Data Mining MTAT.03.183

Text Mining
Social Network Analysis
Graph Mining

Jaak Vilo
2009 Fall

Topics

• Information Retrieval
• Text Mining
• Web Mining
• Social Network Analysis
  – friends, epidemiology, co-authoring, co-citation, espionage, ...
• Graph Mining
• STACC
Books

• Feldman, Sanger: The Text Mining Handbook

• Information Retrieval
• Web Mining

Materials

• Modern Information Retrieval by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.
  – http://people.ischool.berkeley.edu/~hearsirbook/
  – http://www.amazon.co.uk/Modern-Information-Retrieval-AKAA
  – http://books.google.com/books?id=AAAAACAAJ&dq=modern+information+retrieval
  – New edition in May 2009 (overdue?)

• Google Books: Information Retrieval
  – http://books.google.com/books?q=information+retrieval

• ESSCaSS’08 : Ricardo Baeza-Yates and Filippo Menczer
  – http://courses.cs.ut.ee/schools/esscass2008/Main/Materials
Topics

• **Information Retrieval**
  • Text Mining
  • Web Mining
  • Social Network Analysis
    – friends, epidemiology, co-authoring, co-citation, espionage, ...
• **Graph Mining**

Concepts

*Information Retrieval* - the study of systems for representing, indexing (organising), searching (retrieving), and recalling (delivering) data.

*Information Filtering* - given a large amount of data, return the data that the user wants to see

*Information Need* - what the user really wants to know; a query is an approximation to the information need.

*Query* - a string of words that characterizes the information that the user seeks

*Browsing* - a sequence of user interaction tasks that characterizes the information that the user seeks
Documents

- News, articles
- Laws, legal documents
- Scientific publications, patents
- E-mail
- Technical documents
- Books
- Encyclopediae
- Dictionaries
- ...

Information Retrieval

- DB of indexed documents
- Query
- Find documents relevant to query
The classic search model

- TASK
- Info Need
- Verbal form
- Mis-conception
- Info about removing mice without killing them
- Mis-translation
- How do I trap mice alive?
- Mis-formulation

Search engine

- Polysomy Synonymy

More features

- Metadata
- Context
- Hypertext – xrefs
- Language
- Structured vs unstructured
- Semantics
- Tags
- ...

An introduction to Web Mining, WWW2009, Beijing
Classic IR Goal

- Classic relevance
  - For each query Q and stored document D in a given corpus assume there exists relevance score \( \text{Score}(Q, D) \)
    - Score is average over users U and contexts C
  - Optimize \( \text{Score}(Q, D) \) as opposed to \( \text{Score}(Q, D, U, C) \)
  - That is, usually:
    - Context ignored
    - Individuals ignored
    - Corpus predetermined

The Notion of Relevance

- Data retrieval: semantics tied to syntax
- Information retrieval: ambiguous semantics
- Relevance:
  - Depends on the user
  - Depends on the context (task, time, etc)
  - Corollary: The Perfect IR System
does not exist
Evaluation: First Quality, next Efficiency

TP
FP
FN
TN
T=True
F=False
P=Positives
N=Negative

Jaak Vilo and other authors UT: Data Mining 2009 13

User Needs

- **Need (Broder 2002)**
  - **Informational** – want to learn about something (~40% / 65%)
    - Low hemoglobin
  - **Navigational** – want to go to that page (~25% / 15%)
    - United Airlines
  - **Transactional** – want to do something (web-mediated) (~35% / 20%)
    - Edinburgh weather
    - Male surface images
    - Canon S410
  - Gray areas
    - Find a good hub
    - Explore a search icon, what’s there?

Jaak Vilo and other authors UT: Data Mining 2009 14
Using the Context

Example: I want information about Santiago

- **Context**
  - Family in Chile
  - Catholic
  - Travelling to Cuba
  - Lives in Argentina
  - Located in Santo Domingo
  - Architect
  - Spanish movies fan
  - Baseball fan

- **Probable Answer**
  - Santiago de Chile
  - Santiago de Compostela
  - Santiago de Cuba
  - Santiago del Estero
  - Santiago de los Caballeros
  - Santiago Calatrava
  - Santiago Segura
  - Santiago Benito

---

SEARCH GOAL | DESCRIPTION | EXAMPLES
---|---|---
1. Navigational | My goal is to go to specific known website that I already have in mind. The only reason I’m searching is that it’s more convenient than typing the URL, or perhaps I don’t know the URL. | aloha airlines 
www.govhospital 
Home page

2. Informational | My goal is to learn something by reading or viewing web pages
2.1 Direct | I want to get an answer to a question that has a single, unambiguous answer. | what is a supercharger 2004 election dates
2.1.2 Open | I want to get an answer to an open-ended question, or one with uncontrolled depth. | why are metals shiny
2.2 Undirected | I want to learn anything/everything about my topic. A query for topic X might be interpreted as “tell me about X.” | help quitting smoking
2.3 Advice | I want to get advice, ideas, suggestions, or instructions. | waiting with weights
2.4 Locate | My goal is to find out whether or where some real world service or product can be obtained | yellow pages
2.5 List | My goal is to get a list of plausible suggested web sites (i.e. the search result list itself), each of which might be candidates for helping me achieve some underspecified, unspecified goal. | travel
3. Manipulation | My goal is to obtain a resource (not information) available on web pages. | Page with resources
3.1 Download | My goal is to download a resource that must be on my computer or other device to be useful | www.parole.com
3.2 Entertainment | My goal is to be entertained simply by viewing items available on the result page | www.airline.com
3.3 Interact | My goal is to interact with a resource using another program/service available on the web site I find | wysiwyg editor
3.4 Obtain | My goal is to obtain a resource that does not require a component to use. I may print it out, but I can also just look at it on the screen. I’m not obtaining it to learn some information, but because I want to use the resource itself. | review document

Rose & Levinson, 2004

---

An Introduction to Web Mining, WWW2008, Beijing
Web search

- data mining university of tartu fall 2009
- Java
- Paris Hilton
- climate change
Web search engines

- Rooted in Information Retrieval (IR) systems
  - Prepare a keyword index for corpus
  - Respond to keyword queries with a ranked list of documents.

- ARCHIE
  - Earliest application of rudimentary IR systems to the Internet
  - Title search across sites serving files over FTP
Document representation

- Feature vector
  - Bag of words
    - vector $D_i$ – which words are present
  - Query $Q$
  - similarity $M(Q, D_i) = Q \cdot D_i$ (scalar product)

- Weighted version – easy to generalise

TF-IDF

- Term Frequency – Inverted Document Frequency
- Weights: frequent in doc, infrequent in DB
  - Measures frequency, and relevance
  - $TF-IDF(w, d) = tf(w, d) \cdot idf(w)$

$$TF-IDF\_Weight(w, d) = Term\_Frequency(w, d) \cdot \log(N/DocFreq(w))$$

- $\log( N / 1+DocFreq(w) )$ to avoid division by 0 if not in the database at all...
Boolean queries: Examples

• Simple queries involving relationships between terms and documents
  – Java (documents containing word Java)
  – Java AND (NOT coffee) Java -coffee
• Proximity queries
  – phrase “Java beans” or the term API
  – Documents where Java and island occur in the same sentence

Document preprocessing

• Tokenization
  – Filtering away tags
  – Tokens regarded as nonempty sequence of characters excluding spaces and punctuations.
  – Token represented by a suitable integer, \( tid \), typically 32 bits
  – Optional: stemming/conflation of words
  – Result: document (did) transformed into a sequence of integers \( (tid, pos) \)
Storing tokens

• Straight-forward implementation using a relational database
  – Example figure
  – Space scales to almost 10 times
• Accesses to table show common pattern
  – reduce the storage by mapping tids to a lexicographically sorted buffer of \((did, pos)\) tuples.
  – Indexing = transposing document-term matrix

Two variants of the inverted index data structure, usually stored on disk. The simpler version in the middle does not store term offset information; the version to the right stores term offsets. The mapping from terms to documents and positions (written as "document/position") may be implemented using a B-tree or a hash-table.
Storage

• For dynamic corpora
  – Berkeley DB2 storage manager
  – Can frequently add, modify and delete documents

• For static collections
  – Index compression techniques (to be discussed)

Stopwords

• Function words and connectives
• Appear in large number of documents and little use in pinpointing documents

• Indexing stopwords
  – Stopwords not indexed
    • For reducing index space and improving performance
  – Replace stopwords with a placeholder (to remember the offset)

• Issues
  – Queries containing only stopwords ruled out
  – Polysemous words that are stopwords in one sense but not in others
    • E.g.; can as a verb vs. can as a noun
Stemming

- Conflating words to help match a query term with a morphological variant in the corpus.
- Remove inflections that convey parts of speech, tense and number
- E.g.: university and universal both stem to universe.
- Techniques
  - morphological analysis (e.g., Porter's algorithm)
  - dictionary lookup (e.g., WordNet).
- Stemming may increase recall but at the price of precision
  - Abbreviations, polysemy and names coined in the technical and commercial sectors
  - E.g.: Stemming "ides" to "IDE", “SOCKS” to "sock", “gated” to “gate”, may be bad!

Batch indexing and updates

- Incremental indexing
  - Time-consuming due to random disk IO
  - High level of disk block fragmentation
- Simple sort-merges.
  - To replace the indexed update of variable-length postings
- For a dynamic collection
  - single document-level change may need to update hundreds to thousands of records.
  - Solution: create an additional “stop-press” index.
Relevance ranking

- Keyword queries
  - In natural language
  - Not precise, unlike SQL
    - Boolean decision for response unacceptable
  - Solution
    - Rate each document for how likely it is to satisfy the user's information need
    - Sort in decreasing order of the score
    - Present results in a ranked list.
- No algorithmic way of ensuring that the ranking strategy always favors the information need
  - Query: only a part of the user's information need
Responding to queries

- Set-valued response
  - Response set may be very large
    - (E.g., by recent estimates, over 12 million Web pages contain the word java.)
- Demanding selective query from user
- Guessing user's information need and ranking responses
- Evaluating rankings

<table>
<thead>
<tr>
<th>$k$</th>
<th>$r_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

Precision and interpolated precision plotted against recall for the given relevance vector. Missing $r_k$ are zeroes.
Beyond keyword search

Bayesian Inferencing

Bayesian inference network for relevance ranking. A document is relevant to the extent that setting its corresponding belief node to true lets us assign a high degree of belief in the node corresponding to the query.

Manual specification of mappings between terms to approximate concepts.
Bayesian Inferencing (contd.)

• Four layers
  1. Document layer
  2. Representation layer
  3. Query concept layer
  4. Query

• Each node is associated with a random Boolean variable, reflecting belief

• Directed arcs signify that the belief of a node is a function of the belief of its immediate parents (and so on..)

NLP – Natural Language Processing

• Tokenisation – words, sentences
• Part-of-Speech (POS) tagging
  – articles, nouns, verbs, adjective, number, … (87+?)
• Syntactical Parsing
• Shallow Parsing
Text Categorisation

- Indexing texts with controlled vocabularies
- Document sorting, text filtering
  - topic
  - spam
- Hierarchical web page categorisation
  - Yahoo!, etc

• Single label vs multilabel
  - overlapping
  - binary
  - Multilabel can be represented by binary classifiers
• Document or category – pivoted
• Hard vs Soft
Bridging the gap: **raw data** vs **actionable information**

- Document Management Systems
- Technical documentation
- Email
- CRM
- News Stories
- Web pages

**Tagging**

- Search
- Personalisation
- Analysis
- Alerting
- Decision Support

**Machine Readable** **Machine Understandable**

---

**Topics**

- Information Retrieval
- **Text Mining**
- Web Mining
- Social Network Analysis
  - friends, epidemiology, co-authoring, co-citation, espionage, ...
- Graph Mining
Information Extraction (IE)

• Information extraction tasks...

Typical subtasks of IE are

• **Content Noise Removal**
  – remove noise contents. For example, *tagclouds*, *navigational menu*, *related contents*, and context related advertisements.

• **Named Entity Recognition** (NER)
  – recognition of entity names (for people and organizations), place names, temporal expressions, and certain types of numerical expressions.

• **Coreference resolution**:
  – detection of *coreference* and *anaphoric* links between text entities. In IE tasks, this is typically restricted in finding links between previously extracted named entities. For example, "International Business Machines" and "IBM" refer to the same real world entity.

• **Terminology extraction**: finding the relevant terms for a given *corpus*

• **Relationship Extraction**: identification of relations between entities, such as:
  – PERSON works for ORGANIZATION (extracted from the sentence "Bill works for IBM.")
  – PERSON located in LOCATION (extracted from the sentence "Bill is in France.")
Example from Feldman

Intelligent Auto-Tagging


By Stephen J. Hedges and Cam Simpson

The Finsbury Park Mosque is the center of radical Muslim activism in England. Through its doors have passed at least three of the men now held on suspicion of terrorist activity in France, England and Belgium, as well as one Algerian man in prison in the United States.

The mosque’s chief cleric, Abu Hamza al-Masri, lost two hands fighting the Soviet Union in Afghanistan and he advocates the elimination of Western influence from Muslim countries. He was arrested in London in 1999 for his alleged involvement in a Yemen bomb plot and, like the other Yemeni leaders, he was granted special residence in Saudi Arabia before being extradited.

"..."
## IE Accuracy by Information Type

<table>
<thead>
<tr>
<th>Information Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>90-98%</td>
</tr>
<tr>
<td>Attributes</td>
<td>80%</td>
</tr>
<tr>
<td>Facts</td>
<td>60-70%</td>
</tr>
<tr>
<td>Events</td>
<td>50-60%</td>
</tr>
</tbody>
</table>
Ontologies

- Names, synonyms, objects...

- Controlled vocabularies

- Ontologies – definitions and relationships
Topics

• Information Retrieval
• Text Mining
• **Web Mining**
• Social Network Analysis
  – friends, epidemiology, co-authoring, co-citation, espionage, ...
• Graph Mining

History of Hypertext

• Citation,
  – Hyperlinking
• **Ramayana, Mahabharata, Talmud**
  – branching, non-linear discourse, nested commentary,
• **Dictionary, encyclopedia**
  – self-contained networks of textual nodes
  – joined by referential links
Hypertext systems

• Memex [Vannevar Bush, 1945]
  – stands for “memory extension”
  – photoelectrical-mechanical storage and computing device
  – Aim: to create and help follow hyperlinks across documents

• Hypertext
  – Coined by Ted Nelson
  – Xanadu hypertext (1981): system with
    • robust two-way hyperlinks, version management, controversy management, annotation and copyright management.

• Hypertext Mining

  • http://www.robotwisdom.com/web/timeline.html
  • HyperTerrorist's Timeline of Hypertext History
  • Jorn Barger . . . . . . . . . . . . . . . . . . . . . . . . . 4 March 1996
World-wide Web

- Initiated at CERN (the European Organization for Nuclear Research)
  - By Tim Berners-Lee
- GUIs
  - Berners-Lee (1990)
  - Erwise and Viola (1992), Midas (1993)
- Mosaic (1993)
  - a hypertext GUI for the X-window system
  - HTML: markup language for rendering hypertext
  - HTTP: hypertext transport protocol for sending HTML and other data over the Internet
  - CERN HTTPD: server of hypertext documents

World Wide Web

- Hypertext documents
  - Text
  - Links
- Web
  - billions of documents
  - authored by millions of diverse people
  - edited by no one in particular
  - distributed over millions of computers, connected by variety of media

Chakrabarti
http://www.cse.iitb.ac.in/~soumen/mining-the-web/
The early days of the Web: CERN HTTP traffic grows by 1000 between 1991-1994 (image courtesy W3C)

The early days of the Web: The number of servers grows from a few hundred to a million between 1991 and 1997 (image courtesy Nielsen)
1994: the landmark year

- Foundation of the “Mosaic Communications Corporation"
- first World-wide Web conference
- MIT and CERN agreed to set up the World-wide Web Consortium (W3C).

Web: A populist, participatory medium

- number of writers = (approx) number of readers.
- the evolution of MEMES
  - ideas, theories etc that spread from person to person by imitation.
  - Now they have constructed the Internet !!
  - E.g.: “Free speech online”, chain letters, and email viruses
Abundance and authority crisis

• liberal and informal culture of content generation and dissemination.
• Very little uniform civil code.
• redundancy and non-standard form and content.
• millions of qualifying pages for most broad queries
  – Example: java or kayaking
• no authoritative information about the reliability of a site

Problems due to Uniform accessibility

• little support for adapting to the background of specific users.
• commercial interests routinely influence the operation of Web search
  – “Search Engine Optimization“ !!
Hypertext data

• Semi-structured or unstructured
  – No schema
• Large number of attributes

![What is in the Web?](image)

- Information
- Porn

- + On-line casinos + Free movies + Cheap software
  + Buy a MBA diploma + Prescription - free drugs +
  VI-4-gra + Get rich now now now!!!
What is in the Web?

Current challenges (1)

- **Scraper spam**
  - Copies good content from other sites, adds monetization
    (most often Google AdSense)
  - Hard to identify at the page level (indistinguishable from
    original source), monetization not reliable clue (there is
    actually good content on the web that uses AdSense/YPN!)

- **Synthetic text**
  - Boilerplate text, randomized, built around key phrases
  - Avoids duplicate detection

- **Query-targeted spam**
  - Each page targets a single tail query (anchortext, title, body,
    URL). Often in large auto-constructed hosts, host-level
    analysis most helpful

- **DNS spam**
Current challenges (2)

- Blog spam
  - Continued trend toward blog “ownership” rather than comment spam
  - Orthogonal to other categories (scrapers, synthesizers). Just a hosting technique, plus exploiting blog interest
- Example:
  - 68,000 blogspot.com hosts all generated by the same spammer
    - 1) nursingschoolresources.blogspot.com
    - 2) transplantresources.blogspot.com
    - ... 67,798) beachesresourcesforyou.blogspot.com
    - 67,799) startrekresourcesforyou.blogspot.com

Community Dynamics

Next generation products will blur distinctions between Creators, Synthesizers, and Consumers

Example: Launchcast
Every act of consumption is an explicit act of production that requires no incremental effort.
Listening itself implicitly creates a radio station...
Crawling and indexing

• Purpose of crawling and indexing
  – quick fetching of large number of Web pages into a local repository
  – indexing based on keywords
  – Ordering responses to maximize user’s chances of the first few responses satisfying his information need.

• Earliest search engine: Lycos (Jan 1994)
• Followed by….
  – Alta Vista (1995), HotBot and Inktomi, Excite

Topic directories

• Yahoo! directory
  – to locate useful Web sites
• WWW-Wärk (www.ee), Neti (www.neti.ee)
• Efforts for organizing knowledge into ontologies
  – Centralized: (Yahoo!)
  – Decentralized: About.COM and the Open Directory
Clustering and classification

• Clustering
  – discover groups in the set of documents such that documents within a group are more similar than documents across groups.
  – Subjective disagreements due to
    • different similarity measures
    • Large feature sets

• Classification
  – For assisting human efforts in maintaining taxonomies
  – E.g.: IBM's Lotus Notes text processing system & Universal Database text extenders

Hyperlink analysis

• Take advantage of the structure of the Web graph.
  – Indicators of prestige of a page (E.g. citations)
  – HITS & PageRank

• Bibliometry
  – bibliographic citation graph of academic papers

• Topic distillation
  – Adapting to idioms of Web authorship and linking styles
Resource discovery and vertical portals

• Federations of crawling and search services
  – each specializing in specific topical areas.
• Goal-driven Web resource discovery
  – language analysis does not scale to billions of documents
  – counter by throwing more hardware

Structured vs. Web data mining

• Traditional data mining
  – data is structured and relational
  – well-defined tables, columns, rows, keys, and constraints.
• Web data
  – readily available data rich in features and patterns
  – spontaneous formation and evolution of
    • topic-induced graph clusters
    • hyperlink-induced communities
• Goal of book: discovering patterns which are spontaneously driven by semantics,
Crawl “all” Web pages?

- Problem: no catalog of all accessible URLs on the Web.
- Solution:
  - start from a given set of URLs
  - Progressively fetch and scan them for new outlinking URLs
  - fetch these pages in turn…..
  - Submit the text in page to a text indexing system
  - and so on………..

Crawling procedure

- Simple
  - Great deal of engineering goes into industry-strength crawlers
  - Industry crawlers crawl a substantial fraction of the Web
  - E.g.: Alta Vista, Northern Lights, Inktomi
- No guarantee that all accessible Web pages will be located in this fashion
- Crawler may never halt ……
  - pages will be added continually even as it is running.
Crawling overheads

• Delays involved in
  – Resolving the host name in the URL to an IP address using DNS
  – Connecting a socket to the server and sending the request
  – Receiving the requested page in response
• Solution: Overlap the above delays by
  – fetching many pages at the same time

Anatomy of a crawler.

• Page fetching threads
  – Starts with DNS resolution
  – Finishes when the entire page has been fetched
• Each page
  – stored in compressed form to disk/tape
  – scanned for outlinks
• Work pool of outlinks
  – maintain network utilization without overloading it
    • Dealt with by load manager
• Continue till the crawler has collected a sufficient number of pages.
Large-scale crawlers: performance and reliability considerations

- Need to fetch many pages at the same time
  - Utilize network bandwidth
  - Single page fetch may involve several seconds of network latency

- Highly concurrent and parallelized DNS lookups

- Use of asynchronous sockets
  - Explicit encoding of the state of a fetch context in a data structure
  - Polling socket to check for completion of network transfers
  - Multi-processing or multi-threading: impractical

- Care in URL extraction
  - Eliminating duplicates to reduce redundant fetches
  - Avoiding “spider traps”
Computing Similarity

- Features:
  - Segments of a document (natural or artificial breakpoints) [Brin95]
  - Shingles (Word N-Grams) [Brin95, Brod98]
    
    \[
    \begin{align*}
    a \text{ rose is a rose is a rose} & \Rightarrow \\
    & a \text{ _rose_is_a} \\
    & \text{rose_is_a_rose} \\
    & \text{is_a_rose_is} \\
    \end{align*}
    \]
    
    are all added in the bag of word representation

- Similarity Measure
  - TFIDF [Shiv95]
  - Set intersection [Brod98]
    (Specifically, Size_of_Intersection / Size_of_Union )

An Introduction to Web Mining, WWW2008, Beijing

Content mining

- Duplicate and near-duplicate document detection
- Content-based spam detection
What does WWW look like?
Fig. 7. The bowtie structure of the Web: Plot (a) shows the 4 parts: IN, OUT, SCC, and TENDRILS [Broder et al. 2000]. Plot (b) shows recursive bowties: subgraphs of the WWW can each be considered a bowtie. All these smaller bowties are connected by the navigational backbone of the main SCC of the Web [Dill et al. 2001].
Graph $G=(V,E)$

- $V$ – vertices (nodes)
- $E$ – edges (arcs, connections)
Find a shortest path from station A to station B. 
-need serious thinking to get a correct algorithm.

Representation For An Undirected Graph

(a) (b) (c)

Figure 22.1 Two representations of an undirected graph. (a) An undirected graph $G$ having five vertices and seven edges. (b) An adjacency-list representation of $G$. (c) The adjacency-matrix representation of $G$. 
Co-citation

- Measure of similarity between nodes
- If nodes $v$ and $w$ are both linked by node $u$, then they are co-cited
- If $E$ is the adjacency matrix of the graph, the number of nodes that co-cite both $v$ and $w$ is
  
  $$p[u] = \sum_u E[u,v] E[u,w] = \sum_u E^T[v,u] E[u,w] = (E^T E)[v,w]$$

- Thus similarity is captured in the entries of matrix $E^T E$

An Introduction to Web Mining, WWW2008, Beijing
PageRank

- [Brin and Page, 1998]
- Algorithm suggested for ranking results in web search
- An authority score is assigned to each Web page
- Authority scores independent of the query

- Authority scores corresponds to the stationary distribution of a random walk on the graph:
  - With probability $\alpha$ follow a link in the graph
  - With probability $1-\alpha$ go to a node chosen uniformly at random (teleportation)

- Random walk also known as random surfer model

An introduction to Web Mining, WWW2006, Beijing

PageRank

- Let $E$ be the adjacency matrix of the graph, and $L$ the row-stochastic version of $E$
- Each row of $E$ is normalized so that it sums to 1
- Authority score defined by
  $$ P_{(n+1)} = L^T p_{(n)} $$
- problematic if the graph is not strongly connected, So:
  $$ P_{(n+1)} = \alpha L^T p_{(n)} + (1-\alpha) \frac{1}{n} 1 $$
- where $1$ is the matrix with all entries equal to 1
- and $\alpha \in [0,1]$, common value $\alpha = 0.85$

An introduction to Web Mining, WWW2006, Beijing
PageRank variants and enhancements

- Personalized PageRank
  - Teleportation to a set of pages defining the preferences of a particular user
- Topic-sensitive PageRank [Haveliwala 02]
  - Teleportation to a set of pages defining a particular topic
- TrustRank [Gyöngyí 04]
  - Teleportation to “trustworthy” pages

- Many papers on analyzing PageRank and numerical methods for efficient computation

HITS

- [Kleinberg 1998]
- Exploit the intuition that there are:
  - pages that contain high-quality information (authorities)
  - pages with good navigational properties (hubs)

Good hubs point to good authorities and good authorities are pointed by good hubs
HITS algorithm

- Given a query $q$
- Use a standard web IR system to find a set of pages $R$ relevant to $q$ (root set)
- Expand to the set of pages connected to $R$ (expanded set) and form the graph $G=(V,E)$
- $a$ authority vector: $a[u]$ the authority score of node $u$
- $h$ hub vector: $h[u]$ the hub score of node $u$
  \[
  a = E^T h \\
  h = E a
  \]
- $a$ converges to the principal eigenvector of $E^T E$
- $h$ converges to the principal eigenvector of $E E^T$

An introduction to Web Mining, WWW2006, Beijing

HITS

- HITS is related to SVD on the graph matrix $E$
- non-principal eigenvectors provide different topics
- HITS sensitive to local-topology
- PageRank is more stable – due to random jump step
- Researchers attempted to make HITS more stable
  - SALSA stochastic algorithm for link analysis [Lempel and Moran, 01]:
  - A random surfer model in which the surfer follows alternatively random inlinks and outlinks
  - [Ng et al. 01] introduce a random jump step in the HITS model

An introduction to Web Mining, WWW2006, Beijing
Discussion

- HITS introduces the notion of hub, which does not exist in PageRank
- HITS is query sensitive
- PageRank does not depend on the query; thus the authority scores can be pre-computed
- Nepotism, two-host nepotism, and clique attacks

Clustering coefficient

\[ C_1 = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}} \]

- How to compute it?
- How to compute the number of triangles in a graph?
- Assume that the graph is very large, stored on disk
• How to compute the diameter of a graph?
• Matrix multiplication in $O(n^{2.376})$ time, but $O(n^2)$ space
• BFS from a vertex takes $O(n + m)$ time,
• but need to do it from every vertex, so $O(mn)$
• Resort to approximations again

Topics

• Information Retrieval
• Text Mining
• Web Mining
• Social Network Analysis
  – friends, epidemiology, co-authoring, co-citation, espionage, ...
• Graph Mining
Social network of 9/11 hijackers

the drawing. Many of the “higher level” aesthetic criteria are implicit consequences of the
- minimized number of edge crossings,
- evenly distributed edge length,
- evenly distributed vertex positions on the graph area,
- sufficiently large angular resolution between edges.
Running Example

Hijackers by Flight

<table>
<thead>
<tr>
<th>Flight 77: Pentagon</th>
<th>Flight 11: WTC 1</th>
<th>Flight 175: WTC 2</th>
<th>Flight 93: PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khalid Al-Midhar</td>
<td>Satam Al Suqami</td>
<td>Marwan Al-Shehhi</td>
<td>Saeed Alghamdi</td>
</tr>
<tr>
<td>Majed Moqed</td>
<td>Waleed M. Alshehri</td>
<td>Fayez Ahmed</td>
<td>Ahmed Alhaznawi</td>
</tr>
<tr>
<td>Nawaq Alhamzi</td>
<td>Wail Alshehri</td>
<td>Ahmed Alghamdi</td>
<td>Ahmed Alnami</td>
</tr>
<tr>
<td>Salem Alhamzi</td>
<td>Mohamed Atta</td>
<td>Hamza Alghamdi</td>
<td>Ziad Jarrahi</td>
</tr>
<tr>
<td>Hani Hanjour</td>
<td>Abdulaziz Alomari</td>
<td>Mohald Alshehri</td>
<td></td>
</tr>
</tbody>
</table>
a spring that pulls the vertices together), whereas distinct vertices are pushed apart by some constraint to help prevent them from being drawn at the same point. The method seeks equilibrium of these contradicting constraints. The first such algorithm was introduced by Eades (Eades 1984). Following Eades, two additional layout algorithms were introduced by Kamada and Kawai (KK) (Kamada and Kawai 1989) and Fruchterman and Reingold (FR) (Fruchterman and Reingold 1991).

Kamada and Kawai’s (KK) Method

Utilizing Hooke’s law, Kamada and Kawai modeled a graph as a system of springs. Every two vertices are connected by a spring whose rest length is proportional to the graph-theoretic distance between its two endpoints. Each spring’s stiffness is inversely proportional to the square of its rest length. The optimization algorithm

Figure XI.3. KK layout of the hijackers’ graph.

XI.5 Partitioning of Networks

Figure XI.7. Summary diagram of centrality measures (solid arrows point to highest value; dashed arrows point to second largest (done using Netminer (Cyrann 2004)).
Figure XI.B. Core partitioning of the hijackers’ graph.

XI.5.1 Cores

ABSTRACT

We present a study of anonymized data capturing a month of high-level communication activities within the whole of the Microsoft Messenger instant-messaging system. We examine characteristics and patterns that emerge from the collective dynamics of large numbers of people, rather than the actions and characteristics of individuals. The dataset contains summary properties of 30 billion conversations among 240 million people. From the data, we construct a communication graph with 180 million nodes and 1.3 billion undirected edges, creating the largest social network constructed and analyzed to date. We report on multiple aspects of the dataset and synthesized graph. We find that the graph is well-connected and robust to node removal. We investigate on a planetary-scale the oft-cited report that people are separated by “six degrees of separation” and find that the average path length among Messenger users is 6.6. We also find that people tend to communicate more with each other when they have similar age, language, and location, and that cross-gender conversations are both more frequent and of longer duration than conversations with the same gender.

Categories and Subject Descriptors: I.2.8 Database Management: Database applications – Data mining

General Terms: Measurement; Experimentation.

Keywords: Social networks; Communication networks; User demographics; Large data; Online communication.

Figure 4: World and Messenger user population age pyramid. Ages 15–30 are overrepresented in the Messenger population.
Figure 2: (a) Distribution of the number of people participating in a conversation. (b) Distribution of the durations of conversations. The spread of durations can be described by a power-law distribution.

Figure 3: (a) Distribution of login duration. (b) Duration of times when people are not logged into the system (times between logout and login).

Figure 6: Communication characteristics of users by reported age. We plot age vs. age and the color (z-axis) represents the intensity of communication.
Figure 7: Number of users at a particular geographic location. Color of data represents the number of users.

Figure 8: Number of Messenger users per capita. Color intensity corresponds to the number of users per capita in the cell of the grid.

Figure 9: A communication heat map.

Figure 10: (a) Communication among countries with at least 10 million conversations in June 2006. (b) Countries by average length of the conversation. Edge widths correspond to logarithms of intensity of links.
Smaller scale questions

http://yury.name/webguide/02webguide.pdf
Possible questions in SNA

• Prestige
• Centrality
• Co-citation/co-occurrence
• Radius/diameter
• ...
Prestige

• \( p \) = vector of prestige values
• \( E \) - citations
• \( E[i,j] \) – document \( i \) cites document \( j \)

• Calculate a new prestige where prestige of citing documents is added to current prestige
Prestige

- \( p' = E^T p \)

\[
p'[v] = \sum_{u} E^{T}[v,u] p[u] = E^T p
\]

- Find fixpoint, starting with \( p=(1,...n)^T \)
- Ensure \( p \) is normalised at each step, \( ||p||=1 \)
- \( p \) converges to the principal eigenvector of \( ET \)
Centrality

- **Importance notion** based on centrality
- removing a central node disconnects the graph to large components

- $d(u,v)$ the shortest-path distance between $u$ and $v$
- $r(u) = \max_v d(u,v)$ - radius of node $u$
- $\arg \min_u r(u)$ - center of the graph

- Various other notions of centrality in the literature
Topics

• Information Retrieval
• Text Mining
• Web Mining
• Social Network Analysis
  – friends, epidemiology, co-authoring, co-citation, espionage, ...

• Graph Mining
Finding the modules

Public datasets for H. sapiens
- IntAct: Protein interactions (PPI), 18773 interactions
- IntAct: PPI via orthologs from IntAct, 6705 interactions
- MEM: gene expression similarity over 89 tumor datasets, 46286 interactions
- Transfac: gene regulation data, 5183 interactions
Module evaluation

MCL clustering algorithm

- Markov (Chain Monte Carlo) Clustering
  - [http://www.micans.org/mcl/](http://www.micans.org/mcl/)

- Random walks according to edge weights

- Follow the different paths according to their probability

- Regions that are traversed “often” form clusters
http://www.micans.org/mcl/intro.html

With this, the MCL algorithm can be written as

G is a graph
add loops to G  # see below
set \( P \) to some value  # affects granularity
set \( M_1 \) to be the matrix of random walks on \( G \)

while (change) {
    \( M_2 = M_1 \times M_1 \)  # expansion
    \( M_1 = \Gamma(M_2) \)  # inflation
    change = difference (M_1, M_2)
}

set CLUSTERING as the components of \( M_1 \)  # see below
Approximation and compression = learning

Less is More: Compact Matrix Decomposition for Large Sparse Graphs

Jimeng Sun  Yinqian Xie  Hu Zhang  Christos Faloutsos
Carnegie Mellon University  jimeng,yxie,zhang,Christos}@cs.cmu.edu

Abstract
Given a large sparse graph, how can we find patterns and anomalies? Several important applications can be modeled as large sparse graphs, e.g., network traffic monitoring, research citation network analysis, social network analysis, and regulatory networks in genes. Low rank decompositions, such as SVD and CUR, are powerful techniques for revealing latent hidden variables and associated patterns from high dimensional data. However, these methods often ignore the sparsity property of the graph, and hence usually incur too high memory and computational cost to be practical.

We propose a novel method, the Compact Matrix Decomposition (CMD), to compute sparse low rank approximations. CMD dramatically reduces both the computation cost and the space requirements over existing decomposition methods.

We need to understand

- how networks emerge
- what are their properties
- features
- how to calculate interesting/important characteristics

Graph Mining: Laws, Generators, and Algorithms

DEEPAV CHAKRABARTI AND CHRISTOS FALOUTSOS

Yahoo! Research and Carnegie Mellon University

How does the Web look? How could we tell an abnormal social network from a normal one? These and similar questions are important in many fields where the data can intuitively be cast as a graph; examples range from computer networks to sociology to biology and many more. Indeed, any “N:N” relation in database terminology can be represented as a graph. A lot of these questions boil down to the following: “How can we generate synthetic but realistic graphs?” To answer this, we must first understand what patterns are common in real-world graphs and can thus be considered a mark of normality/unormality. This survey gives an overview of the incredible variety of work that has been done on these problems. One of our main contributions is the integration of points of view from physics, mathematics, sociology, and computer science. Further, we briefly describe recent advances on some related and interesting graph problems.

Categories and Subject Descriptions: E.1 [Data Structures]
General Terms: Algorithms, Measurement

Additional Key Words and Phrases: Generators, graphs, patterns, social networks

http://www.cs.cmu.edu/~deepay/mywww/papers/csur06.pdf
ACM Computing Surveys, Vol. 38, March 2006,
Graph Generators

- Random Graph Generators
  - Connect nodes using random probabilities

- Preferential Attachment Generators
  - Give preference to nodes with more edges

- Optimization-based Generators
  - Minimize risks under limited resources

- Geographical Models
  - Geography affects network growth and topology

- Internet-specific Generators
  - Fit special features of the Internet

Fig. 8. Overview of graph generators. Current generators can be mostly placed under one of these categories, though there are some hybrids such as BRITE and Inet.

- Graphs and networks – also other types

- http://www.weizmann.ac.il/mcb/UriAlon/
  - => Complex networks
Superfamilies of Evolved and Designed Networks

Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, Uri Alon

Complex biological, technological, and sociological networks can be of very different sizes and connectivities, making it difficult to compare their structures. Here we present an approach to systematically study similarity in the local structure of networks, based on the significance profile (SP) of small subgraphs in the network compared to randomized networks. We find several superfamilies of previously unrelated networks with very similar SPs. One superfamily, including transcription networks of microorganisms, represents "rate-limited" information-processing networks strongly constrained by the response time of their components. A distinct superfamily includes protein signaling, developmental genetic networks, and neuronal wiring. Additional superfamilies include power grids, protein-structure networks and geometric networks, World Wide Web links and social networks, and word-adjacency networks from different languages.
Bias on the Web

“Googlearchy”:
How a Few Heavily-Linked Sites Dominate Politics
on the Web*

Matthew Hindman1, Kostas Tsoutsourelakis2, Judy A. Johnson3
March 31, 2003

Popularity and PageRank

"long tail"
"scale-free"
"rich-get-richer"

PageRank

\[ p(i) = \frac{\alpha}{N} + (1 - \alpha) \sum_{j:j \rightarrow i} \frac{p(j)}{|\ell : j \rightarrow \ell|} \]

Brin & Page 1998

Broder & al. 2000
Modeling search engine bias from the relationship between indegree and traffic:

1. traffic ~ P(click)
2. P(click) ~ f(rank)
3. rank ~ f(PageRank)
4. PageRank ~ f(indegree)
Surfing without search engines: popularity reflects rich-get-richer bias of the Web.

Googlearchy: search engines amplify rich-get-richer bias of the Web.
Empirical data: search mitigates rich-get-richer bias of the Web
Conclusions

- The use of search engines partially mitigates the rich-get-richer nature of the Web, giving new sites an increased chance of being discovered (compared to surfing alone), as long as they are about specific topics that match the interests of users.

- The combination of (i) how search engines index and rank results, (ii) what queries users submit, and (iii) how users view the results, leads to an egalitarian effect (“Googlocracy”).

Three Lectures

1. The effect of search engines and their ranking algorithms on the dynamics and evolution of the Web

2. Analyzing and modeling the Web traffic network and the avalanche dynamics of online popularity

3. Two applications of the social Web:
   - mining bookmark annotations to build a bottom-up semantic network
   - distributed collaborative Web search by adaptive intelligent peers
The power of social media

- Flickr – community phenomenon
- Millions of users share and tag each others’ photographs (why???)
- The *wisdom of the crowds* can be used to search
  - Ranking features to Yahoo! Answers
- The principle is not new – anchor text used in “standard” search
- What about generating pseudo-semantic resources?

An Introduction to Web Mining, 2008

Ricardo Baeza-Yates

Jaak Vilo and other authors

UT: Data Mining 2009

STACC
Software Technologies and Applications Competence Center

Jaak Vilo

November 26, 2009
Partners

Research Tracks and Programs

Data Integration and Mining (DIM)

1.1 Web Analytics and Social Network Analysis
Delfi, Logica, Quretec, Regio, Skype

1.2 Biomedical data integration and mining
ITK, Cybernetica, Quretec

1.3 Privacy-Preserving Data Mining
Cybernetica, Swedbank, Quretec

Software and Services Engineering

2.1 Smart Internet Interfaces
Delfi, Regio, Webmedia

2.2 Smart Services
Webmedia, Regio, Logica, Cybernetica

2.3 Software Development Productivity
Webmedia, Logica, Cybernetica, KnowIT
1.1 Web Analytics and Social Network Analysis

- Methods for analysing the structure and dynamics of very large social networks and raw web usage data in order to discover user clusters, user goals, service or product consumption patterns, customer churning patterns, spam and fraud patterns and other patterns of individual or collective user behaviour.
- Application to user interface personalization and re-organization, personalized search, targeted advertising, peer-to-peer network monitoring, and derivation of e-Business metrics associated to advertising, customer acquisition and retention/churning, and business intelligence.

Delfi, Logica, Quretec, Regio, Skype 163
Web usage mining

STACC
Software Technology and Applications Competence Centre
Activities

1. Social Network Analysis
   - mining of very large (400M nodes, x 10 edges) graphs for elementary properties

2. Web (and software) log mining
   - warehousing
   - clustering, visualisation, decision support ...

3. User Intent Prediction
   - real time learning of user intent and goals
   - collaborative filtering
1.2 Biomedical data integration and mining

- To develop **data integration and analysis methods** for **electronic patient records** and **biomarker data** with the goal of improving the (early) diagnosis and medical treatment of the diseases.
  - Complex disease of COPD as the proof of principle
  - Mining Electronic patient records and E-Health data
- UT – data mining; TTU – medical know-how and data
- Hospital and e-health solutions

ITK, Cybernetica, Quretec, + E-Health, Medicum, Finnish NIH
Activities

1. COPD patient cohort buildup based on smoking etc characteristics
2. Data warehousing and decision support for hospital e-health data
3. Introducing ontologies into clinic
4. Text mining of medical records

1.3 Privacy-Preserving Data Mining

To develop and to evaluate privacy-enhancing methods for data storage and processing, along two complementary directions:

1. **Security of micro-data releases and query auditing:**
   Detection and elimination of possible privacy breaches in data to be published, and detection of queries to (medical and financial) databases that may breach the privacy of individuals.

2. **Privacy-preserving data aggregation:** Development of secure and practically efficient methods for aggregating data from multiple sources, that leak nothing beyond the end aggregate results.
Privacy preserving DM

Q: sexual behavior of HIV patients?

# sex partners:

2

34

0

AVERAGE = 12

876113

126531

764224

9823

1235

68612

# sex partners:

2

34

0
Privacy preserving DM

ShareMind
Activities in 1.3

1. Micro-data protection mechanisms
2. An environment for developing privacy-preserving applications
3. Demonstration: questionnaire system
4. Protocol analysis for secret-shared applications
The first record is dated 10/01/2009, last record is dated 10/31/2009.

<table>
<thead>
<tr>
<th>Username</th>
<th>Group</th>
<th>#jobs</th>
<th>Wallclock</th>
<th>Average</th>
<th>Average</th>
<th>Full name</th>
</tr>
</thead>
<tbody>
<tr>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>TOTAL</td>
<td>-</td>
<td>36187</td>
<td>7293.91</td>
<td>100.00</td>
<td>1.64</td>
<td>0.07</td>
</tr>
<tr>
<td>cipo</td>
<td>users</td>
<td>363</td>
<td>3445.66</td>
<td>47.24</td>
<td>1.85</td>
<td>0.07 Heiki Kasemagi</td>
</tr>
<tr>
<td>thcmob</td>
<td>users</td>
<td>33397</td>
<td>1427.46</td>
<td>19.57</td>
<td>1.00</td>
<td>0.06 Juri Reimand</td>
</tr>
<tr>
<td>bronto</td>
<td>hirlam</td>
<td>119</td>
<td>759.92</td>
<td>10.42</td>
<td>8.21</td>
<td>0.12 Andres Luhamae</td>
</tr>
<tr>
<td>ivan</td>
<td>users</td>
<td>102</td>
<td>566.69</td>
<td>7.77</td>
<td>7.77</td>
<td>0.13 Ivan Suphonisko</td>
</tr>
<tr>
<td>bayaziti</td>
<td>users</td>
<td>649</td>
<td>510.45</td>
<td>7.00</td>
<td>1.00</td>
<td>0.31 Bayazit Yunusbaev</td>
</tr>
<tr>
<td>pirit85</td>
<td>users</td>
<td>69</td>
<td>184.23</td>
<td>2.53</td>
<td>1.00</td>
<td>0.17 Pirit Pimargi</td>
</tr>
<tr>
<td>maatb</td>
<td>users</td>
<td>467</td>
<td>106.21</td>
<td>1.01</td>
<td>1.01</td>
<td>0.21 Mait Metspalu</td>
</tr>
<tr>
<td>alfonso</td>
<td>users</td>
<td>69</td>
<td>57.53</td>
<td>0.79</td>
<td>13.80</td>
<td>0.04 Jaas Jezov</td>
</tr>
<tr>
<td>jaas</td>
<td>users</td>
<td>58</td>
<td>21.12</td>
<td>0.29</td>
<td>1.00</td>
<td>0.01 Slim Sober</td>
</tr>
<tr>
<td>siima</td>
<td>users</td>
<td>629</td>
<td>16.87</td>
<td>0.23</td>
<td>1.00</td>
<td>0.08 Tonu Zako</td>
</tr>
<tr>
<td>tluusko</td>
<td>users</td>
<td>40</td>
<td>13.72</td>
<td>0.19</td>
<td>1.00</td>
<td>0.03 Kaur Alasoo</td>
</tr>
<tr>
<td>a72094</td>
<td>users</td>
<td>43</td>
<td>2.07</td>
<td>0.03</td>
<td>1.00</td>
<td>0.00 Lauri Juhani Liivamagi</td>
</tr>
<tr>
<td>lauria</td>
<td>users</td>
<td>95</td>
<td>0.89</td>
<td>0.01</td>
<td>9.31</td>
<td>0.00 Lauri Anton</td>
</tr>
<tr>
<td>reidar</td>
<td>users</td>
<td>46</td>
<td>0.81</td>
<td>0.01</td>
<td>1.00</td>
<td>1.29 Reidar Anderson</td>
</tr>
<tr>
<td>kasak105</td>
<td>users</td>
<td>4</td>
<td>0.17</td>
<td>0.00</td>
<td>8.00</td>
<td>0.09 Kait Kasak</td>
</tr>
<tr>
<td>eero</td>
<td>users</td>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00 Eero Vainikko</td>
</tr>
</tbody>
</table>
Conclusions

• Data Mining is a rich and diverse field

• Driven by data, curiosity, business value

• Algorithmics, Statistics, Visualisation, ...

• Decision support/actionable information