Text Mining for Social Media

Dr. Madhu
NIT Hamirpur

Social Media

Social media are media for social interaction, using highly accessible and scalable communication techniques. It is the use of web-based and mobile technologies to turn communication into interactive dialogue.

Top Sites 2013

Rank	*Website*	*Rank*	*Website*
1	Google	6	Blogger
2	Facebook	7	Baidu
3	Youtube	8	Wikipedia
4	Yahoo!	9	Twitter
5	Windows Live	10	QQ.com

Types

Category	*Representative Sites*
Wiki	Wikipedia, Scholarpedia
Blogging	Blogger, LiveJournal, WordPress
Social News	Digg, Mixx, Slashdot
Micro Blogging	Twitter, Google Buzz
Opinion & Reviews	ePinions, Yelp
Question Answering	Yahoo! Answers, Baidu Zhidao
Media Sharing	Flickr ,Youtube
Social Bookmarking	Delicious, CiteULike
Social Networking	Facebook, LinkedIn, MySpace

Big Data and Small World

Small world Experiments – Stanley Miligram

6 degrees of separation

Small world phenomenon on web - Jure Leskovec

- 2008, largest social network of that time,
- The communication network of 240 million users of Microsoft Instant Messenger verified
- It is indeed a small world after all

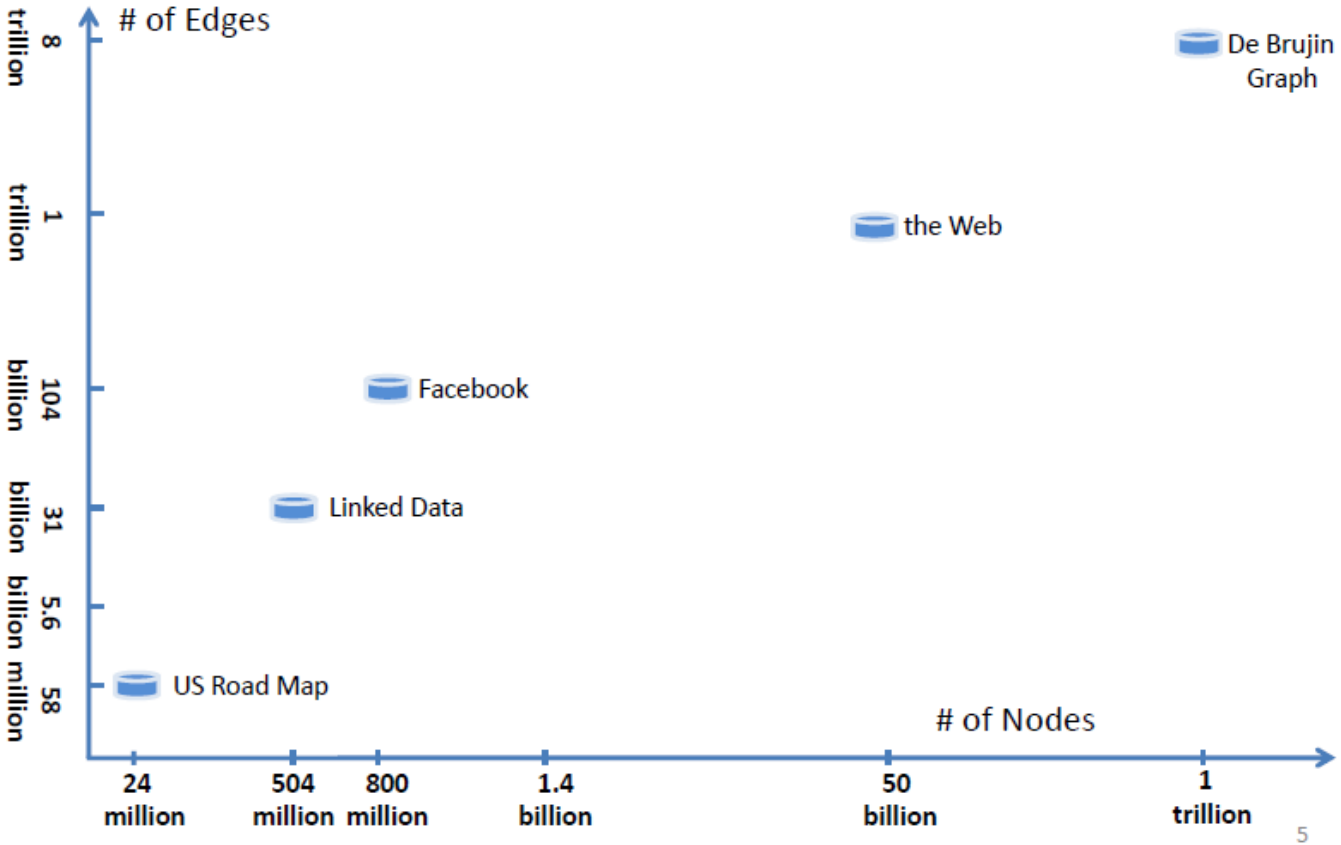
  average degree of separation is 6.6.

How Big Data IS?

2.5 Billion Pieces Of Content And 500+ Terabytes Ingested Every Day on Facebook

Graph Data Size

Graphs encode rich relationships



Why Social Network Analysis?

- Different perspectives
- Link analysis
- Crowd's wisdom
- Business Intelligence
- Online buzz

SNA's Findings

- Community Analysis
- Opinion and Sentiment Analysis
- Social Recommendation
- Influence Modelling
- Information Diffusion and Provenance
- Privacy, Security, and Trust

Structural Analysis

- Degree centrality
- Betweeness
- Closeness centrality
- Eigenvalue

Outline

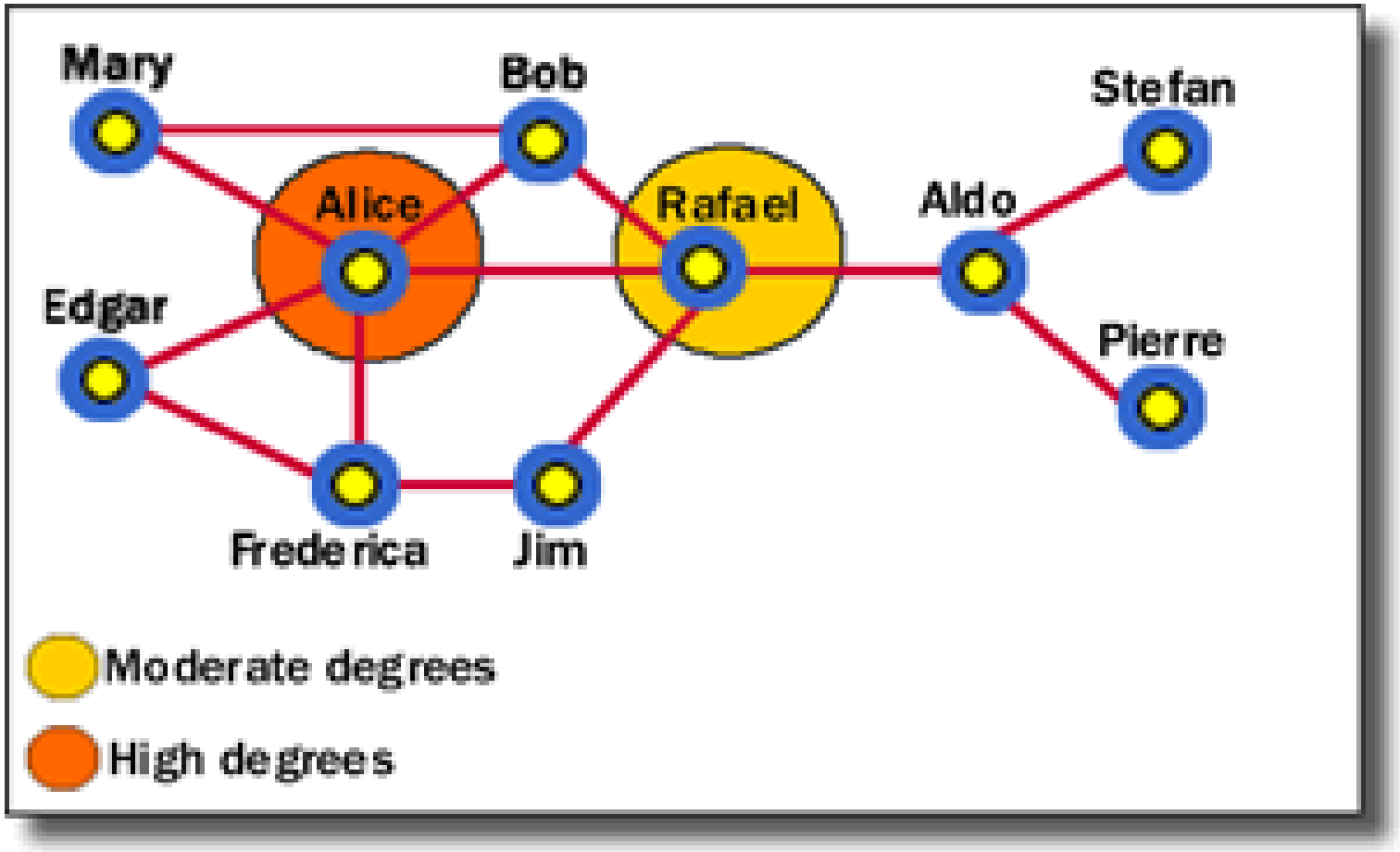
- Structural Analysis
- Classical SNA Tasks
- Data analytic view
- Text Mining + Structural SNA
- Challenges
- Some solutions
- Future

Centrality

- Central people have more influence in their network.
- tend to receive better performance reviews
- tend to be more satisfied with their jobs than people who are less central.

Measures of centrality: degree, betweenness and closeness centrality

Example


Mary
Bob
Stefan
Alice
Rafael
Aldo
Edgar
Pierre
Frederica
Jim
Moderate degrees
High degrees

Strength of Weak Ties

Cohesion : There are several measures of cohesion, including density. However, one common measure is the average number of ties it takes for a person in the group to "reach" another person in the group. If Adam is connected to Bill who is connected to Cindy, then Adam is at a distance of 2 from Cindy. The average distance for the group gives an indication of the group's cohesion.

Subgroup Identification: SNAs can identify the number of closely knit subgroups or "cliques" in a network. Within a clique, every unit is connected to every other unit. These subgroups can then be analysed to see if they share overlapping members. A network that contains highly segregated subgroups is not as well integrated as a network in which individuals belong to several overlapping subgroups.

SNA

- Content Analysis (High EQ)

  Images, Texts, Tags, Symbols

- Structural Analysis (High Relevance)

Social Media: Beyond Simple Sentiment

- **Analysis of Conversations-** Higher level context
  - Techniques: self-revelation, humor, sharing of secrets,
- **Establishment of informal agreements**, private language
  - Detect relationships among speakers and changes over time
  - Strength of social ties, informal hierarchies
- **Combination with other techniques**
  - Expertise Analysis – plus Influencers
  - Quality of communication (strength of social ties, extent of Private language, amount and nature of epistemic emotions -confusion.
- **Experiments - Pronoun Analysis – personality types**
  - Analysis of phrases, multiple contexts – conditionals, oblique

Social Media: Beyond Simple Sentiment

Expertise Analysis

- Experts think & write differently – process, chunks
- Categorization rules for documents, authors, communities

Applications:

Business & Customer intelligence, Voice of the Customer

Deeper understanding of communities, customers – better models – Security, threat detection – behavior prediction, Are they experts?

Expertise location- Generate automatic expertise characterization

Crowd Sourcing – technical support to Wiki's

Political – conservative and liberal minds/texts

- Disgust, shame, cooperation, openness

Good News

70-80 % is Text

Why Text Mining for SNA

- Structural Properties: High Relevance
- Text data Rich: High Opinion

SNA = Textual Information + Structural Properties

Semantic Gap

Text Mining For SNA (Major Tasks)

- Costumer Intelligence
- Clustering the Social Community
- Social Influence Models etc.
- Event Detection

Major Challenges

- Time Sensitivity
- Short Length
- Unstructured Phrases
- Abundant Information

More Problems

- Text in social media is not independent and identically distributed (i.i.d.) data anymore.
- Multi-dimensional social networks
- Network representation
- Dynamic networks.

Context information for effective similarity measure

- The first is the basic representation of texts called surface representation , which exploits phrases in the original text from different aspects to preserve the contextual information.
- Fail to perform a deep understanding of the original text.
- Correlation between Phrases

Unstructured Data

- Variance in the quality of the content
- First, the variance of quality originates from people's attitudes when posting a microblogging message or answering a question in a forum.
- The distribution of quality has high variance: from very high-quality items to low-quality, sometimes abusive content.

Further Addition to Chios

- New abbreviations in text
- Folksonomy
- How r u?", "Good 9t
- Previous text analytics sources always appear as ‹user, content›
- structure, while the text analytics in social media is able to derive data
- from various aspects, which include user, content, link, tag, time stamp etc.

Event Detection

- Social text streams
- First, a classifier is trained by using keywords, message length, and corresponding context as features to classify tweets into positive or negative cases.
- Second, they build a probabilistic spatiotemporal model for the target event to identify location of the event

Event Detection

"Breaking News"

- ranked higher and assigned in an important place, like the front page.
- Tracking the diffusion and evolution of a popular event in social media is another interesting direction in this field They tackle the problem of popular
- event tracking in online communities by studying the interplay between textual content and social networks.

"Breaking News"

- first analyse temporal and locational distributions of tag usage.
- Second, they identify tags related with events, and further distinguish if the tags are relevant to aperiodic events or periodic events.

Collaborative Q&A

Collaborative question answering portals are a popular destination for users looking for advice with a particular situation, for gathering opinions, for sharing technical knowledge, for entertainment, for community interaction, and for satisfying one's curiosity about a countless number of things.

What is social tagging?

- Tag photos on Flickr
- Tag URLs on Delicious
- Tag blog posts on Blogger, Wordpress, Livejournal
- Hash tags on twitter
- Annotations on social networks like Orkut, Facebook
- Comment and tag events on event sites
- Tagging books on LibraryThing
- Tagging citations, reviews, news, multimedia, answers …

Why taxonomies?

- Problems with Metadata Generation and Fixed Taxonomies
  - Manual, expensive, different vocabulory
  - fixed static taxonomies are rigid, conservative, and centralized
  - post activation analysis paralysis
- Folksonomies as a Solution
  - folksonomy (folk (people) + taxis (classification) + nomos (management))
  - emergent and iterative system

Tags: why and what?

- Different User Tagging Motivations
  - Future Retrieval (to read)
  - Contribution and Sharing
  - Attract Attention
  - Play and Competition
  - Self Presentation (mystuff, myLaptop)
  - Opinion Expression
  - Task Organization (jobsearch)
  - Social Signaling
  - Money
  - Technological Ease (Phonetags)
- Categorizers Versus Describers

Kinds of Tags

- Content-Based Tags (Autos, Honda, batman, Lucene)
- Context-Based Tags (location, time)
- Attribute Tags (Jeremy's Blog)
- Ownership Tags
- Subjective Tags (opinion, emotion)
- Organizational Tags (mywork, mypaper)
- Purpose Tags (learn_LATEX)
- Factual Tags (people, place, concepts)
- Personal Tags
- Self-referential tags (sometaithurts)
- Tag Bundles (tagging tags)

Broad Significance

Tag Recommendation

Effective Utilization of Tags

Real World Example
Bridging Semantic Gap

Wikipedia

Seed Phrase Extraction: shallow parsing

Semantic Features Generation

Feature Space Construction

Seed Extraction

$$WikiDice(t_i, t_j) = \begin{cases} 0 & \text{if } f(t_i \mid t_j) = 0 \\ & \text{or } f(t_j \mid t_i) = 0 \\ \frac{f(t_i|t_j)+f(t_j|t_i)}{f(t_i)+f(t_j)} & \text{otherwise} \end{cases}$$

$$WikiJaccard(t_i, t_j) = \frac{min(f(t_i \mid t_j), f(t_j \mid t_i))}{f(t_i) + f(t_j) - max(f(t_i \mid t_j), f(t_j \mid t_i))}$$

Seed Extraction

$$WikiOverlap(t_i, t_j) = \frac{min(f(t_i \mid t_j), f(t_j \mid t_i))}{min(f(t_i), f(t_j))}$$

$$WD_{ij} = \frac{WikiDice_{ij} - min(WikiDice_k)}{max(WikiDice_k) - min(WikiDice_k)}$$

Semantic Similarity

$$WikiSem(t_i, t_j) = (1 - \alpha - \beta)WD_{ij} + \alpha WJ_{ij} + \beta WO_{ij}$$

where α and β weight the importance of the three similarity measures.

Semantic Similarity

$$InfoScore(t_i) = \sum_{j=1, j \neq i}^{n} WikiSem(t_i, t_j).$$

$$t^* = \arg \max_{t_i \in \{t_1, t_2, \ldots, t_n\}} InfoScore(t_i)$$

Semantic feature Generation

- Background Knowledge Base
- Index of wiki corpus
- Ranking wiki query

Semantic feature Generation

input : a set S of *seed phrases*
output: *semantic features* SF

SF $\leftarrow$ *null*
for *seed phrase* $s \in S$ **do**
    if $s \in$ Sentence level **then**
        s.Query $\leftarrow$ SolrSyntax(s, OR)
    else
        s.Query $\leftarrow$ SolrSyntax(s, AND)
    WikiPages $\leftarrow$ Retrive(s.Query)
    $SF \leftarrow SF$ + Analyze(WikiPages)
return SF

Feature Space Filtering

Feature Filtering

- Remove features generated from too general *seed phrase* that returns a large number (more than 10,000) of articles from the index corpus.
- Transform features used for Wikipedia management or administration, e.g. "List of hotels"→"hotels", "List of twins"→"twins".
- Apply phrase sense stemming using Porter stemmer , e.g. "fictional books"→"fiction book".
- Remove features related to chronology, e.g. "year", "decade" and "centuries".

Feature Selection

- First, the *tf-idf* weights of all generated features are calculated.
- One *seed phrase* $s_i (0 < i \leq m)$ may generate k *semantic features*, denoted by $\{f_{i1}, f_{i2}, \ldots, f_{ik}\}$.
- S*emantic* diversity of features

$$f_i^* = \arg \max_{f_{ij} \in \{f_{i1}, f_{i2}, \ldots, f_{ik}\}} tf\_idf(f_{ij})$$

Data

http://www.lsi.upc.edu/~nlp/wikicorpus/

http://www.casos.cs.cmu.edu/projects/automap/

Future

Business Intelligence

Social Capital

Models for social footprinting

Social Tagging + folksonomy for Semantic Gap

References

http://www.alexa.com/topsites

http://www.casos.cs.cmu.edu/projects/automap/

Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.

Mining Text Data (Springer) Ed. Charu Aggarwal, ChengXiang Zhai, March 2012.

Social Network Data Analytics (Springer) Ed. Charu Aggarwal, March 2011.

World is Still Big

Home Work

1. Compute all Centrality measures for a randomly generated graph of 50 nodes.
2. Label(tag ramdomly) the Nodes of the above social Graph(add some more info) with 4 tags and try community detection.
3. Try semantic gap bridging for "Text Mining for Machine Intelligence" Phrase from wikicorpus.

Thank You