>-74,40,-73,41</td>
<td>6/22/2012 3:39:28 PM</td>
<td>Running</td>
</tr>
<tr>
<td>A london experiment</td>
<td>Experiment</td>
<td>gb london</td>
<td>-0.56,51.26,0.28,51.68</td>
<td>6/22/2012 8:12:33 PM</td>
<td>Stopped</td>
</tr>
</tbody>
</table>

OSWINDS
OPERATING SYSTEMS WEB/INTERNET DATA SOURCES MANAGEMENT

Cloud4Trends: Localized Real-Time Trend Detection In Social Media

Trends from Blogger

1. kiss single monst

2. scouting report golf

3. childhood retroblog
Emotional Aware Clustering on Micro-Blogging Sources (affect analysis)

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.
The used affect space

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

\[ \sum_{i=1}^{\text{|words|}} \text{Sem}(w_i, \text{emotion}) \times \text{Sent}(w_i) \]

| wi | : Sent(w_i) ≠ 0, for 1 ≤ i ≤ | words |
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 valence of one’s feelings.

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


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


