IR in Social Media

Alexey Maykov, Matthew Hurst, Aleksander Kolcz

Microsoft Live Labs

Outline

- Session 1: Overview, Applications and Architectures (for social media analysis)
- In-Depth 1: Data Acquisition
- Session 2: Methods
  - Graphs
  - Content
- In-Depth 2: Link Counting

Outline

- Session 1: Overview, Applications and Architectures (for social media analysis)
- In-Depth 1: Data Acquisition
- Session 2: Methods
  - Graphs
  - Content
- In-Depth 2: Data Preparation

Session 1 Outline

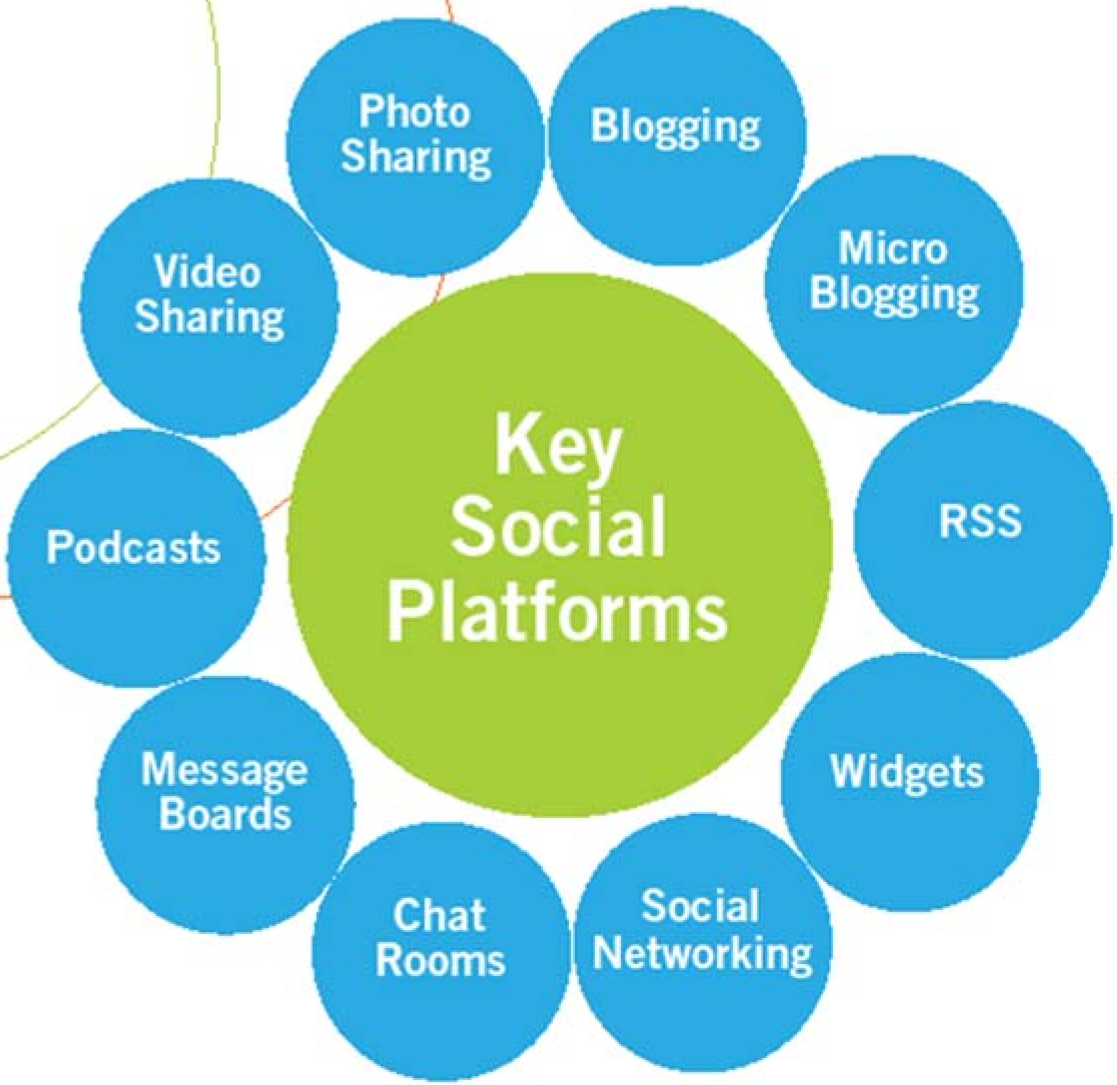
- Introduction
- Applications
- Architectures

Session 1 Outline

- Introduction
- Applications
- Architectures

Definitions

- What is social media?
  - By example: blogs, usenet, forums
  - Anything which can be spammed!
- Social Media vs Mass Media
  - http://caffertyfile.blogs.cnn.com/
  - http://www.exit133.com/



Key Social Platforms

Photo Sharing

Blogging

Micro Blogging

RSS

Widgets

Social Networking

Chat Rooms

Message Boards

Podcasts

Video Sharing

Key Features

- Many commonly cited features:
  - Creator: non professional (generally)
  - Intention: share opinions, stories with small(ish) community.
  - Etc.
- Two Important features:
  - Informal: doesn’t mean low quality, but certainly fewer barriers to publication (c.f. editorial review…)
  - Ability of audience to respond (comments, trackbacks/other blog posts, …)

- And so it went in the US media: silence, indifference, with a dash of perverse misinterpretation. Consider Michael Hirsh's laughably naive commentary that imagined Bush had already succeeded in nailing down SOFA, to the chagrin of Democrats.

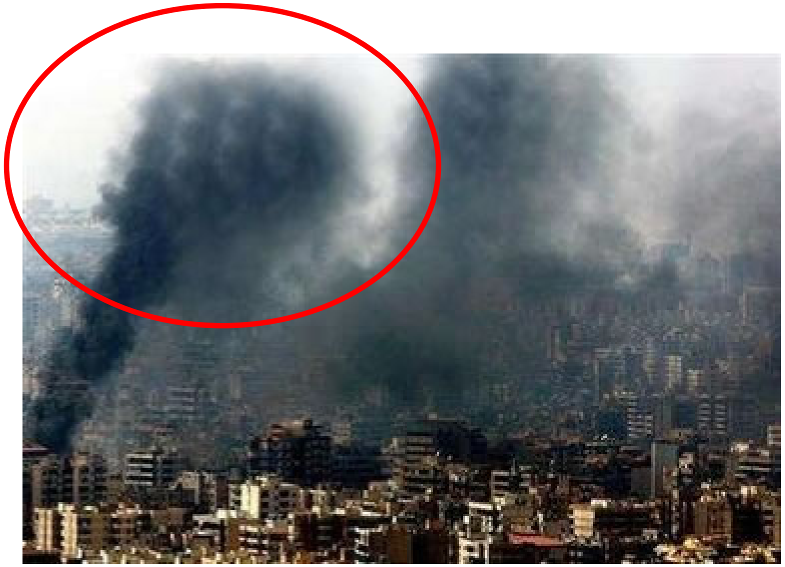
- DailyKos – smintheus, Jun 15 2008

Impact

- New textual web content: social media accounts for 5 times as much as ‘professional’ content now being created (Tomkins et al; ‘People Web’).
- A number of celebrated news related stories surfaced in social media.

Reuters and Photoshop

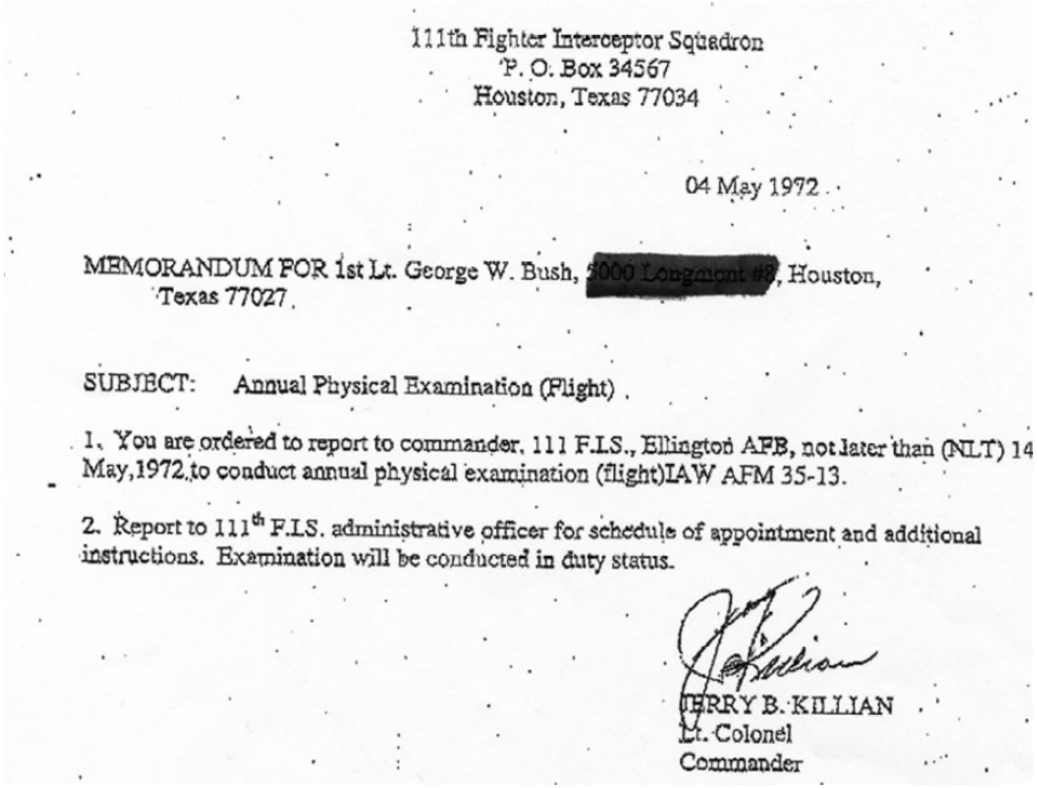
- Note copied smoke areas…



Surfaced on LittleGreenFootballs.com to the embarrassment of Reuters.
http://littlegreenfootballs.com/weblog/?entry=21956_Reuters_Doctoring_Photos_from_Beirut&only

Rathergate

- Bloggers spotted a fake memo which CBS (Dan Rather) had failed to fact check/verify.



111th Fighter Interceptor Squadron
P. O. Box 34567
Houston, Texas 77034

04 May 1972

MEMORANDUM FOR 1st Lt. George W. Bush, [redacted], Houston, Texas 77027

SUBJECT: Annual Physical Examination (Flight)

1. You are ordered to report to commander, 111 F.I.S., Ellington AFB, not later than (NLT) 14 May,1972,to conduct annual physical examination (flight)IAW AFM 35-13.

2. Report to 111th F.I.S. administrative officer for schedule of appointment and additional instructions. Examination will be conducted in duty status.

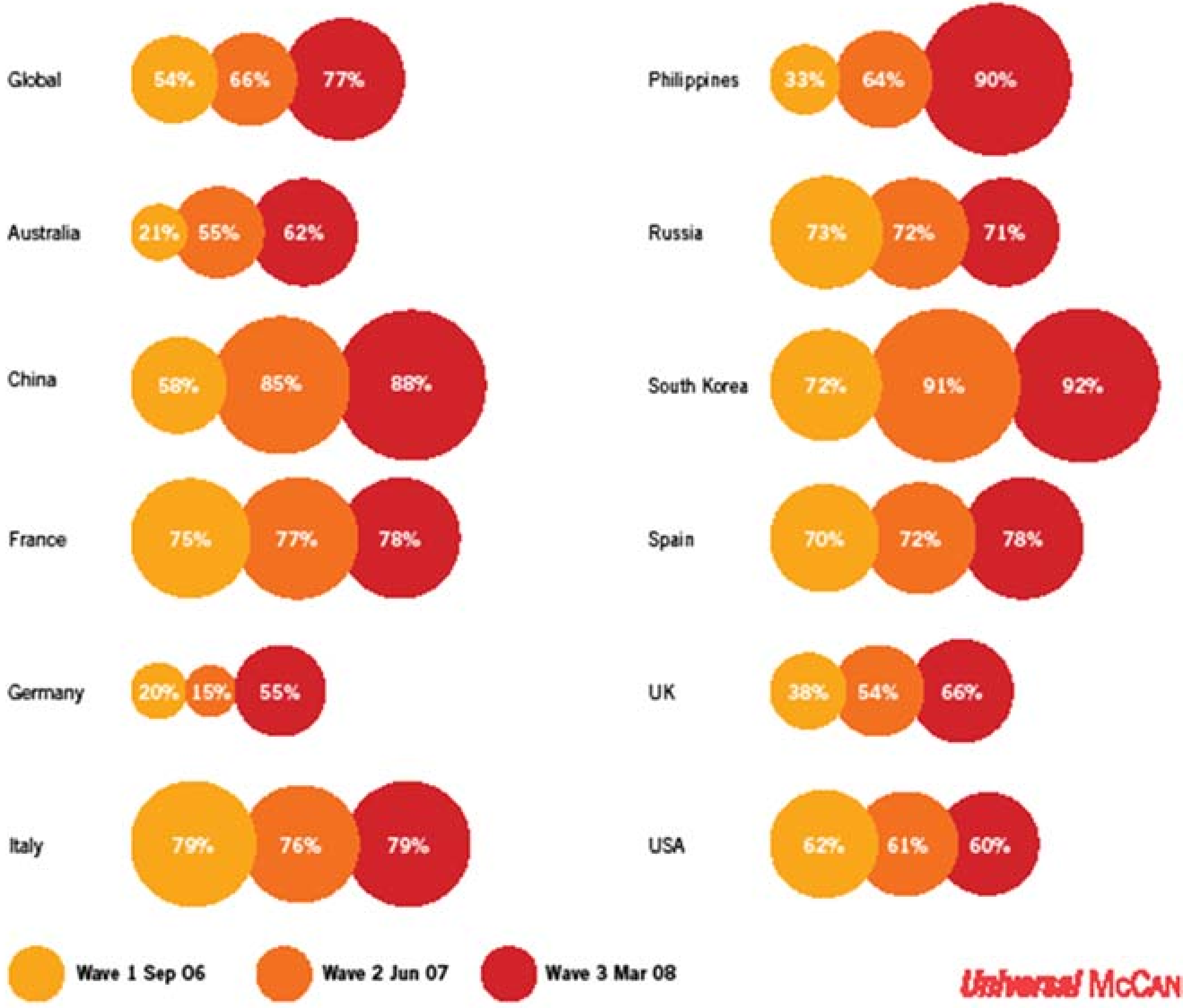
JERRY B. KILLIAN
Lt. Colonel
Commander

Impact Continued

- Recent work (McGlohon) establishes that political Usenet groups have decreasing links to MSM but increasing links to social media (weblogs).

Blog readership Waves 1-3

"Thinking about using the Internet, which of the following have you ever done?" – "Read blogs / weblogs" Active Internet Universe

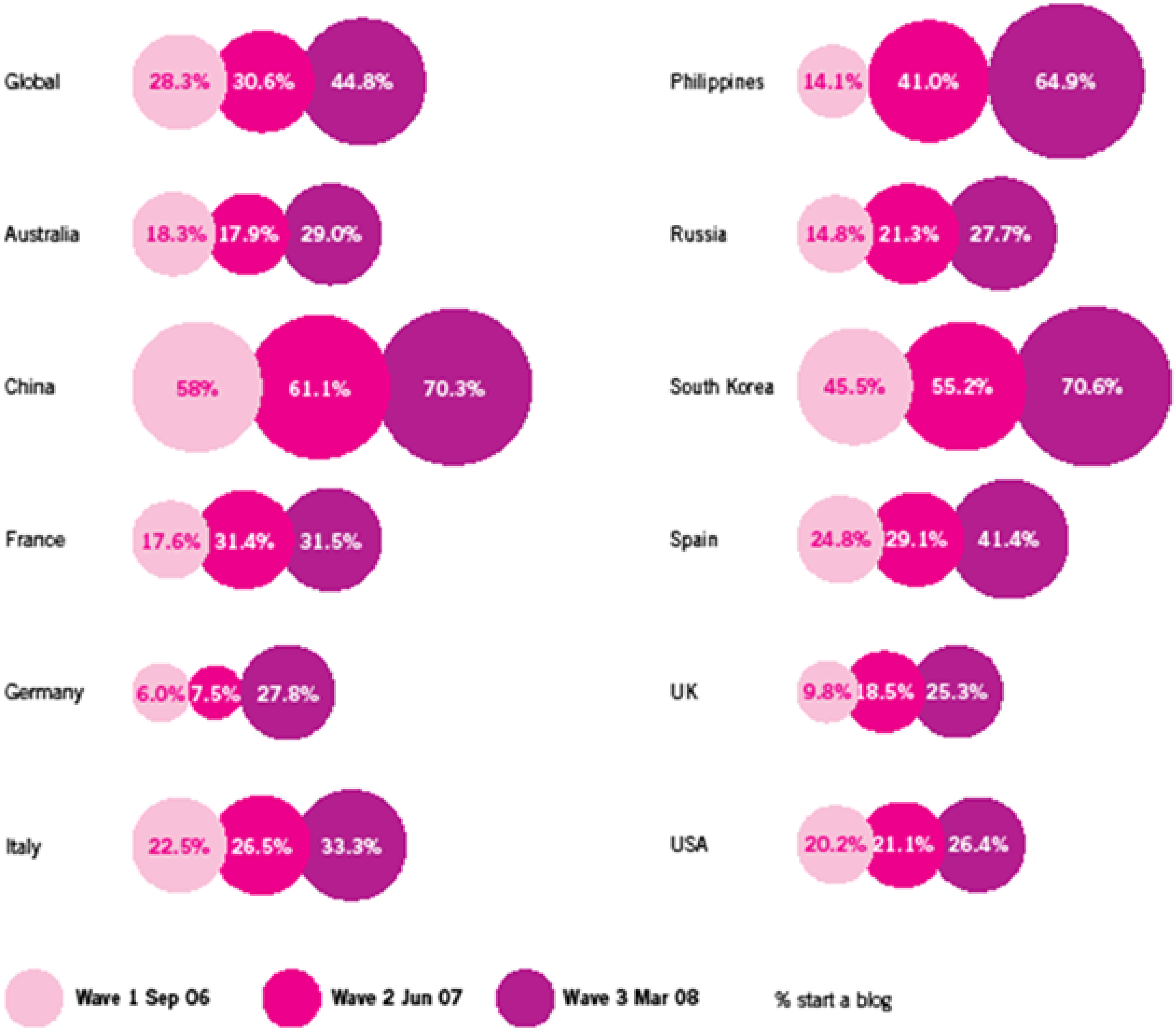


Universal McCANN | NEXT THING NOW

Writing blogs: usage trends

Blog writing Waves 1-3

“Thinking about using the Internet, which of the following have you ever done?” – “Start my own blog / weblog” Active Internet Universe



Power to the people - Social Media Tracker Wave 3

Academia

- <<Analysis of Social Media>> taught by William Cohen and Natalie Glance at CMU
- <<Networks: Theory and Application>> Lada Adamic, U of Mi
- UMBC eBiquity group

@VISIBLE TECHNOLOGIES

nielsen

Cymfony

umbria
A Division of J.D. Power and Associates

Jodange

SENTIMENT METRICS

SentiMetrix

zeta
INTERACTIVE

Shift Happens.
RelevantNoise
is now Zeta
Interactive.

+ Find out
why we're
what's next

BUZZLOGIC
INFLUENCE IS GOLDEN

ATTENTIO

attensity

reputica

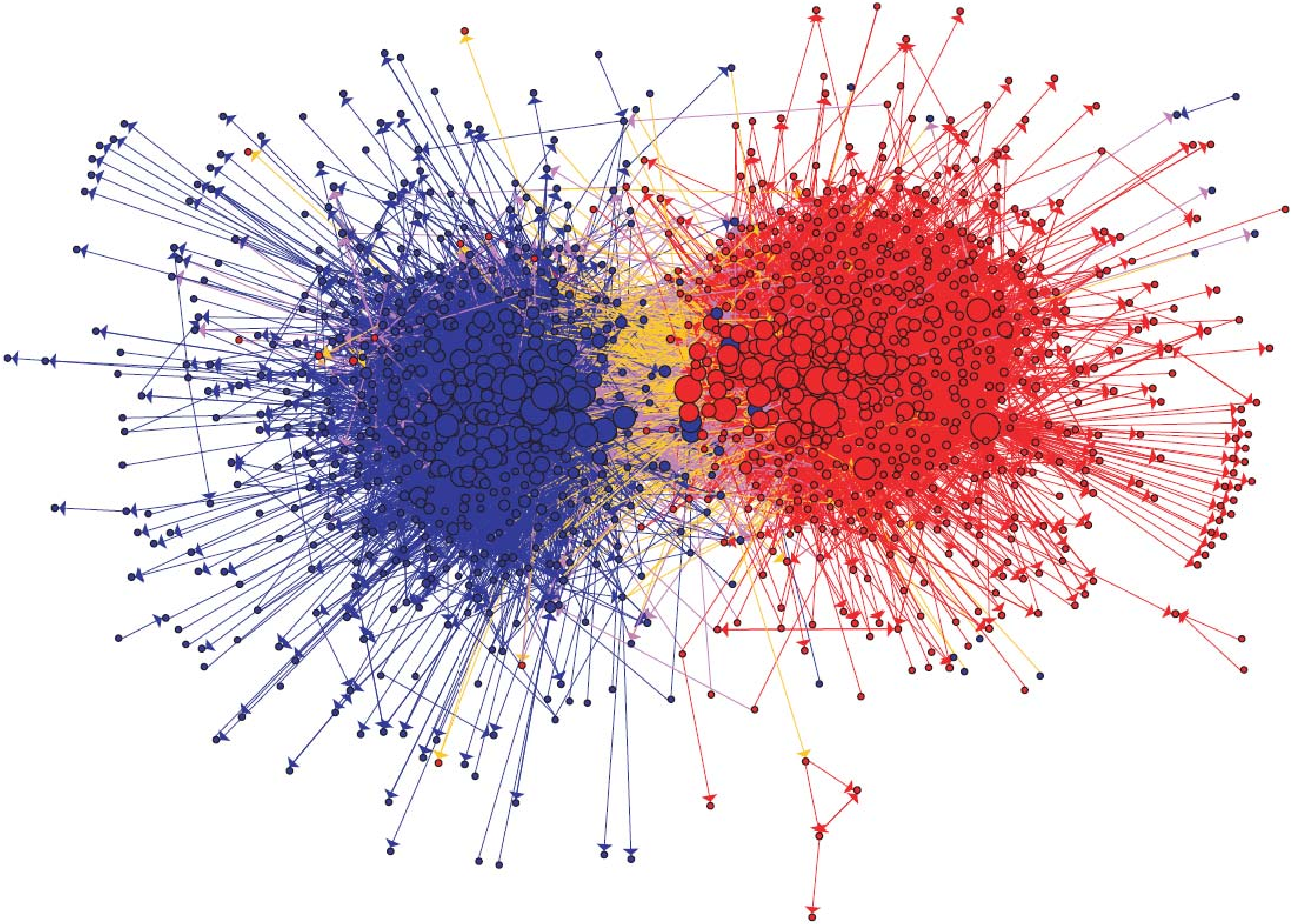
andiamo!™
systems
Social Media Measurement... Simplified

PORTENT INTERACTIVE

pythia
BETA

Conferences

- ICWSM
- Social Networks and Web 2.0 track at WWW



Session 1 Outline

- Introduction
- Applications
- Architectures

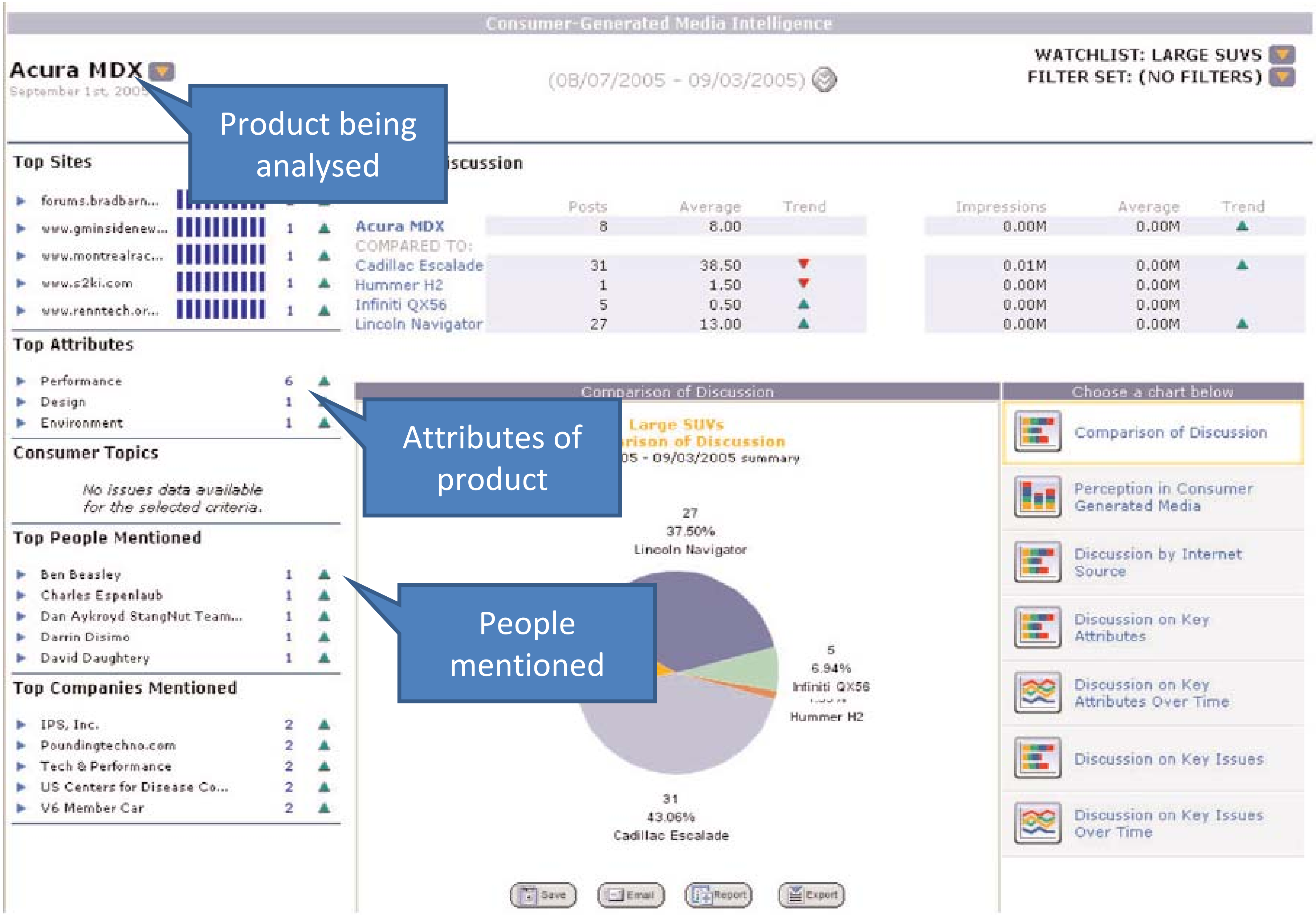
Applications 1: BI

- Business Intelligence over Social Media promises:
  - Tracking attention to your brand or product
  - Assessing opinion wrt brand, product or components of the product (e.g. ‘the battery life sucks!’)
  - Comparing your brand/product with others in the category
  - Finding communities critical to the success of your business.

VISIBLE TECHNOLOGIES
nielsen
Cymfony
umbria
A Division of J.D. Power and Associates
Jodange
SENTIMENT METRICS
SentiMetrix
zeta INTERACTIVE
BUZZLOGIC
INFLUENCE IS GOLDEN
ATTENTIO
attensity
reputica
andiamo! systems
Social Media Measurement... Simplified
PORTENT INTERACTIVE
pythia BETA

Shift Happens. RelevantNoise is now Zeta Interactive.

+ Find out why we're what's next



Consumer-Generated Media Intelligence

Acura MDX ▾

September 1st, 2005

(08/07/2005 - 09/03/2005)

WATCHLIST: LARGE SUVS ▾

FILTER SET: (NO FILTERS) ▾

Top Sites

forums.bradbarn...		
www.gminsidenew...	1	▲
www.montrealrac...	1	▲
www.s2ki.com	1	▲
www.renntech.or...	1	▲

Top Attributes

Performance	6	▲
Design	1	
Environment	1	▲

Consumer Topics

No issues data available for the selected criteria.

Top People Mentioned

Ben Beasley	1	▲
Charles Espenlaub	1	▲
Dan Aykroyd StangNut Team...	1	▲
Darrin Disimo	1	▲
David Daughtery	1	▲

Top Companies Mentioned

IPS, Inc.	2	▲
Poundingtechno.com	2	▲
Tech & Performance	2	▲
US Centers for Disease Co...	2	▲
V6 Member Car	2	▲

iscussion

	Posts	Average	Trend	Impressions	Average	Trend
Acura MDX	8	8.00		0.00M	0.00M	▲
COMPARED TO:						
Cadillac Escalade	31	38.50	▼	0.01M	0.00M	▲
Hummer H2	1	1.50	▼	0.00M	0.00M	
Infiniti QX56	5	0.50	▲	0.00M	0.00M	
Lincoln Navigator	27	13.00	▲	0.00M	0.00M	▲

Comparison of Discussion

Large SUVs

rison of Discussion

05 - 09/03/2005 summary

Choose a chart below

- Comparison of Discussion
- Perception in Consumer Generated Media
- Discussion by Internet Source
- Discussion on Key Attributes
- Discussion on Key Attributes Over Time
- Discussion on Key Issues
- Discussion on Key Issues Over Time

Product being analysed

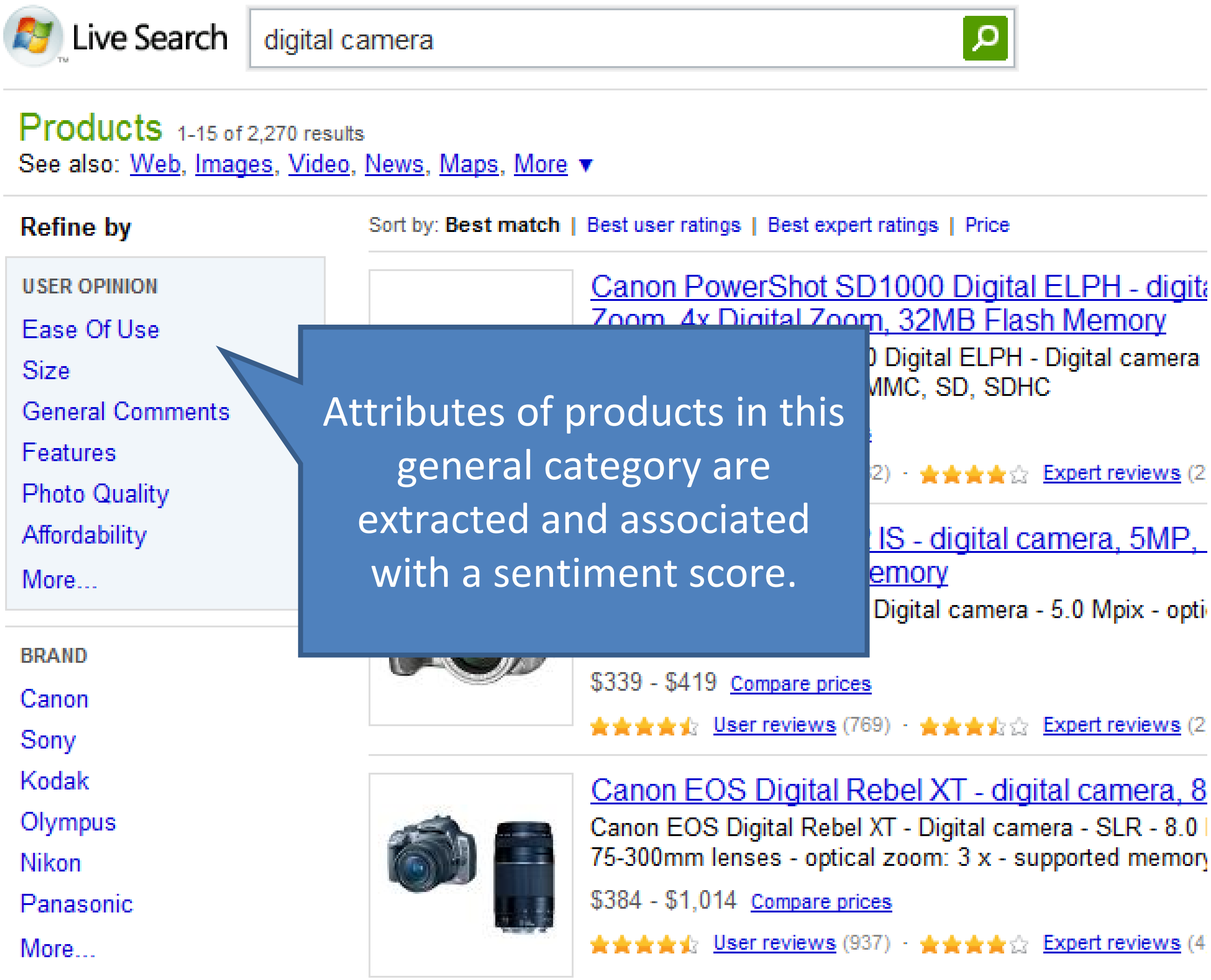
Attributes of product

People mentioned

Source: Cymfony, Inc.

Applications 2: Consumer

- Aggregating reviews to provide consumers with summary insights to help with purchase decisions.



Live Search

digital camera

Products 1-15 of 2,270 results

See also: Web, Images, Video, News, Maps, More ▼

Refine by

USER OPINION

- Ease Of Use
- Size
- General Comments
- Features
- Photo Quality
- Affordability
- More...

BRAND

- Canon
- Sony
- Kodak
- Olympus
- Nikon
- Panasonic
- More...

Sort by: **Best match** | Best user ratings | Best expert ratings | Price

Canon PowerShot SD1000 Digital ELPH - digit
Zoom, 4x Digital Zoom, 32MB Flash Memory

) Digital ELPH - Digital camera
MMC, SD, SDHC

2) · ★★★★☆ Expert reviews (2

IS - digital camera, 5MP,
emory

Digital camera - 5.0 Mpix - opti

$339 - $419 Compare prices

★★★★½ User reviews (769) · ★★★½☆ Expert reviews (2

Canon EOS Digital Rebel XT - digital camera, 8

Canon EOS Digital Rebel XT - Digital camera - SLR - 8.0
75-300mm lenses - optical zoom: 3 x - supported memor

$384 - $1,014 Compare prices

★★★★½ User reviews (937) · ★★★★☆ Expert reviews (4

Attributes of products in this general category are extracted and associated with a sentiment score.

Applications (addtl)

- Trend Analysis
- Ad selection
- Search
- Many more!

Session 1 Outline

- Introduction
- Applications
- Architectures

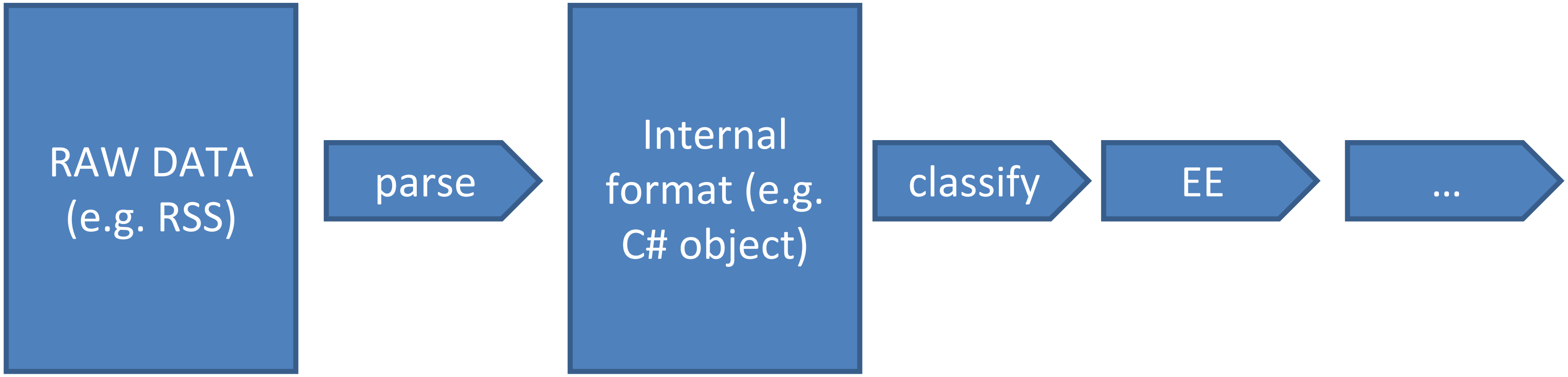
Functional Components

- Acquisition: getting that data in from the cloud.
- Content Preparation: translating the data in to an internal format; enriching the data.
- Content Storage: preserving that data in a manner that allows for access via an API.
- Mining/Applications

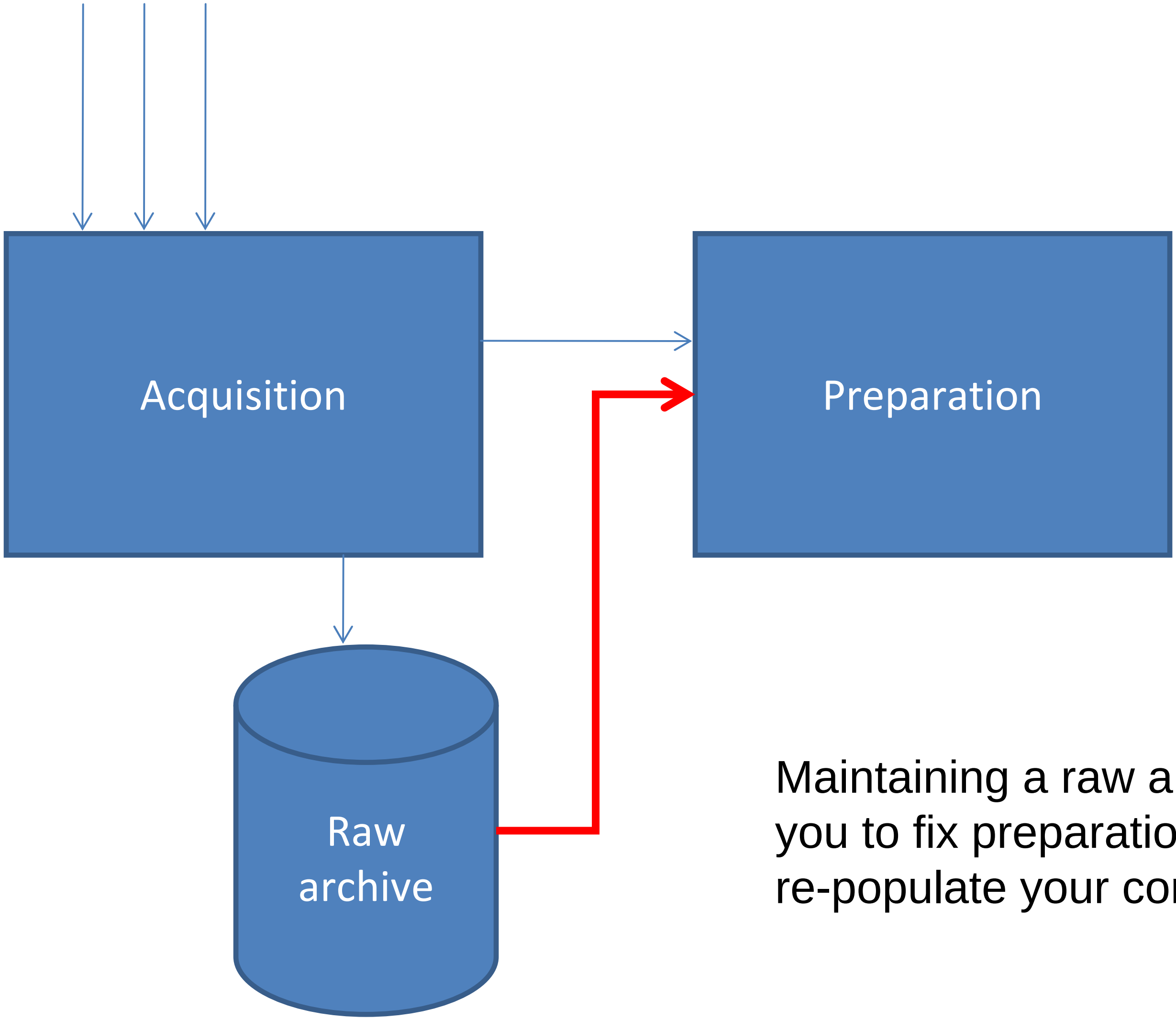
Focus on Content Preparation

- In general, it is useful to have a richly annotated content store:
  - Language of each document
  - Content annotations (named entities, links, keywords)
  - Topical and other classifications
  - Sentiment
- However, committing these processes higher up stream means that fixing issues with the data may be more expensive.

Focus on Content Preparation (cont)



Challenge: what happens if you improve your classifier, or if your EE process contains a bug?



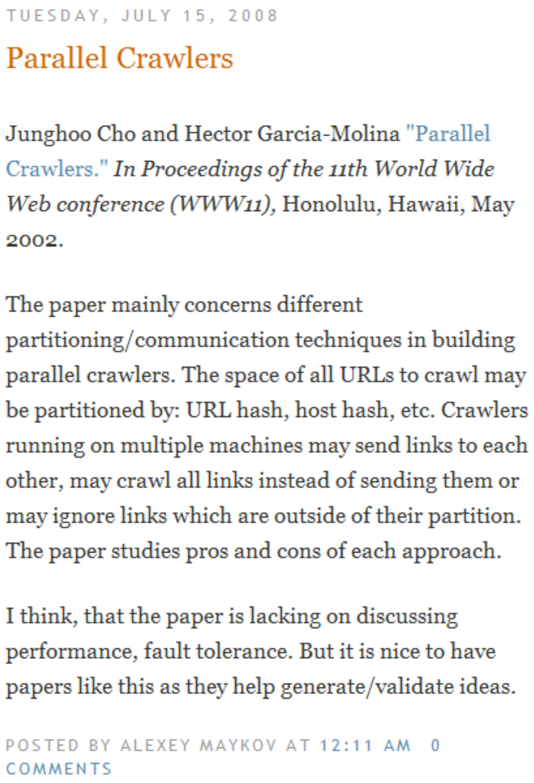
Maintaining a raw archive allows you to fix preparation issue and re-populate your content store.

Challenges

- How to deal with new data types
- How to deal with heterogeneous data (a weblog is not a message board)
- What are duplicates?
    - How does their definition impact analysis

New Data

Blog



TUESDAY, JULY 15, 2008

Parallel Crawlers

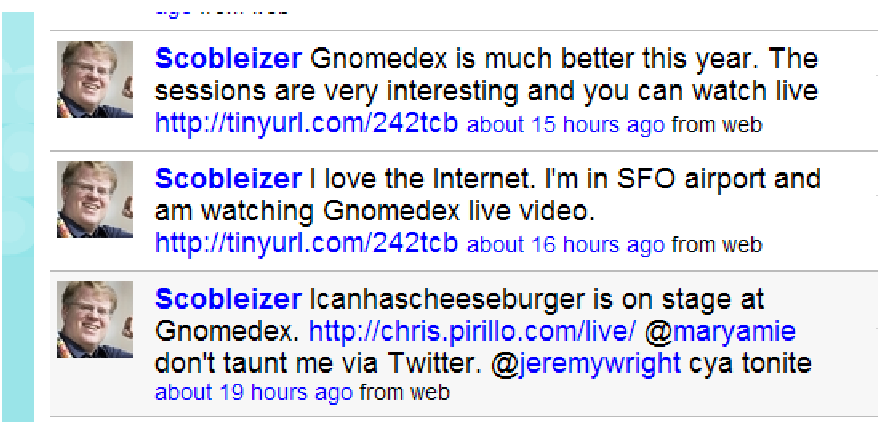
Junghoo Cho and Hector Garcia-Molina "Parallel Crawlers." *In Proceedings of the 11th World Wide Web conference (WWW11),* Honolulu, Hawaii, May 2002.

The paper mainly concerns different partitioning/communication techniques in building parallel crawlers. The space of all URLs to crawl may be partitioned by: URL hash, host hash, etc. Crawlers running on multiple machines may send links to each other, may crawl all links instead of sending them or may ignore links which are outside of their partition. The paper studies pros and cons of each approach.

I think, that the paper is lacking on discussing performance, fault tolerance. But it is nice to have papers like this as they help generate/validate ideas.

POSTED BY ALEXEY MAYKOV AT 12:11 AM 0 COMMENTS

Microblog



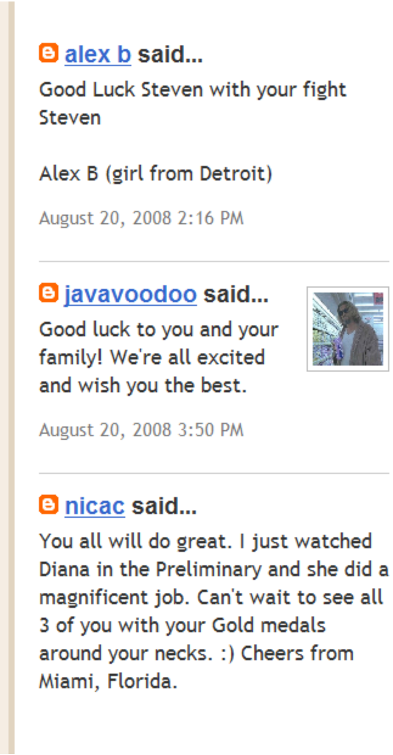
Scobleizer Gnomedex is much better this year. The sessions are very interesting and you can watch live http://tinyurl.com/242tcb about 15 hours ago from web

Scobleizer I love the Internet. I'm in SFO airport and am watching Gnomedex live video. http://tinyurl.com/242tcb about 16 hours ago from web

Scobleizer Icanhascheeseburger is on stage at Gnomedex. http://chris.pirillo.com/live/ @maryamie don't taunt me via Twitter. @jeremywright cya tonite about 19 hours ago from web

Heterogeneous Data

Blogger comments



alex b said...

Good Luck Steven with your fight Steven

Alex B (girl from Detroit)

August 20, 2008 2:16 PM

javavoodoo said...

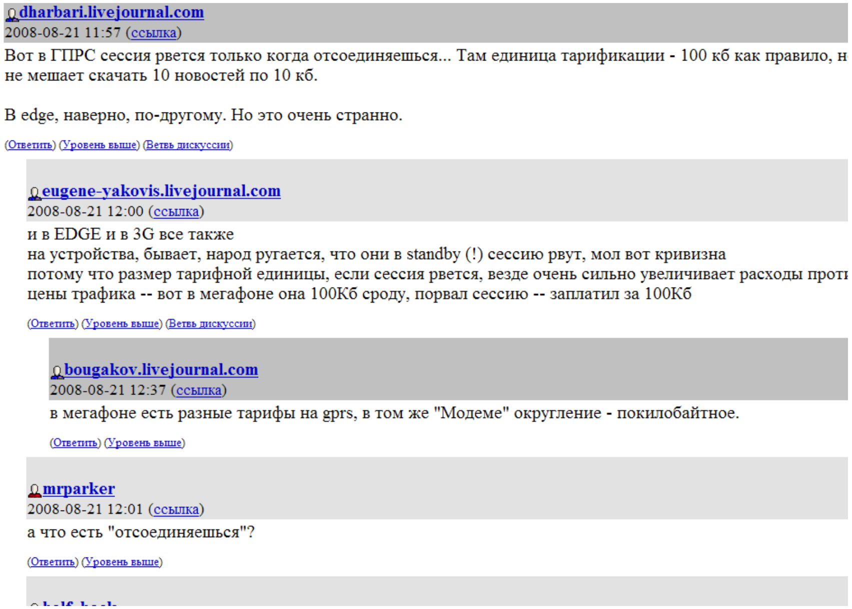
Good luck to you and your family! We're all excited and wish you the best.

August 20, 2008 3:50 PM

nicac said...

You all will do great. I just watched Diana in the Preliminary and she did a magnificent job. Can't wait to see all 3 of you with your Gold medals around your necks. :) Cheers from Miami, Florida.

Forum, LJ comments



dharbari.livejournal.com
2008-08-21 11:57 (ссылка)

Вот в ГПРС сессия рвется только когда отсоединяешься... Там единица тарификации - 100 кб как правило, н
не мешает скачать 10 новостей по 10 кб.

В edge, наверно, по-другому. Но это очень странно.

(Ответить) (Уровень выше) (Ветвь дискуссии)

eugene-yakovis.livejournal.com
2008-08-21 12:00 (ссылка)

и в EDGE и в 3G все также
на устройства, бывает, народ ругается, что они в standby (!) сессию рвут, мол вот кривизна
потому что размер тарифной единицы, если сессия рвется, везде очень сильно увеличивает расходы проти
цены трафика -- вот в мегафоне она 100Кб сроду, порвал сессию -- заплатил за 100Кб

(Ответить) (Уровень выше) (Ветвь дискуссии)

bougakov.livejournal.com
2008-08-21 12:37 (ссылка)

в мегафоне есть разные тарифы на gprs, в том же "Модеме" округление - покилобайтное.

(Ответить) (Уровень выше)

mrparker
2008-08-21 12:01 (ссылка)

а что есть "отсоединяешься"?

(Ответить) (Уровень выше)

half-back

Heterogeneous Data (solution)

- Containment Hierarchy
  - BlogHost->Blog->Post->Comment
  - ForumHost->Forum->Topic->Post*
- Contributors
  - name@container

Sources of Duplication

- Multiple crawl of the same content
- Cross-postings
- Signature lines

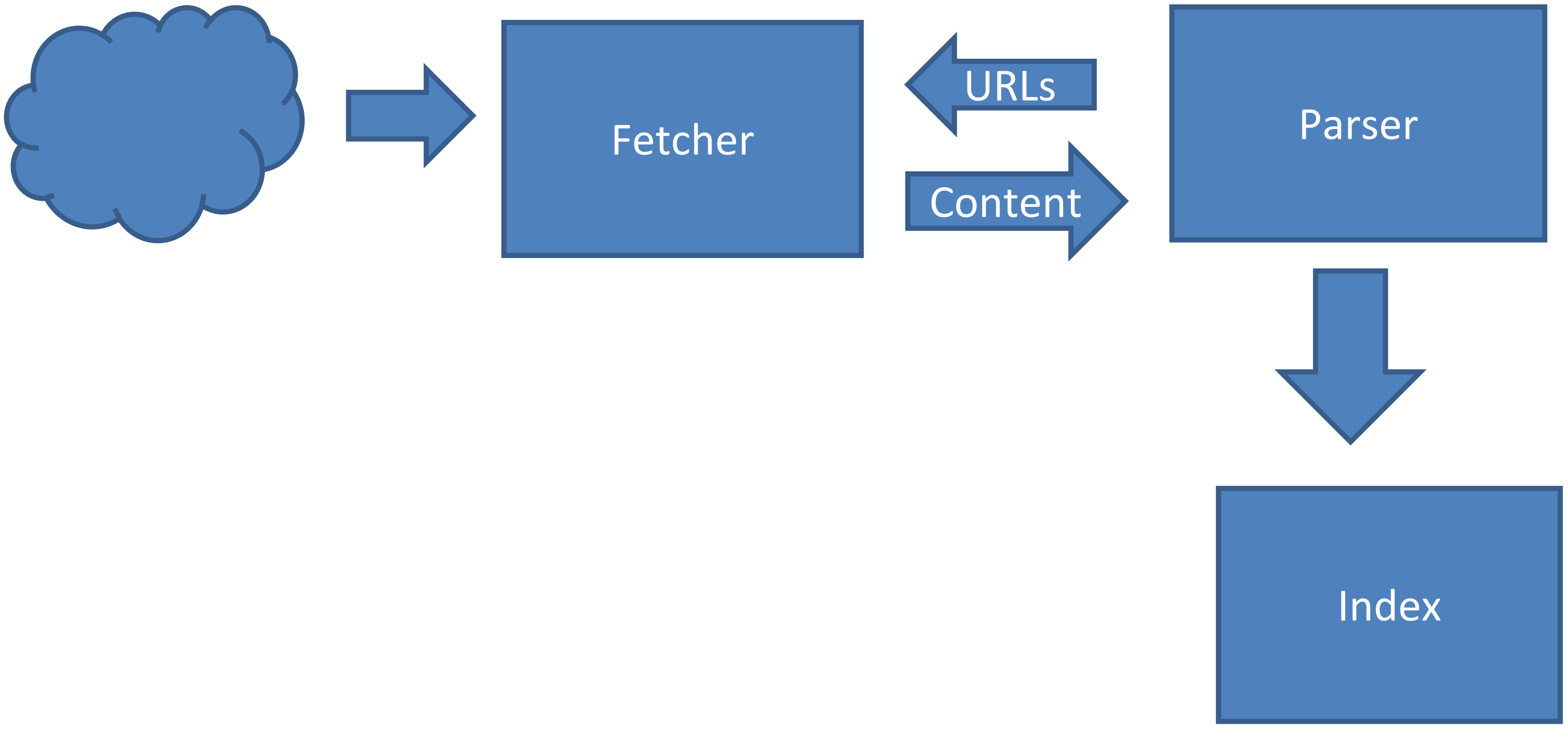
Outline

- Session 1: Overview, Applications and Architectures (for social media analysis)
- In-Depth 1: Data Acquisition
- Session 2: Methods
  - Graphs
  - Content
- In-Depth 2: Link Counting

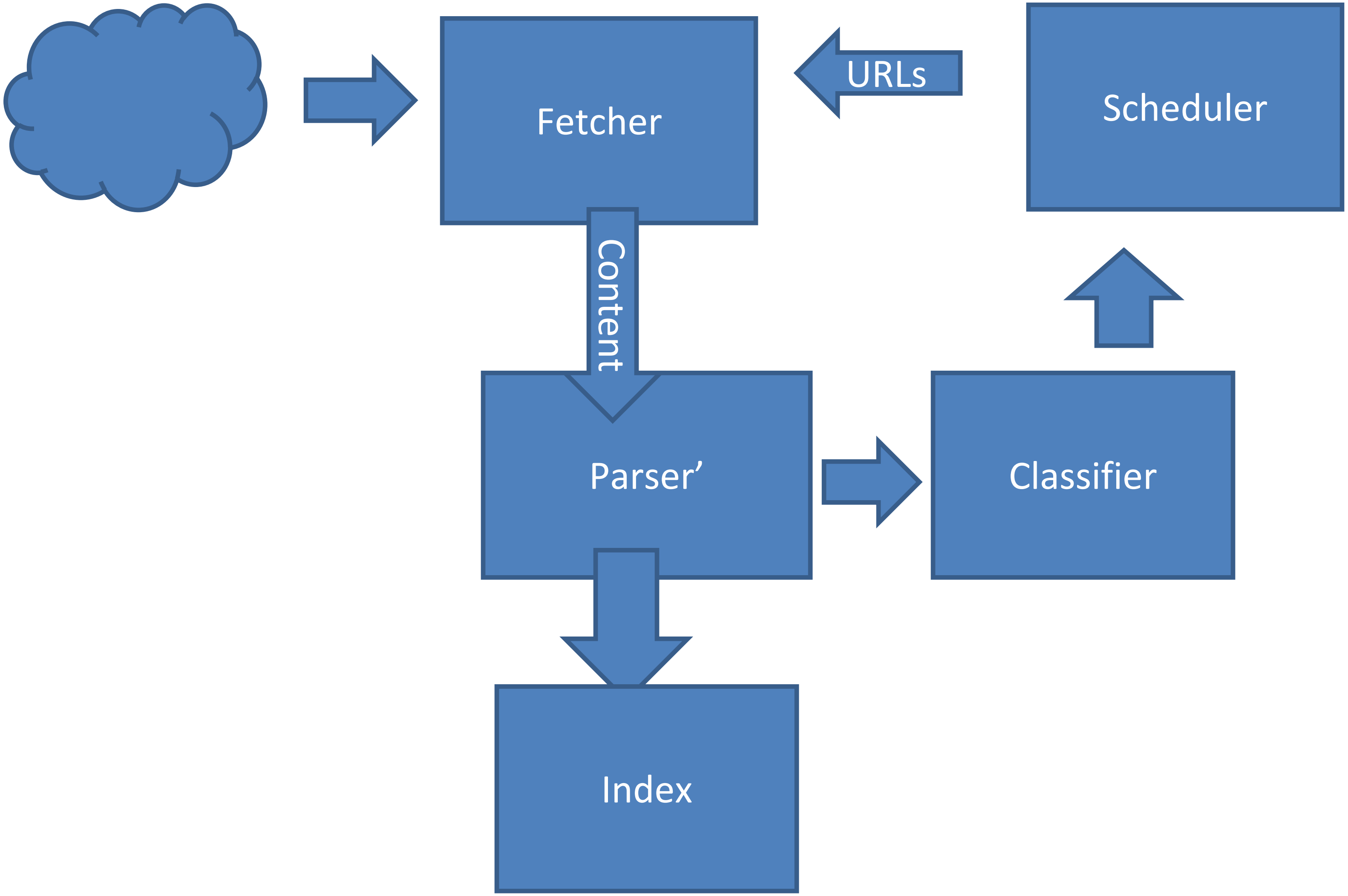
What to Crawl

- HTML
- RSS/Atom
- Private Feeds
  - 6apart: LiveJournal, TypePad, VOX
  - Twitter

Web Crawler



Blog Crawler



Fetcher

URLs

Scheduler

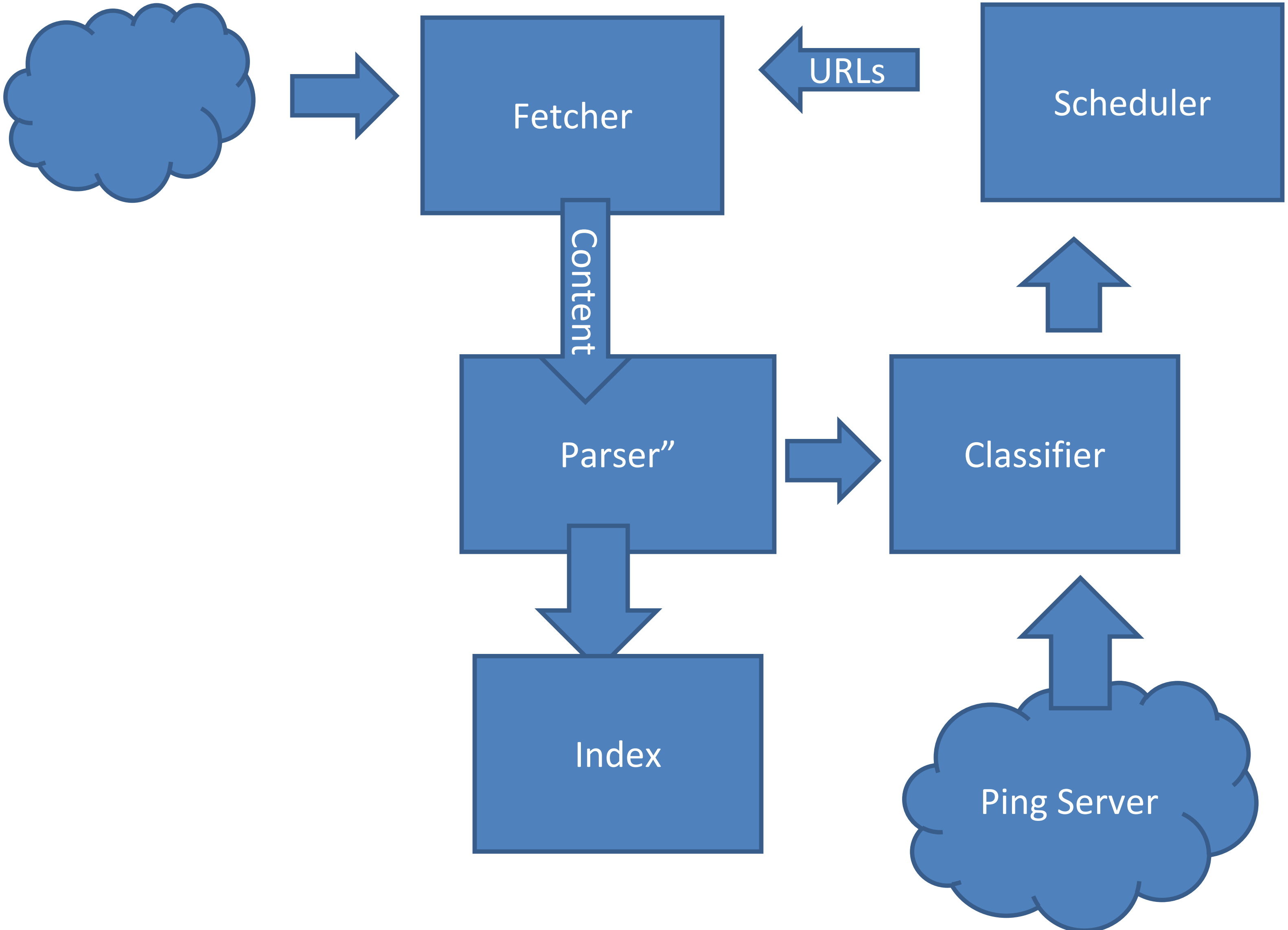
Content

Parser’

Classifier

Index

Blog Crawler (2)



Crawl Issues

- Politeness
  - Robots.txt
  - Exclusions
- Cost
  - Hardware
  - Traffic
- Spam

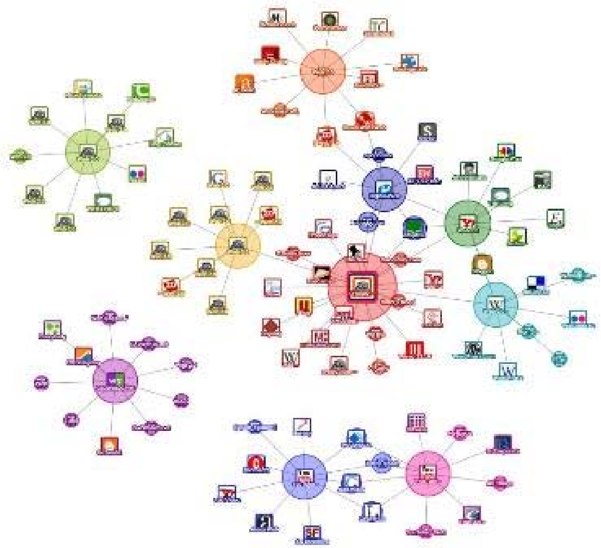
Bibliography

- A. Heydon and M. Najork, \Mercator: A Scalable,Extensible Web Crawler," *World Wide Web, vol. 2, no. 4,*pp. 219{229, Dec. 1999.
- H.-T. Lee, D. Leonard, X. Wang, and D. Loguinov, "IRLbot: Scaling to 6 Billion Pages and Beyond,'' WWW, April 2008 (best paper award).
- Ka Cheung Sia, Junghoo Cho, Hyun-Kyu Cho "Efficient Monitoring Algorithm for Fast News Alerts." IEEE Transactions on Knowledge and Data Engineering, 19(7): July 2007

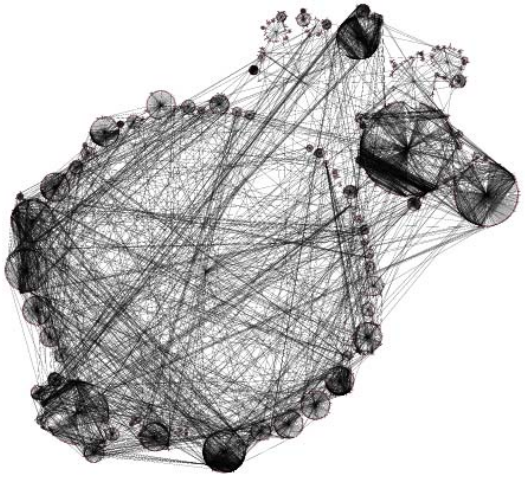
Outline

- Session 1: Overview, Applications and Architectures (for social media analysis)
- In-Depth 1: Data Acquisition
- Session 2: Methods
  - Graph Mining
  - Content Mining
- In-Depth 2: Data Prepartion

Social Media Graphs



Facebook graph, via Touchgraph



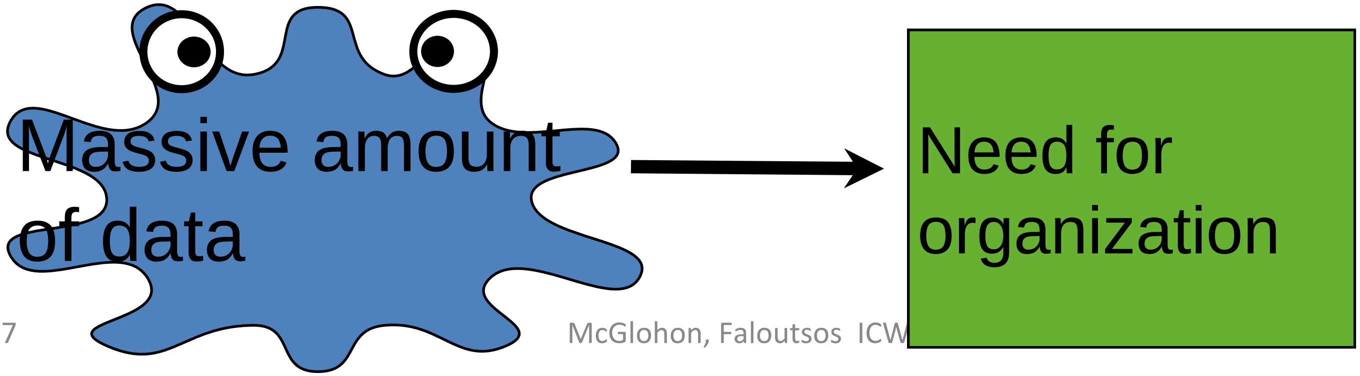
Livejournal, via Lehman and Kottler

Examples of Graph Mining

- Example: Social media host tries to look at certain online groups and predict whether the group will flourish or disband.
- Example: Phone provider looks at cell phone call records to determine whether an account is a result of identity theft.

Why graph mining?

- Thanks to the web and social media, for the first time we have easily accessible network data on a large-scale.
- Understand relationships (links) as well as content (text, images).
- Large amounts of data raise new questions.



Motivating questions

- Q1: How do networks form, evolve, collapse?
- Q2: What tools can we use to study networks?
- Q3: Who are the most influential/central members of a network?
- Q4: How do ideas diffuse through a network?
- Q5: How can we extract communities?
- Q6: What sort of anomaly detection can we perform on networks?

Outline

- Graph Theory
- Social Network Analysis/Social Networks Theory
- Social Media Analysis<-> SNA

Graph Theory

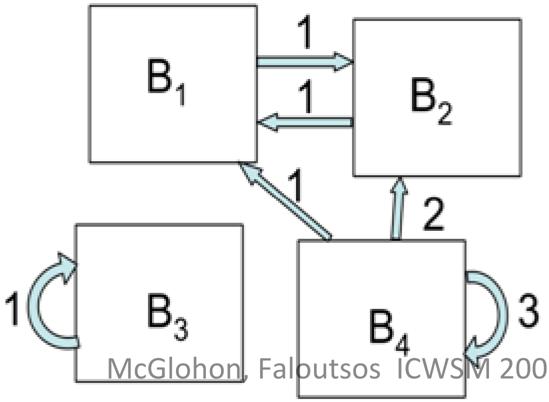
- Network
- Adjacency matrix
- Bipartite Graph
- Components
- Diameter
- Degree Distribution

Graph Theory (Ctd)

- BFS/DFS
- Dijkstra
- etc

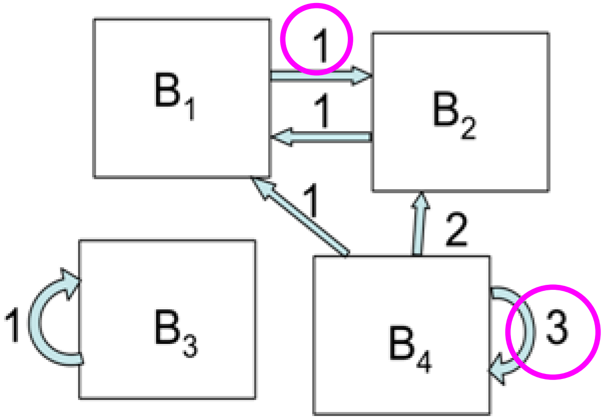
D1: Network

- A network is defined as a graph ***G=(V,E)***
  - – ***V*** : set of vertices, or nodes.
  - – ***E*** : set of edges.
- Edges may have numerical weights.



D2: Adjacency matrix

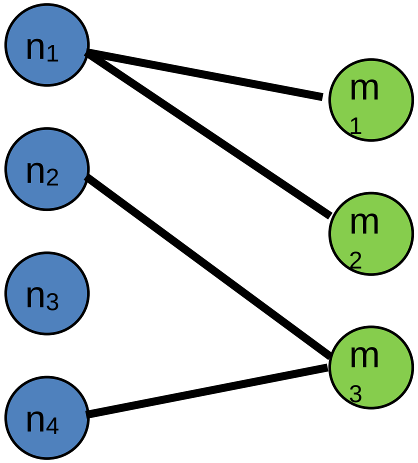
- To represent graphs, use adjacency matrix
- Unweighted graphs: all entries are 0 or 1
- Undirected graphs: matrix is symmetric



$$\begin{array}{cc|cccc} & & & \textit{to} & & \\ & & B_1 & B_2 & B_3 & B_4 \\ \hline & B_1 & 0 & 1 & 0 & 0 \\ \textit{from} & B_2 & 1 & 0 & 0 & 0 \\ & B_3 & 0 & 0 & 1 & 0 \\ & B_4 & 1 & 2 & 0 & 3 \end{array}$$

D3: Bipartite graphs

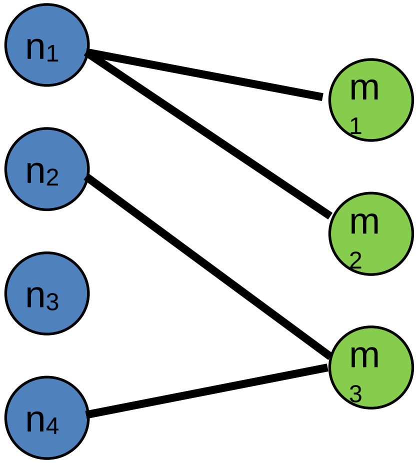
- In a bipartite graph,
  - 2 sets of vertices
  - edges occur between different sets.
- If graph is undirected, we can represent as a non-square adjacency matrix.



$$\begin{array}{c|ccc} & m_1 & m_2 & m_3 \\ \hline n_1 & 1 & 1 & 0 \\ n_2 & 0 & 0 & 1 \\ n_3 & 0 & 0 & 0 \\ n_4 & 0 & 0 & 1 \end{array}$$

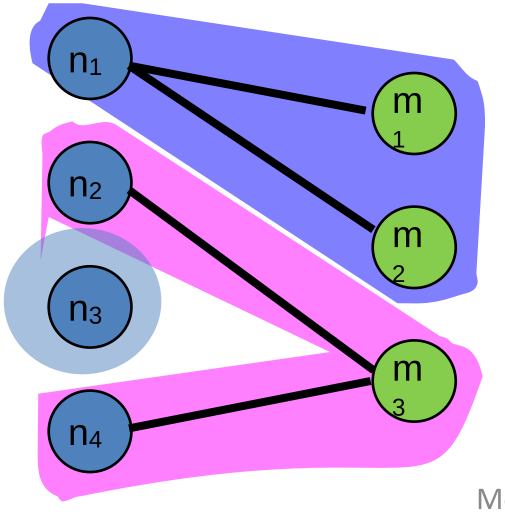
D4: Components

- Component: set of nodes with paths between each.



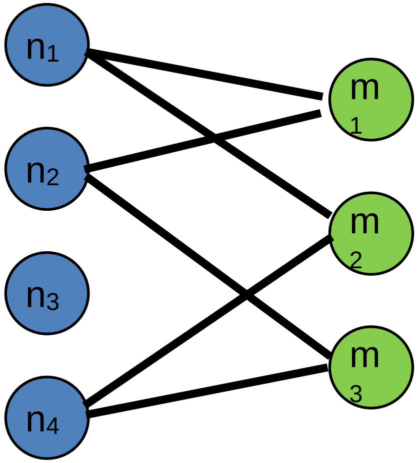
D4: Components

- Component: set of nodes with paths between each.
- We will see later that often real graphs form a giant connected component.



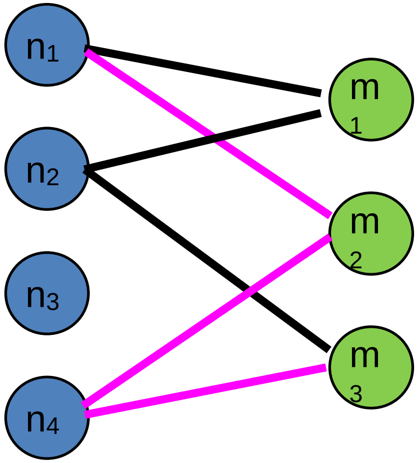
D5: Diameter

- Diameter of a graph is the “longest shortest path”.



D5: Diameter

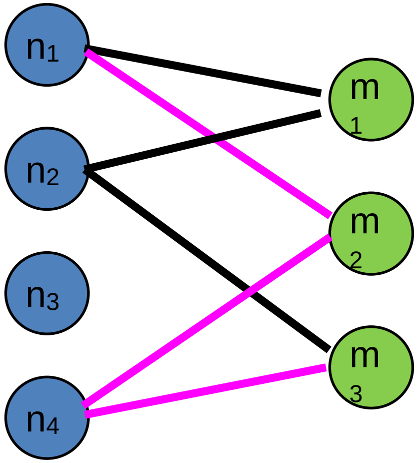
- Diameter of a graph is the “longest shortest path”.



diameter=3

D5: Diameter

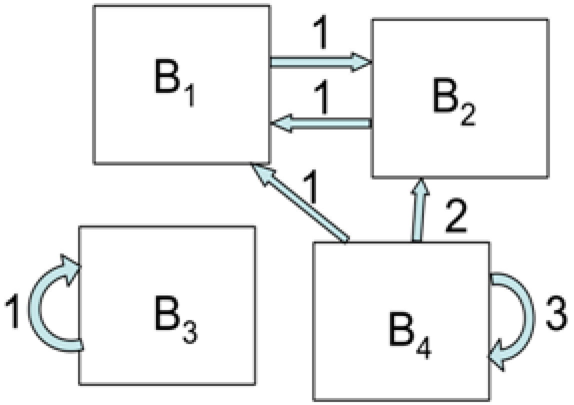
- Diameter of a graph is the “longest shortest path”.
- We can estimate this by sampling.
- Effective diameter is the distance at which 90% of nodes can be reached.



diameter=3

D6: Degree distribution

- We can find the degree of any node by summing entries in the (unweighted) adjacency matrix.



out-degree

$$\begin{array}{cc|cccc|c}
 & & & \textit{to} & & & \\
 & & B_1 & B_2 & B_3 & B_4 & \text{out-degree} \\ \hline
 & B_1 & 0 & 1 & 0 & 0 & 1 \\
\textit{from} & B_2 & 1 & 0 & 0 & 0 & 1 \\
 & B_3 & 0 & 0 & 1 & 0 & 1 \\
 & B_4 & 1 & 1 & 0 & 1 & 3 \\ \hline
\text{in-degree} & & 2 & 2 & 1 & 1 &
\end{array}$$

Graph Methods

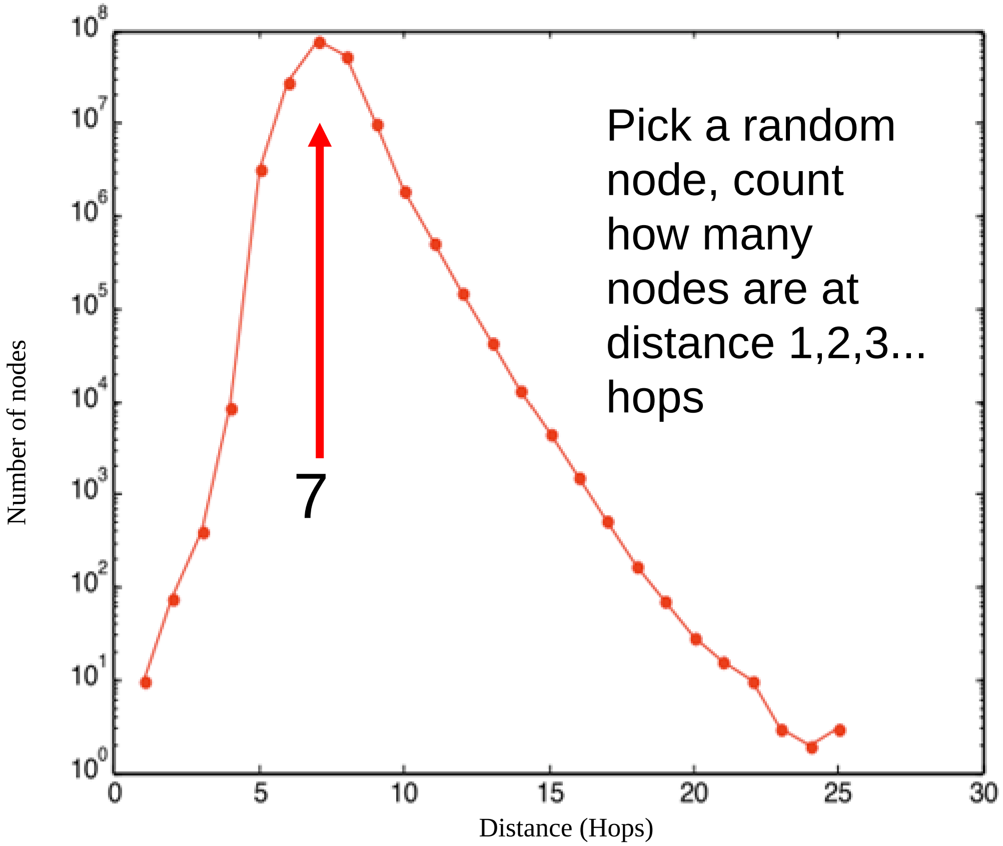
- SVD
- PCA
- HITS
- PageRank

Small World

- Stanley Milgram, 1967: six degrees of separation
- WEB: 18.59, Barabasi 1999
- Erdos number. AVG < 5

[Leskovec & Horvitz 07]

- Distribution of shortest path lengths
- Microsoft Messenger network
  - 180 million people
  - 1.3 billion edges
  - Edge if two people exchanged at least one message in one month period



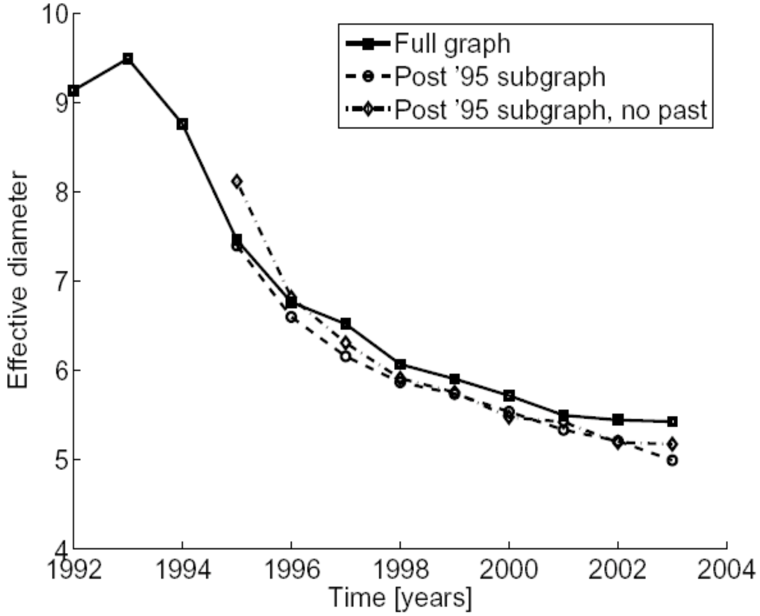
Pick a random node, count how many nodes are at distance 1,2,3... hops

Shrinking diameter

[Leskovec, Faloutsos, Kleinberg KDD 2005]

- Citations among physics papers
- 11yrs; @ 2003:
  - 29,555 papers
  - 352,807 citations
- For each month *M*, create a graph of all citations up to month *M*

diameter



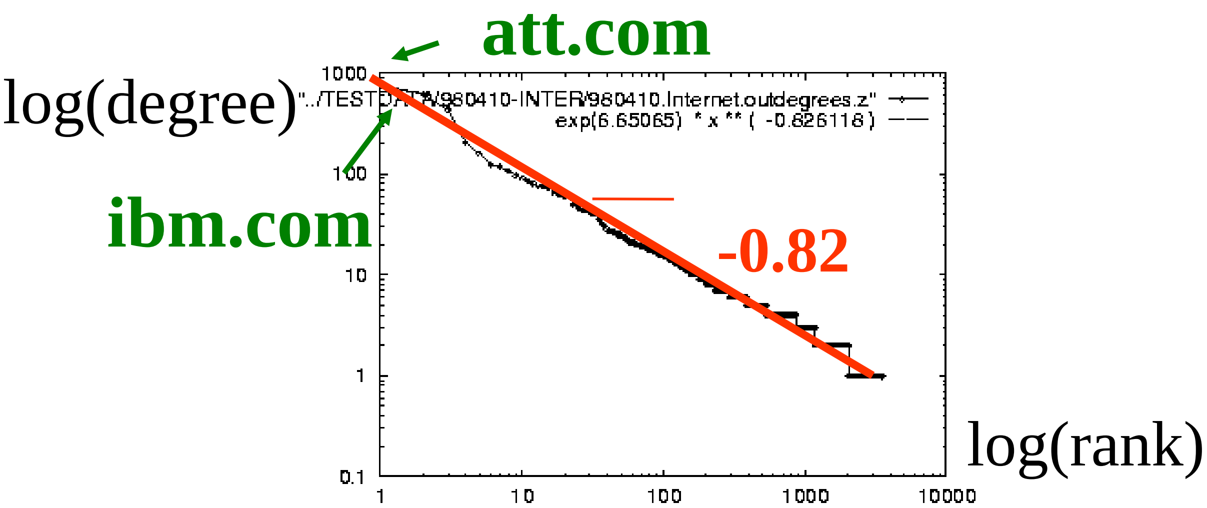
(a) arXiv citation graph

time

Power law degree distribution

- Measure with rank exponent R
- Faloutsos et al [SIGCOMM99]

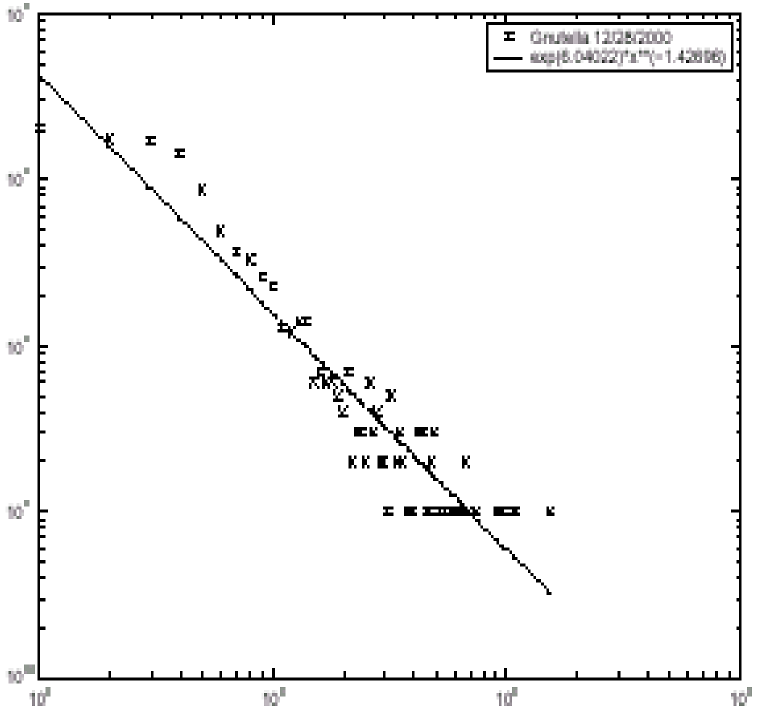
internet domains



The Peer-to-Peer Topology

count

[Jovanovic+]



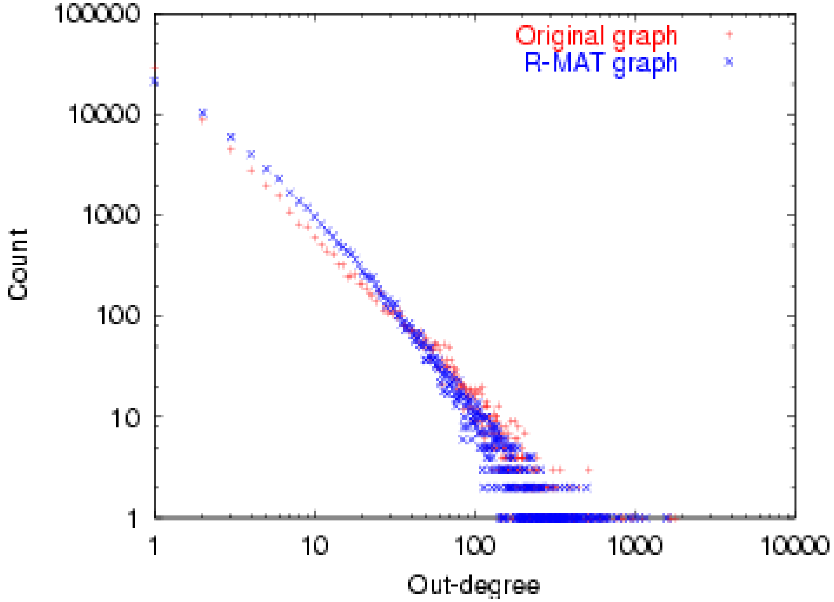
degree

(a) Gnutella snapshot from Dec. 28, 2000 (|r|=0.94)

- Number of immediate peers (= degree), follows a power-law

epinions.com

count

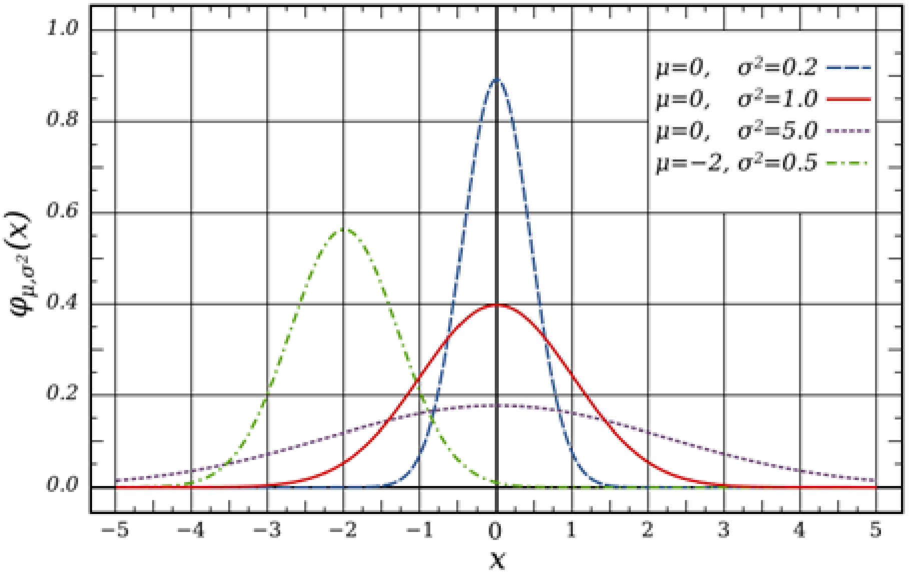


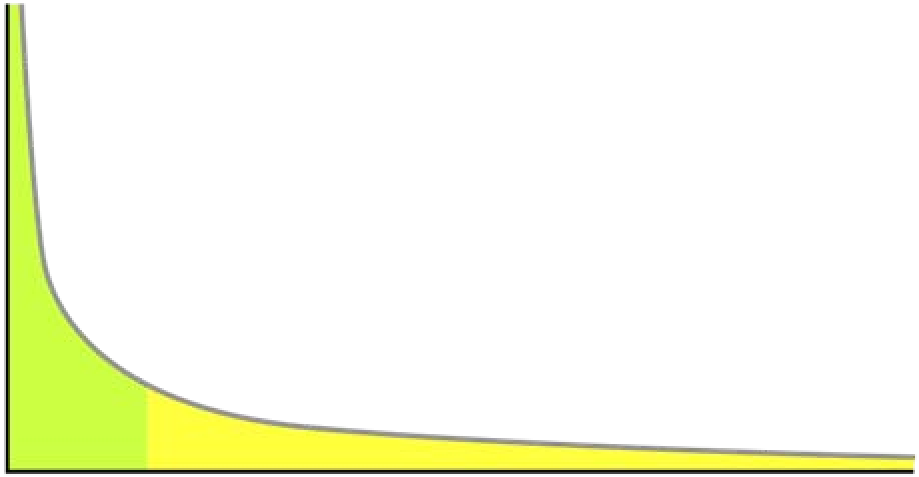
(out) degree

- who-trusts-whom [Richardson + Domingos, KDD 2001]

Power Law

- Normal vs Power





- Head and Tail

Preferential Attachment

- Albert-László Barabási ,Réka Albert: 1999
- Generative Model
- The probability of a node getting linked is proportional to a number of existing links
- Results in Power Law degree distribution
- Average Path length Log(|V|)

SNA/SNT

Well established field

Centrallity

- Degree
- Betweennes

SMA<->SNA

- Real World Networks
- Online Social Networks
  - Explicit
  - Implicit

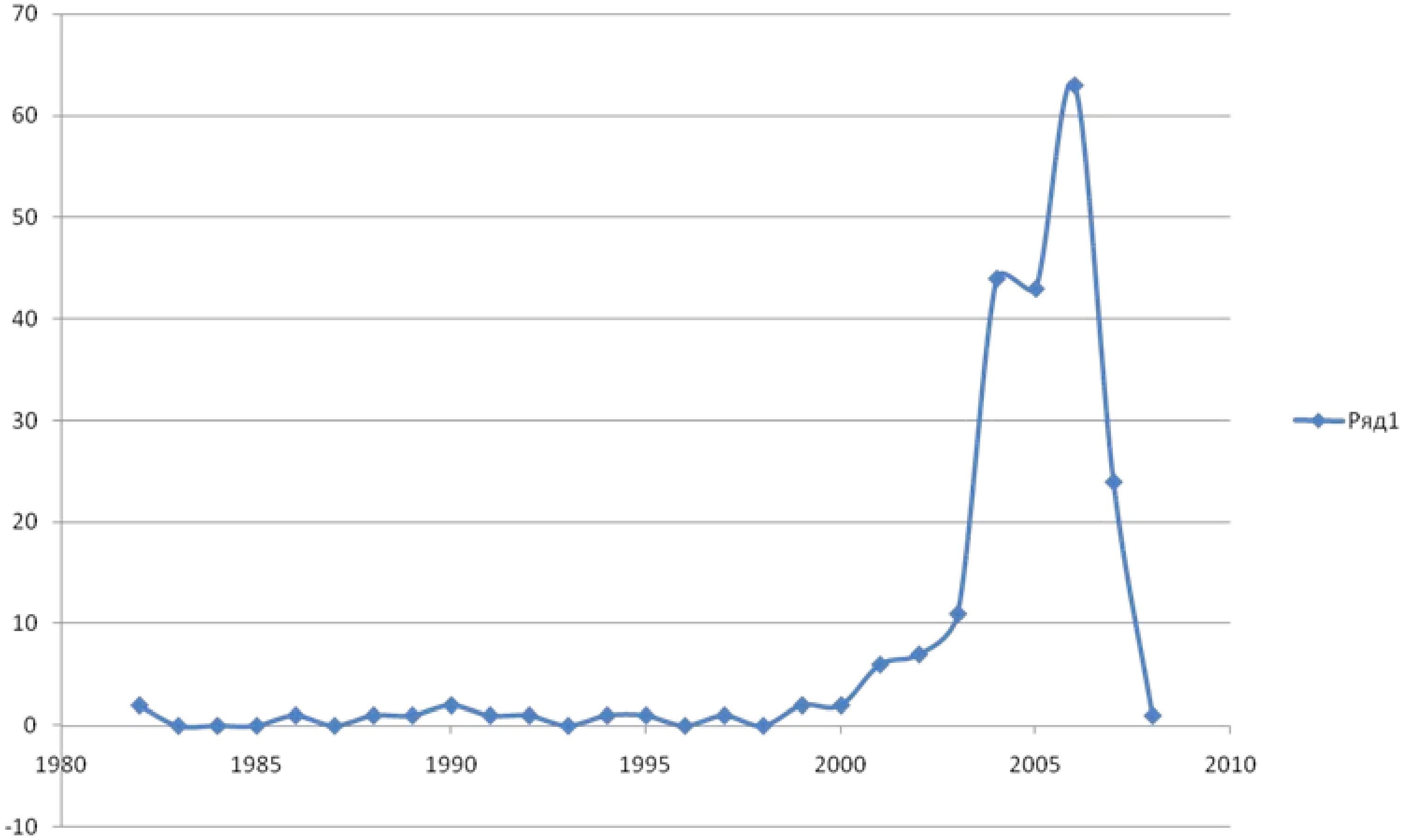
Outline

- Session 1: Overview, Applications and Architectures (for social media analysis)
- In-Depth 1: Data Acquisition
- Session 2: Methods
  - Graphs
  - Content (Subjectivity)
- In-Depth 2: Link Counting

Outline

- Overview
- Problem Statement
- Applications
- Methods
  - Sentiment classification
  - Lexicon generation
  - Target discovery and association

Subjectivity Research



Taxonomy of Subjectivity

Subjective Statement: <holder, <belief>, time>	The moon is made of green cheese..
Opinion: <holder, <prop, orientation>, time>	He should buy the Prius..
Sentiment: <holder, <target, orientation>, time>	I loved Raiders of the Lost Arc!.

Problem Statement(s)

- For a given document, determine if it is positive or negative
- For a given sentence, determine if it is positive or negative wrt some topic.
- For a given topic, determine if the aggregate sentiment is positive or negative.

Applications

- **Product review mining:** Based on what people write in their reviews, what features of the ThinkPad T43 do they like and which do they dislike?
- **Review classification:** Is a review positive or negative toward the movie?
- **Tracking sentiments toward topics over time:** Based on sentiments expressed in text, is anger ratcheting up or cooling down?
- **Prediction (election outcomes, market trends):** Based on opinions expressed in text, will Clinton or Obama win?
- ***Etcetera!***

Jan Wiebe, 2008

Problem Statement

- Scope:
  - Clause, Sentence, Document, Person
- Holder: who is the holder of the opinion?
- What is the thing about which the opinion is held?
- What is the direction of the opinion?
- Bonus: what is the intensity of the opinion?

Challenges

- Negation: I liked X; I didn’t like X.
- Attribution: I think you will like X. I heard you liked X.
- Lexicon/Sense: This is wicked!
- Discourse: John hated X. I liked *it*.
- Russian language is even more complex

Lexicon Discovery

- Lexical resources are often used in sentiment analysis, but how can we create a lexicon?
- Unsupervised Learning of semantic orientation from a Hundred Billion Word Corpus, Turney et al, 2002 (http://arxiv.org/ftp/cs/papers/0212/0212012.pdf)
- Learning subjective adjectives from Corpora, Wiebe, 2000
- Predicting the semantic orientation of adjectives, Hatzivassiloglou and McKeown, 1997, ACL-EACL (http://acl.ldc.upenn.edu/P/P97/P97-1023.pdf)
- Effects of adjective orientation and gradability on sentence subjectivity, Hatzivassiloglou et al, 2002

Using Mutual Information

- Intuition: if words are more likely to appear together than apart they are more likely to have the same semantic orientation.
- (Pointwise) Mutual information is an appropriate measure:

$$\log(\frac{p(x,y)}{p(x)\cdot p(y)})$$

SO-PMI

- Positive paradigm = good, nice, excellent, …
- Negative paradigm = bad, nasty, poor, …

$$\mathrm{PMI}(word_1, word_2) = \log_2\left(\frac{\mathrm{p}(word_1 \ \& \ word_2)}{\mathrm{p}(word_1)\ \mathrm{p}(word_2)}\right)$$

$$\begin{aligned}\text{SO-PMI-IR}(word) = {} & \mathrm{PMI}(word, \{positive \ \ paradigms\}) \\ & - \mathrm{PMI}(word, \{negative \ \ paradigms\})\end{aligned}$$

Graphical Approach

- Intuition: in expressions like ‘it was both *adj*$_1$ and *adj*$_2$’ the adjectives are more likely than not to have the same polarity (both positive or both negative).

Graphical Approach

- Approach 1: look at coordinations independently – 82% accuracy.
- Approach 2: build a complete graph (where nodes are adjectives and edges indicate coordination); then cluster – 90%.

DOCUMENT CLASSIFICATION

Pang, Lee, Vaithyanathan

- Thumbs up?: sentiment classification using machine learning techniques, ACL 2002
- Document level classification of movie reviews.
- Data from rec.arts.movies.reviews (via IMDB)
- Features: unigrams, bigrams, POS
- Conclusions: ML better than human, but sentiment harder than topic classification.

TARGET ASSOCIATION

Determining the Target

- Mining and summarizing customer reviews, KDD 2004, Hu & Liu (http://portal.acm.org/citation.cfm?id=1014073&dl=)
- Retrieving topical sentiments from an online document collection, SPIE 2004, Hurst & Nigam (http://www.kamalnigam.com/papers/polarity-DRR04.pdf)
- Towards a Robust Metric of Opinion, AAAI-SS 2004, Nigam & Hurst (http://www.kamalnigam.com/papers/metric-EAAT04.pdf)

Opinion mining – the **abstraction**

(Hu and Liu, KDD-04; Web Data Mining book 2007)

- Basic components of an opinion
  - Opinion holder: The person or organization that holds a specific opinion on a particular object.
  - Object: on which an opinion is expressed
  - Opinion: a view, attitude, or appraisal on an object from the opinion holder.
- Objectives of opinion mining: many ...
- **Let us abstract the problem**
- We use consumer reviews of products to develop the ideas.

Object/entity

- **Definition** (**object**): An object *O* is an entity which can be a product, person, event, organization, or topic. *O* is represented as
  - a hierarchy of components, sub-components, and so on.
  - Each node represents a component and is associated with a set of **attributes** of the component.
  - *O* is the root node (which also has a set of attributes)
- An opinion can be expressed on any node or attribute of the node.
- To simplify our discussion, we use "**features**" to represent both components and attributes.
  - The term "feature" should be understood in a ***broad sense***,
    - Product feature, topic or sub-topic, event or sub-event, etc
  - the object *O* itself is also a feature.

Model of a review

- An object *O* is represented with a finite set of features, *F* = $\{f_1, f_2, \ldots, f_n\}$.
  - Each feature f_i in *F* can be expressed with a finite set of words or phrases W_i, which are **synonyms**.

- **Model of a review**: An opinion holder *j* comments on a subset of the features $S_j \subseteq F$ of object *O*.
  - For each feature $f_k \in S_j$ that *j* comments on, he/she
    - chooses a word or phrase from W_k to describe the feature, and
    - expresses a positive, negative or neutral opinion on f_k.

Opinion mining tasks (contd)

- At the feature level:

  Task 1: Identify and extract object features that have been commented on by an opinion holder (e.g., a reviewer).

  Task 2: Determine whether the opinions on the features are positive, negative or neutral.

  Task 3: Group feature synonyms.

  – Produce a feature-based opinion summary of multiple reviews.

- Opinion holders: identify holders is also useful, e.g., in news articles, etc, but they are usually known in the user generated content, i.e., authors of the posts.

Feature-based opinion summary (Hu and Liu, KDD-04)

GREAT Camera., Jun 3, 2004

Reviewer: **jprice174** from Atlanta, Ga.

I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. …

….

Feature Based Summary:

Feature1: **picture**

Positive: 12

- The pictures coming out of this camera are amazing.
- Overall this is a good camera with a really good picture clarity.

…

Negative: 2

- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

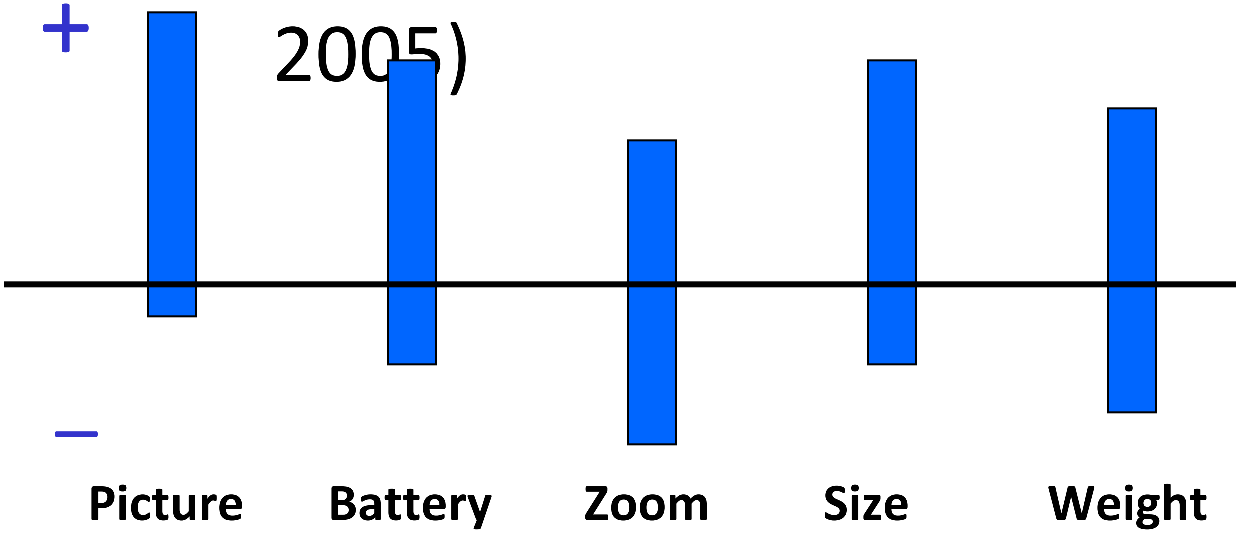
Feature2: **battery life**

…

Visual comparison (Liu et al, WWW-2005)

Summary of reviews of Digital camera

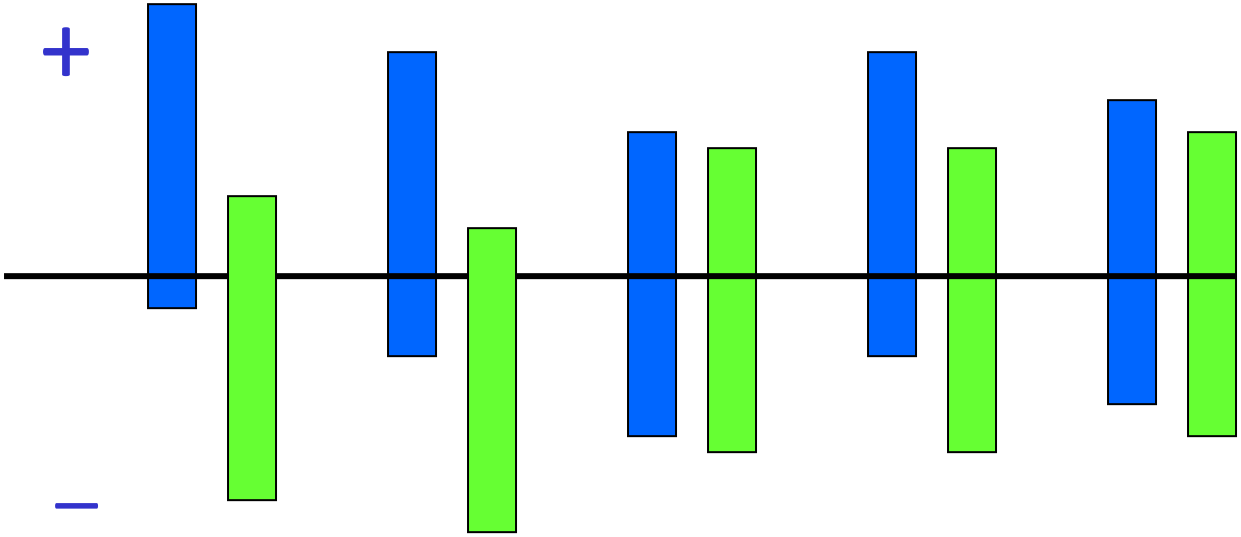
🟦 1



Comparison of reviews of

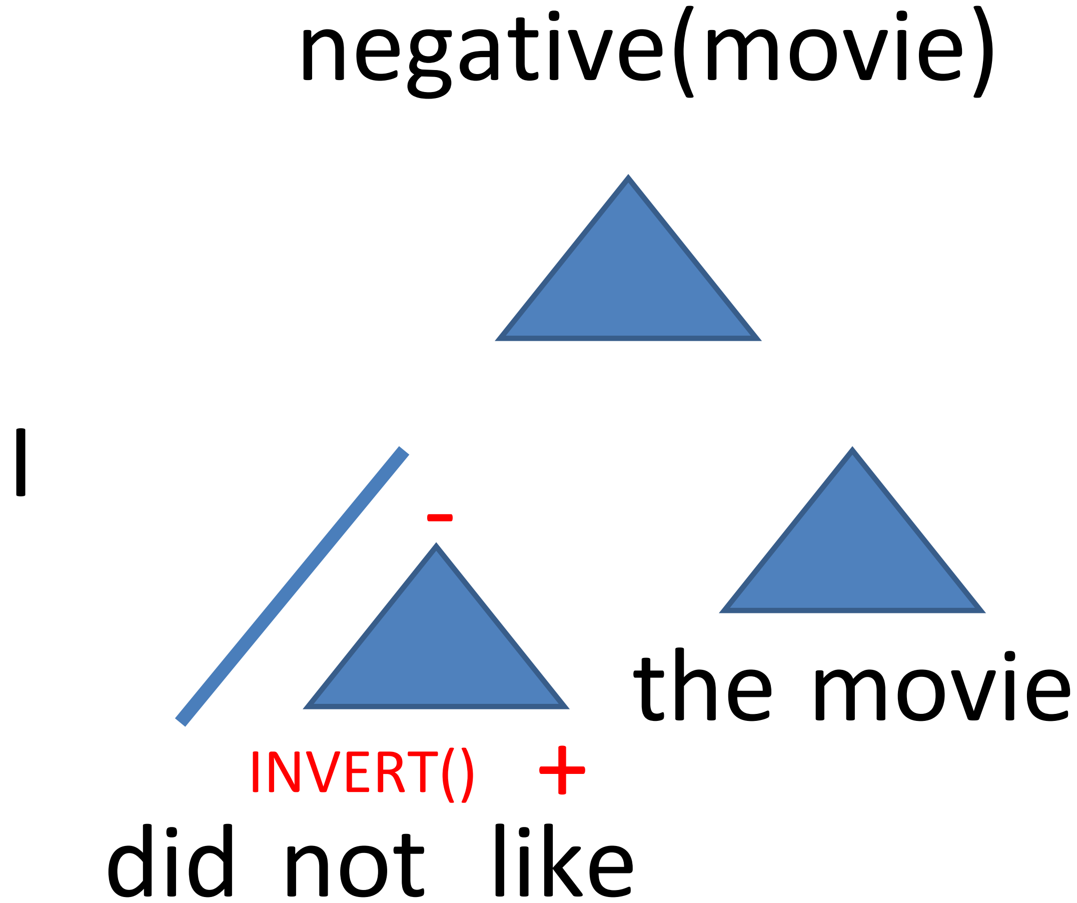
🟦 Digital camera 1

🟩 Digital camera 2



Grammatical Approach

- Hurst, Nigam
- Combine sentiment analysis and topical association using a compositional approach.
- Sentiment as a feature is propagated through a parse tree.
- The semantics of the sentence are composed.



negative(movie)

I

-

the movie

INVERT() +

did not like

Future Directions and Challenges

- Much current work is document focused, but opinions are held by the author, thus new methods should focus on the author.
- More robust methods for handling the informal language of social media.

Outline

- Session 1: Overview, Applications and Architectures (for social media analysis)
- In-Depth 1: Data Acquisition
- Session 2: Methods
  - Graphs
  - Content
- In-Depth 2: Data Preparation

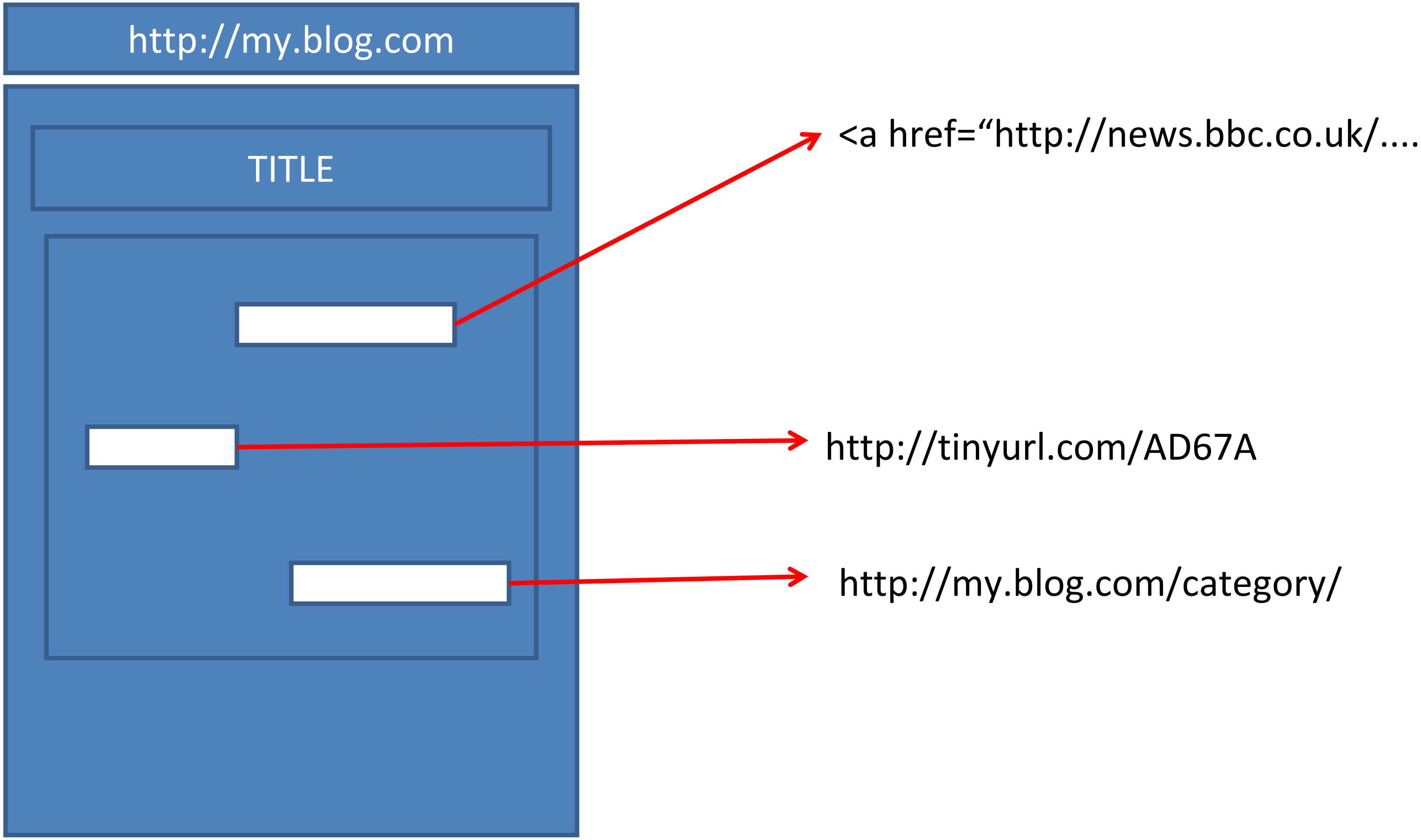
Task Description

- Count every links to a news article in a variety of social media content:
  - Weblogs
  - Usenet
  - Twitter
- Assume that you have a feed of this raw data.

Considerations

- How to extract links.
- Which links to count.
- How to count them.

Weblog Post Links



http://my.blog.com

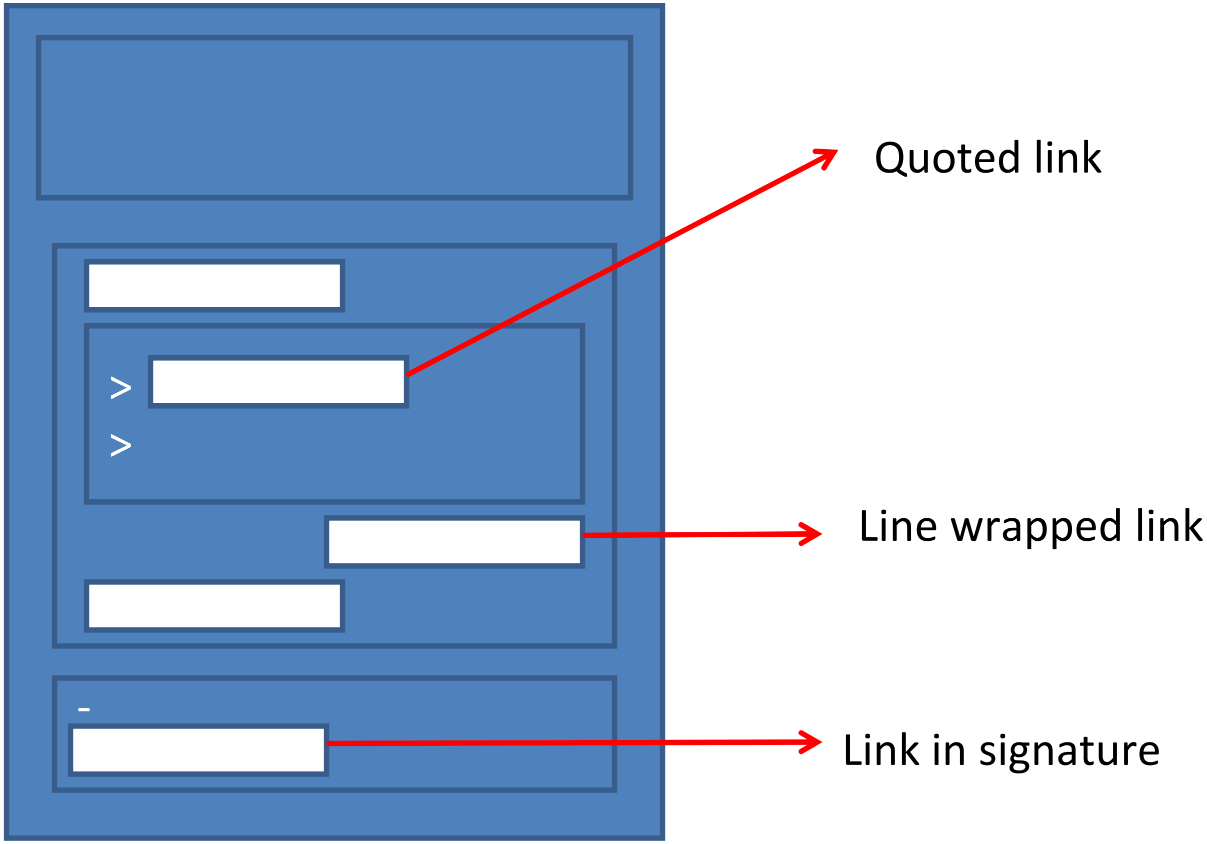
TITLE

<a href="http://news.bbc.co.uk/....

http://tinyurl.com/AD67A

http://my.blog.com/category/

Usenet Post Links



How To Extract Links

- Need to consider how links appear in each medium (in href args, in plain text, …)
- Need to consider cases where the medium can corrupt a link (e.g. forced line breaks in usenet)
- Need to follow some links (tinyurl, feedburner, …)

Which Links to Count (1)

- What is the task of counting links? E.g.: measure how much attention is being paid to what web object (news articles, …)
- Need to distinguish topical links, which are present to reference some topical page, object and links with other rhetorical purposes:
  - Self links (links to other posts in my blog)
  - Links in signatures of Usenet posts

Which Links To Count (2)

- We want to distinguish the type of links:
  - News
  - Weblog posts
  - Company home pages,
  - Etc.
- How can we do this?
  - Crawling and classification?
  - URL based classification?

How to Count

- Often the structure of the medium must be considered:
  - Do we count links in quoted text?
  - Do we count links in cross posted Usenet posts?
  - Do we count self links?

Summary

- All though text and data mining often rely on the law of large numbers, it is vital to get basic issues such as correct URL extraction, link classification, etc. figured out to prevent noise in the results.
- One should consider a methodology to counting (e.g. by modeling the manner in which the author structures their documents and communicates their intentions) so that a) the results can be tested and b) one has a clear picture of the goal of the task.

Research Areas

- Document analysis/parsing: recognizing different areas in a document such as text, quoted material, tables, lists, signatures.
- Link classification: without crawling the link predict some feature of the target based on the URL and context.
- Modeling the content creation process: a clear model is vital for creating and evaluating mining tasks in social media. What was the author trying to communicate?

Conclusion

Thanks

- Mary McGlohon
- Tim Finin
- Lada Adamic
- Bing Liu