Data Warehousing
資料倉儲

Social Network Analysis, Link Mining, Text and Web Mining

992DW08
MI4
Tue. 8,9 (15:10-17:00) L413

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.im.tku.edu.tw/~myday/
2011-05-10

Syllabus

1	100/02/15	Introduction to Data Warehousing
2	100/02/22	Data Warehousing, Data Mining, and Business Intelligence
3	100/03/01	Data Preprocessing: Integration and the ETL process
4	100/03/08	Data Warehouse and OLAP Technology
5	100/03/15	Data Warehouse and OLAP Technology
6	100/03/22	Data Warehouse and OLAP Technology
7	100/03/29	Data Warehouse and OLAP Technology
8	100/04/05	(放假一天) (民族掃墓節)
9	100/04/12	Data Cube Computation and Data Generation
10	100/04/19	Mid-Term Exam (期中考試週)
11	100/04/26	Association Analysis
12	100/05/03	Classification and Prediction, Cluster Analysis
13	100/05/10	Social Network Analysis, Link Mining, Text and Web Mining
14	100/05/17	Project Presentation
15	100/05/24	Final Exam (畢業班考試)

Learning Objective

- Social Network Analysis
- Link Mining
- Text and Web Mining

Social Network Analysis

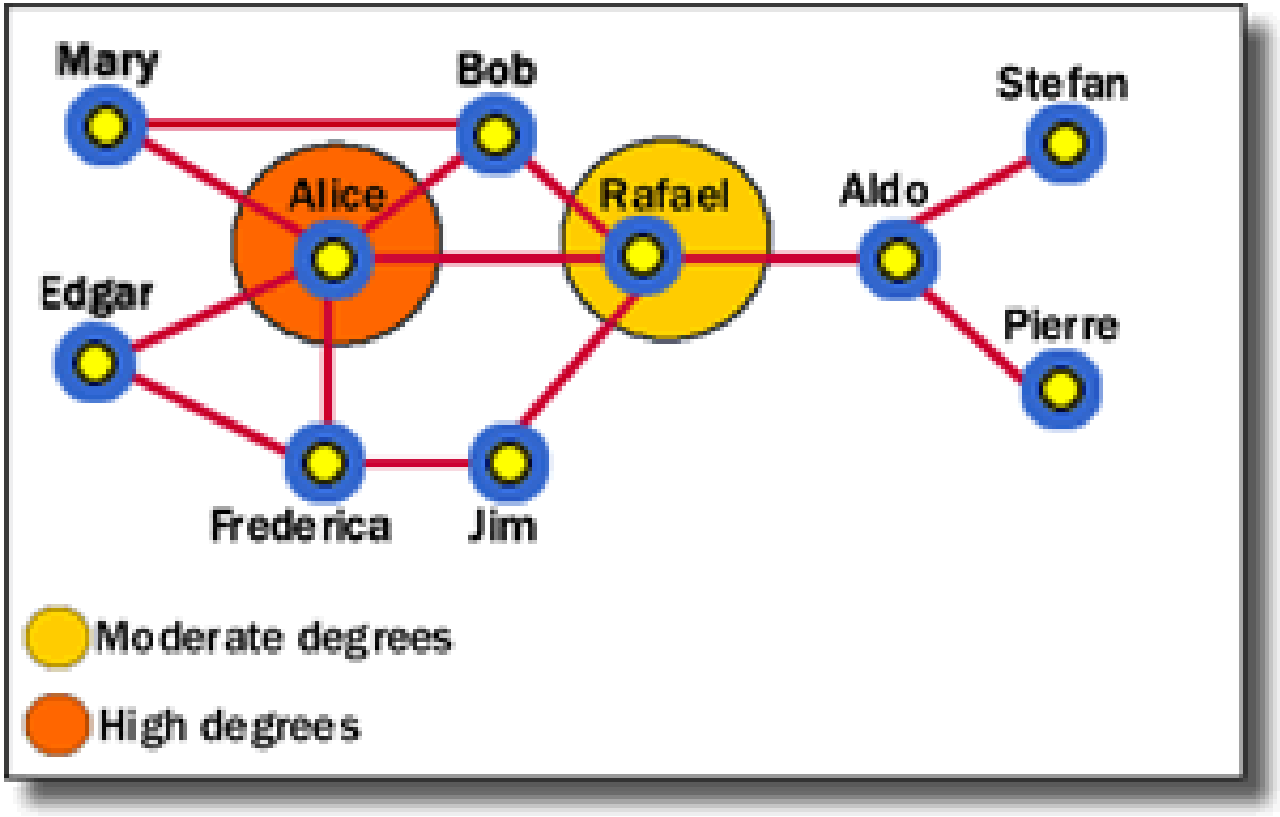
- A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest
- **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior

Social Network Analysis

- Using Social Network Analysis, you can get answers to questions like:
  - How highly connected is an entity within a network?
  - What is an entity's overall importance in a network?
  - How central is an entity within a network?
  - How does information flow within a network?

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

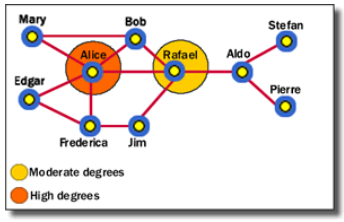
Social Network Analysis: Degree Centrality



Alice has the highest degree centrality, which means that she is quite active in the network. However, she is not necessarily the most powerful person because she is only directly connected within one degree to people in her clique—she has to go through Rafael to get to other cliques.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

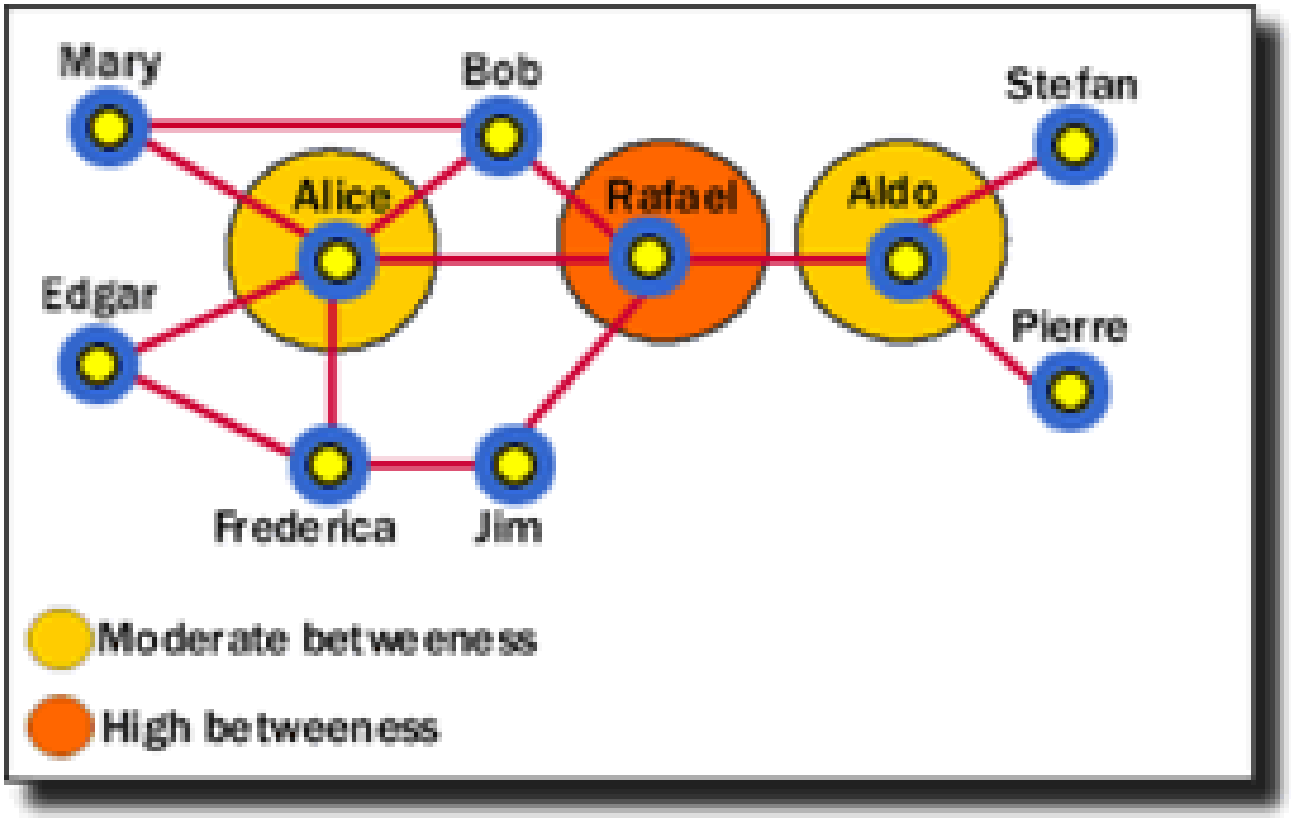
Social Network Analysis: Degree Centrality



- Degree centrality is simply the number of direct relationships that an entity has.
- An entity with high degree centrality:
  - Is generally an active player in the network.
  - Is often a connector or hub in the network.
  - s not necessarily the most connected entity in the network (an entity may have a large number of relationships, the majority of which point to low-level entities).
  - May be in an advantaged position in the network.
  - May have alternative avenues to satisfy organizational needs, and consequently may be less dependent on other individuals.
  - Can often be identified as third parties or deal makers.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

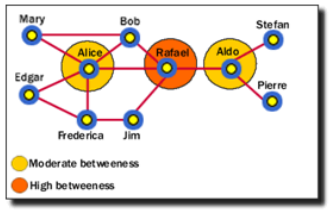
Social Network Analysis: Betweenness Centrality



Rafael has the highest betweenness because he is between Alice and Aldo, who are between other entities. Alice and Aldo have a slightly lower betweenness because they are essentially only between their own cliques. Therefore, although Alice has a higher degree centrality, Rafael has more importance in the network in certain respects.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

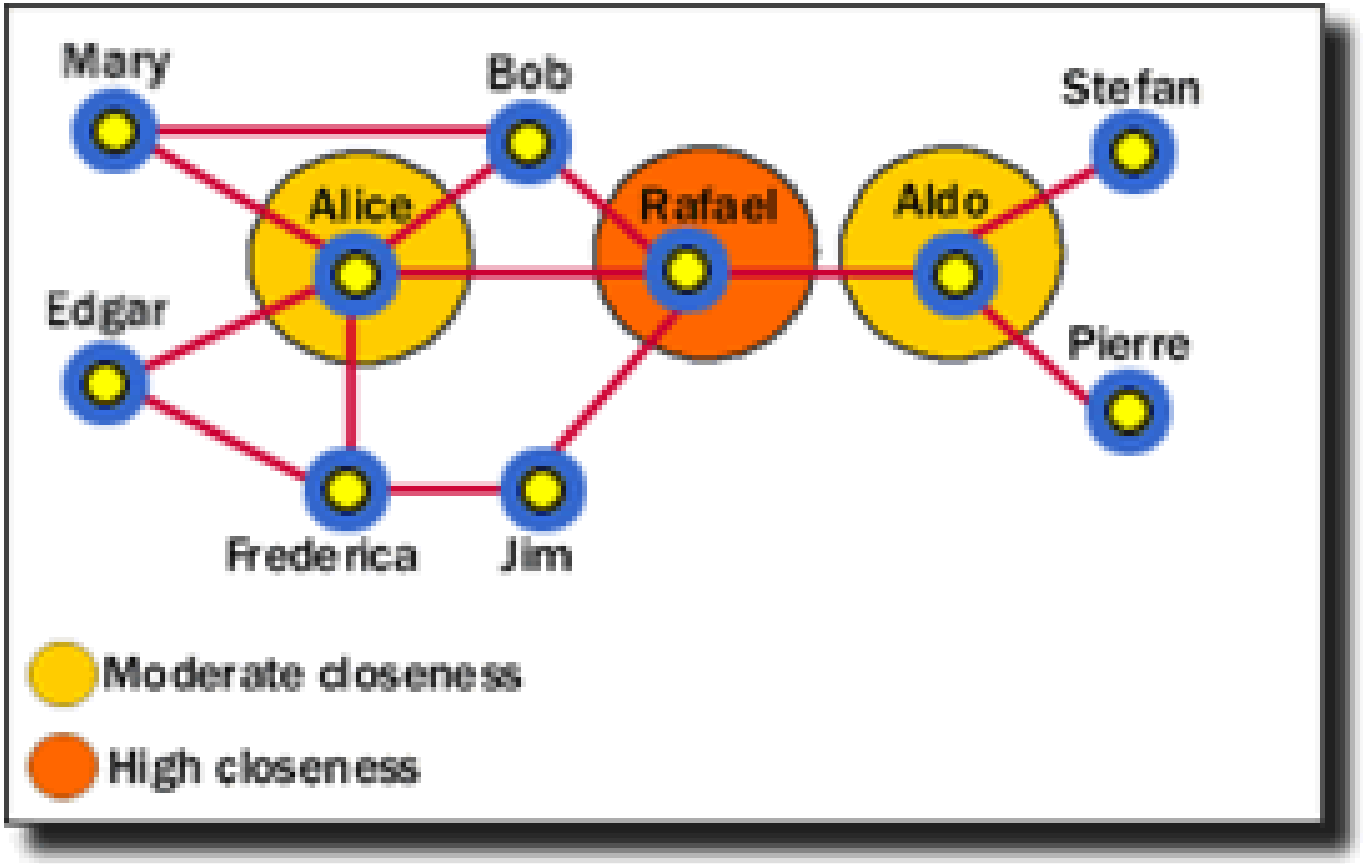
Social Network Analysis: Betweenness Centrality



- Betweenness centrality identifies an entity's position within a network in terms of its ability to make connections to other pairs or groups in a network.
- An entity with a high betweenness centrality generally:
  - Holds a favored or powerful position in the network.
  - Represents a single point of failure—take the single betweenness spanner out of a network and you sever ties between cliques.
  - Has a greater amount of influence over what happens in a network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

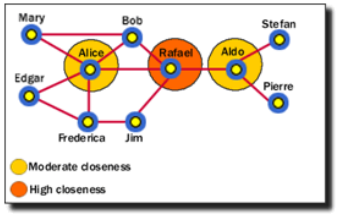
Social Network Analysis: Closeness Centrality



Rafael has the highest closeness centrality because he can reach more entities through shorter paths. As such, Rafael's placement allows him to connect to entities in his own clique, and to entities that span cliques.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

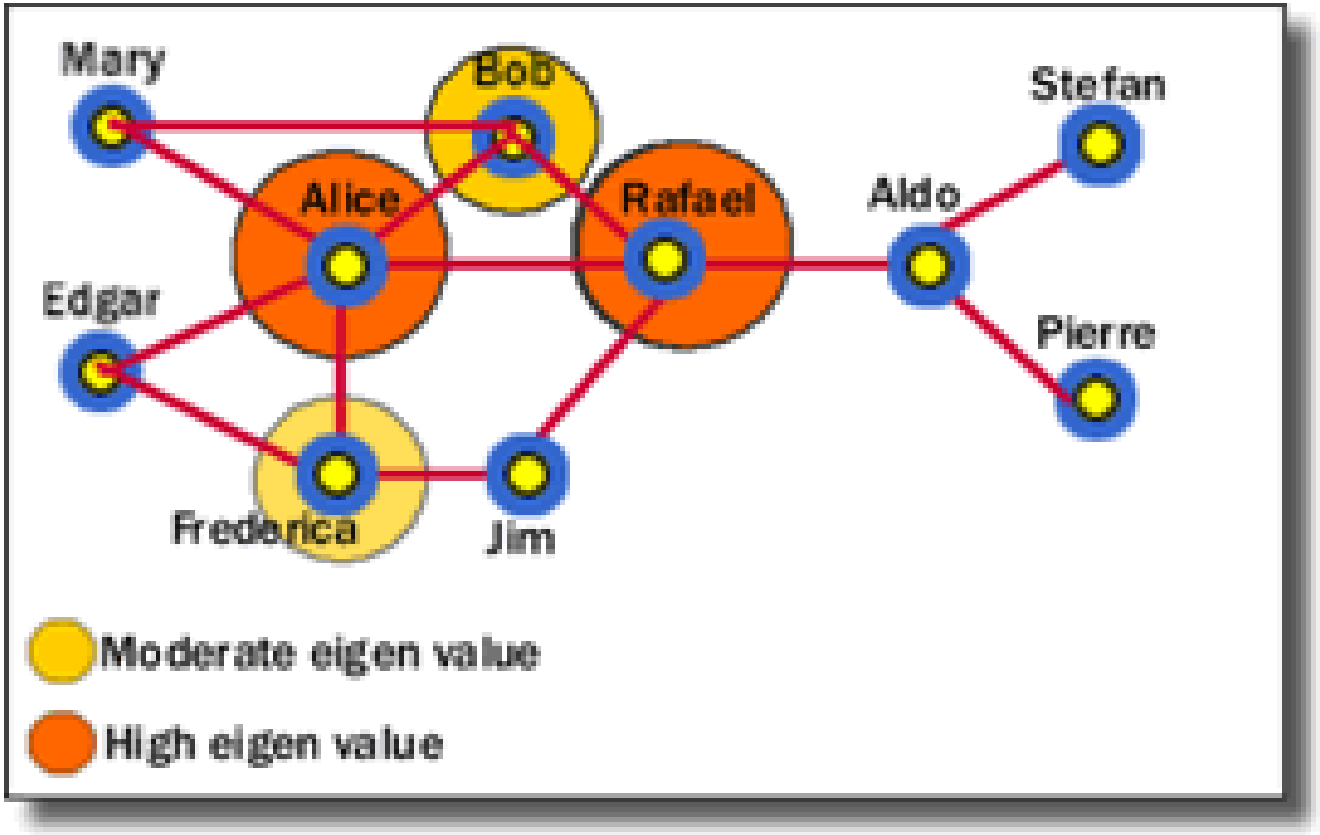
Social Network Analysis: Closeness Centrality



- Closeness centrality measures how quickly an entity can access more entities in a network.
- An entity with a high closeness centrality generally:
  - Has quick access to other entities in a network.
  - Has a short path to other entities.
  - Is close to other entities.
  - Has high visibility as to what is happening in the network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

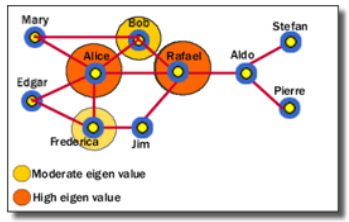
Social Network Analysis: Eigenvalue



Alice and Rafael are closer to other highly close entities in the network. Bob and Frederica are also highly close, but to a lesser value.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

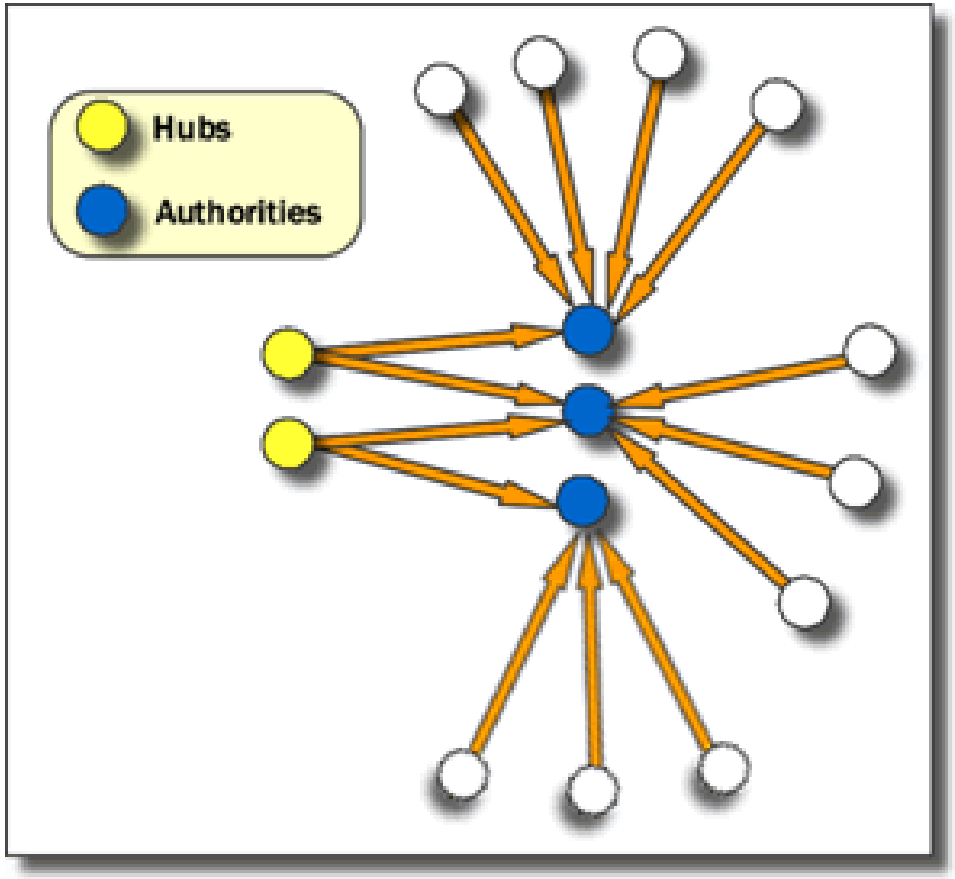
Social Network Analysis: Eigenvalue



- Eigenvalue measures how close an entity is to other highly close entities within a network. In other words, Eigenvalue identifies the most central entities in terms of the global or overall makeup of the network.
- A high Eigenvalue generally:
  - Indicates an actor that is more central to the main pattern of distances among all entities.
  - Is a reasonable measure of one aspect of centrality in terms of positional advantage.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

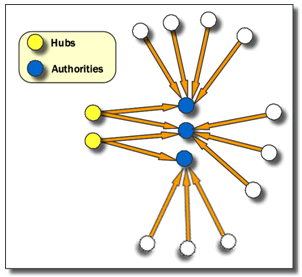
Social Network Analysis: Hub and Authority



Hubs are entities that point to a relatively large number of authorities. They are essentially the mutually reinforcing analogues to authorities. Authorities point to high hubs. Hubs point to high authorities. You cannot have one without the other.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

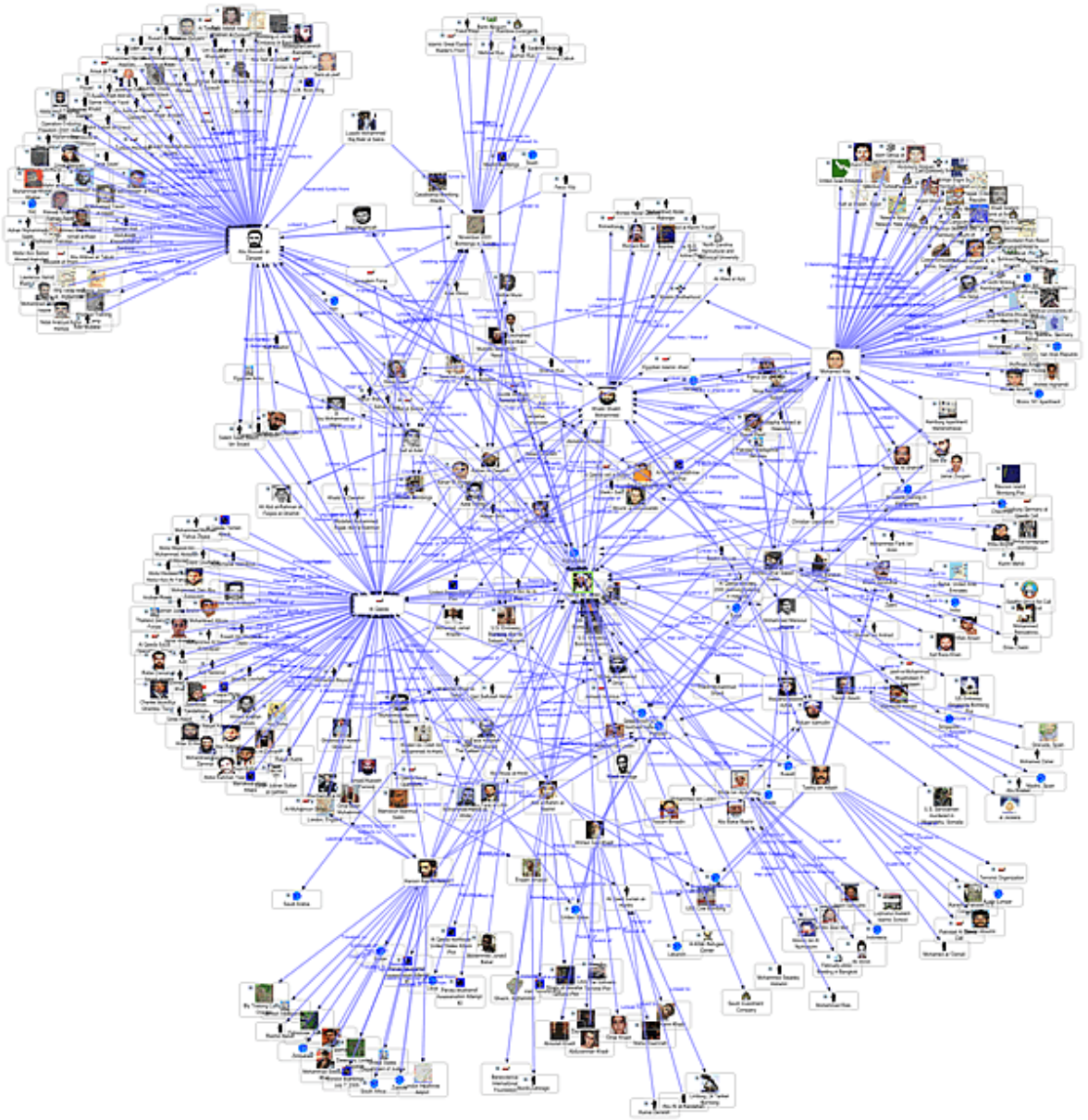
Social Network Analysis: Hub and Authority



- Entities that many other entities point to are called Authorities. In Sentinel Visualizer, relationships are directional—they point from one entity to another.
- If an entity has a high number of relationships pointing to it, it has a high authority value, and generally:
  - Is a knowledge or organizational authority within a domain.
  - Acts as definitive source of information.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis



Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis

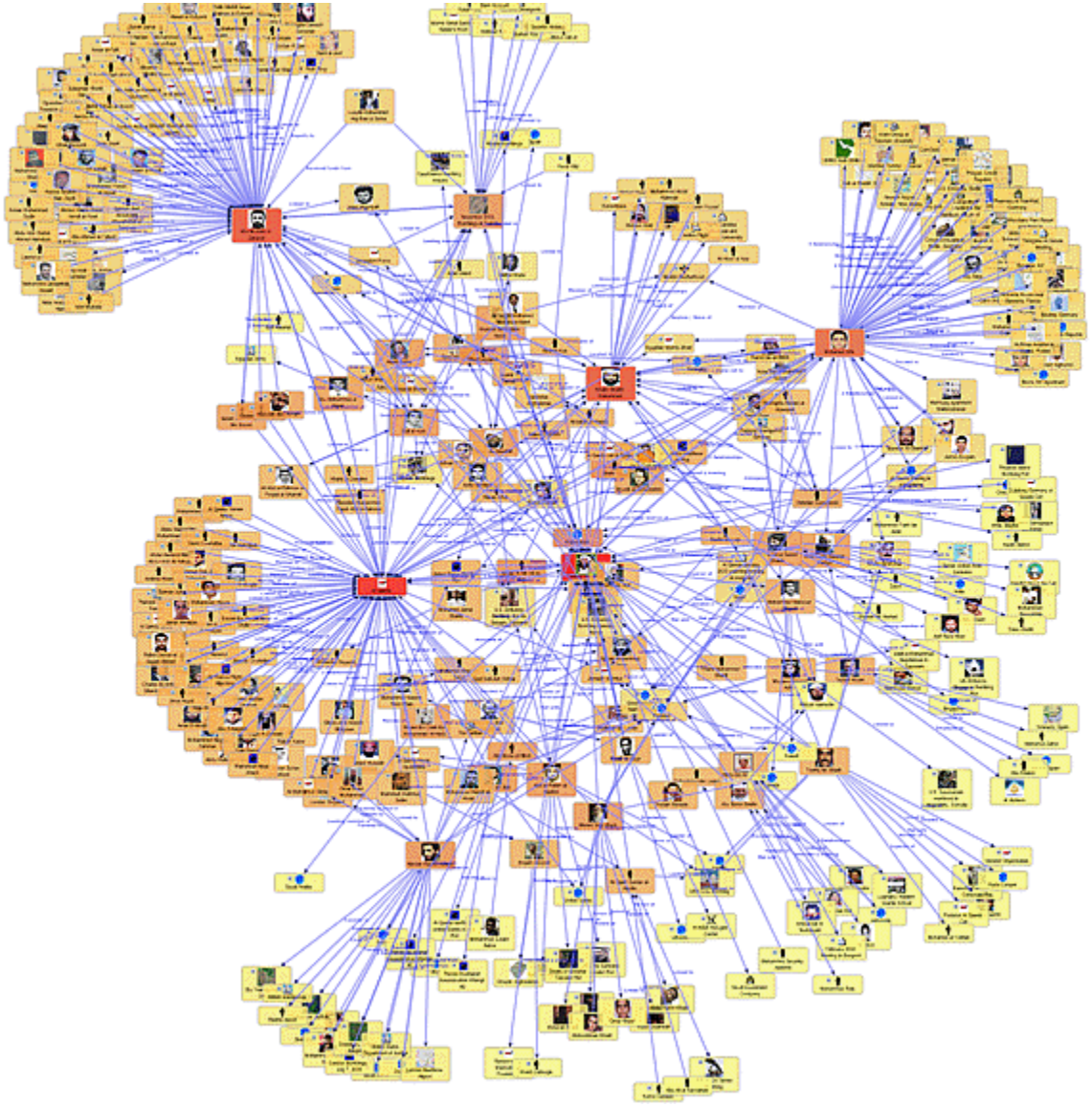
Network Metrics

Calculate | CardView | TableView | Group area | Expand groups | Collapse groups

Name	Type	Degree	Betweenness	Closeness	Eigenvalue	Hub	Authority
Osama bin Laden	Person	44	0.920492092358...	1	0.0271	0	0.011
Abdallah Al-Halabi	Person	2	0	0.654367256637...	0.0001	0	0
Abu Mussab al-Zarqawi	Person	84	0.934837847326...	0.869451697127...	0.7028	0.6572	0.1076
Al Qaeda	Terrorist Organiz...	85	1	0.962427745664...	0.0416	0.3941	0.0166
Ayman Al-Zawahiri	Person	14	0.045794908783...	0.716129032258...	0	0	0.0173
Enaam Arnaout	Person	4	0.031189325814...	0.656804733727...	0.0001	0	0
Imad Eddin Barakat Yarbas	Person	11	0.065049589039...	0.704016913319...	0.0015	0	0.0025
Khalid Shaikh Mohammed	Person	32	0.339916464724...	0.868059817945...	0.002	0	0.1528
Mohamed Atta	Person	61	0.666268740074...	0.820197044334...	0.0015	0	0.6316

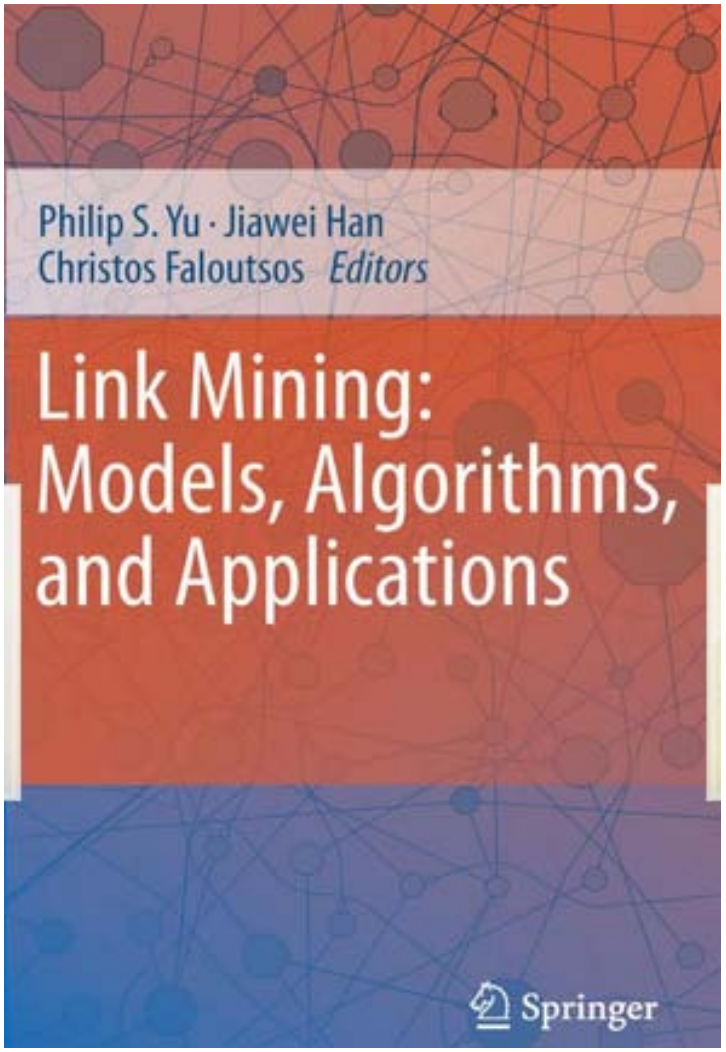
Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis



Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Link Mining



http://www.amazon.com/Link-Mining-Models-Algorithms-Applications/dp/1441965149

Link Mining
(Getoor & Diehl, 2005)

- Link Mining
  - Data Mining techniques that take into account the links between objects and entities while building predictive or descriptive models.
- Link based object ranking, Group Detection, Entity Resolution, Link Prediction
- Application:
  - Hyperlink Mining
  - Relational Learning
  - Inductive Logic Programming
  - Graph Mining

Characteristics of Collaboration Networks

(Newman, 2001; 2003; 3004)

- Degree distribution follows a power-law
- Average separation decreases in time.
- Clustering coefficient decays with time
- Relative size of the largest cluster increases
- Average degree increases
- Node selection is governed by preferential attachment

Social Network Techniques

- Social network extraction/construction
- Link prediction
- Approximating large social networks
- Identifying prominent/trusted/expert actors in social networks
- Search in social networks
- Discovering communities in social network
- Knowledge discovery from social network

Social Network Extraction

- Mining a social network from data sources
- Three sources of social network (Hope et al., 2006)
  - Content available on web pages
    - E.g., user homepages, message threads
  - User interaction logs
    - E.g., email and messenger chat logs
  - Social interaction information provided by users
    - E.g., social network service websites (Facebook)

Social Network Extraction

- IR based extraction from web documents
  - Construct an "actor-by-term" matrix
  - The terms associated with an actor come from web pages/documents created by or associated with that actor
  - IR techniques (TF-IDF, LSI, cosine matching, intuitive heuristic measures) are used to quantify similarity between two actors' term vectors
  - The similarity scores are the edge label in the network
    - Thresholds on the similarity measure can be used in order to work with binary or categorical edge labels
    - Include edges between an actor and its k-nearest neighbors
- Co-occurrence based extraction from web documents

Link Prediction

- Link Prediction using supervised learning (Hasan et al., 2006)
  - Citation Network (BIOBASE, DBLP)
  - Use machine learning algorithms to predict future co-authorship
    - Decision three, k-NN, multilayer perceptron, SVM, RBF network
  - Identify a group of features that are most helpful in prediction
  - Best Predictor Features
    - Keywork Match count, Sum of neighbors, Sum of Papers, Shortest distance

Identifying Prominent Actors in a Social Network

- Compute scores/ranking over the set (or a subset) of actors in the social network which indicate degree of importance / expertise / influence
  - E.g., Pagerank, HITS, centrality measures
- Various algorithms from the link analysis domain
  - PageRank and its many variants
  - HITS algorithm for determining authoritative sources
- Centrality measures exist in the social science domain for measuring importance of actors in a social network

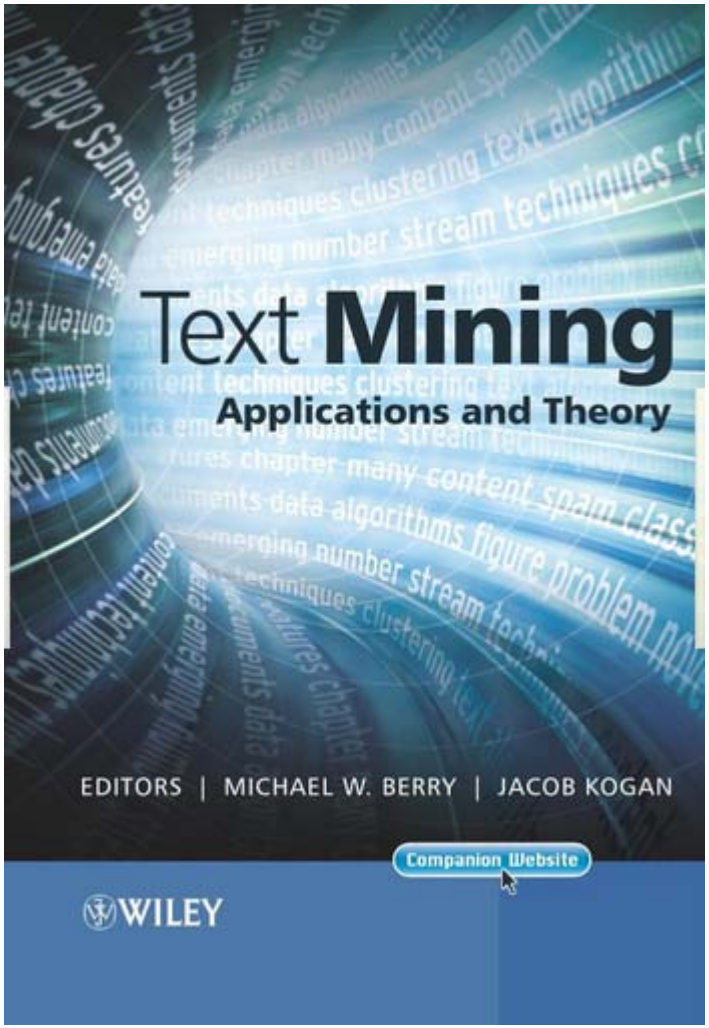
Identifying Prominent Actors in a Social Network

- Brandes, 2011
- Prominence→ high betweenness value
- Betweenness centrality requires computation of number of shortest paths passing through each node
- Compute shortest paths between all pairs of vertices

Text and Web Mining

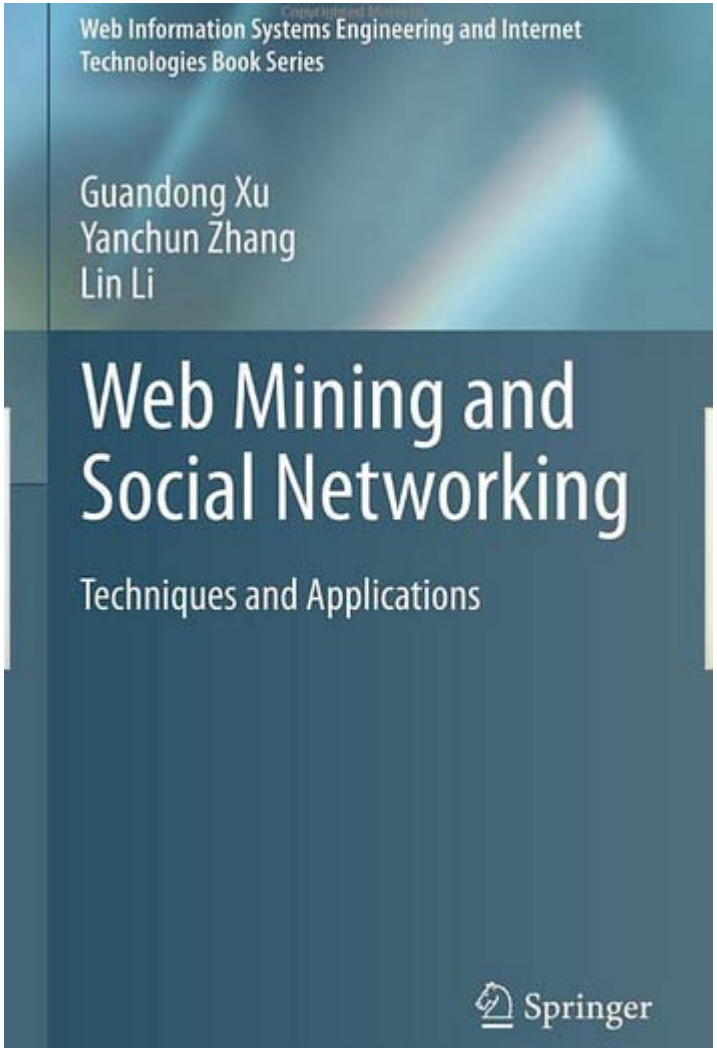
- Text Mining: Applications and Theory
- Web Mining and Social Networking
- Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites
- Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data
- Search Engines – Information Retrieval in Practice

Text Mining



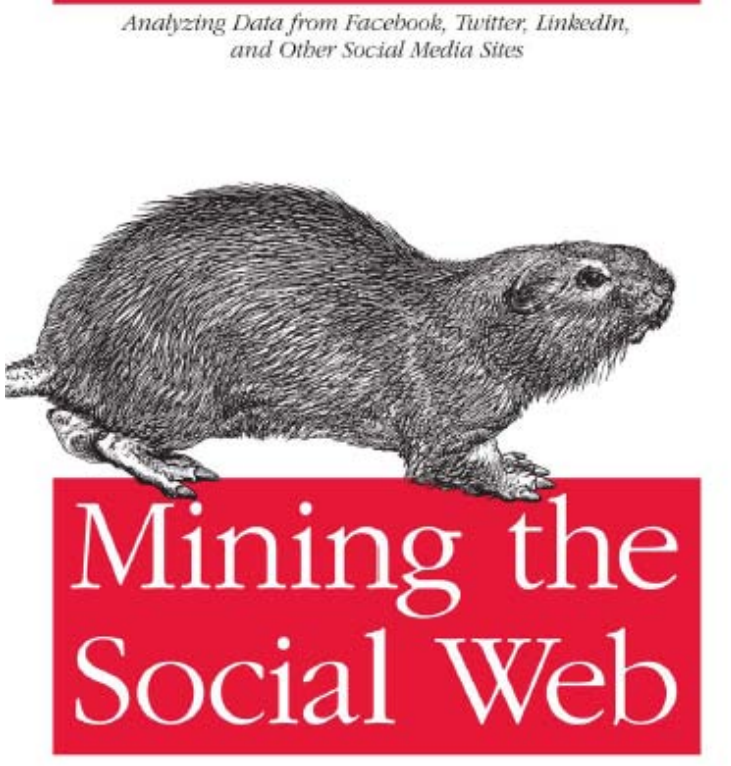
http://www.amazon.com/Text-Mining-Applications-Michael-Berry/dp/0470749822/

Web Mining and Social Networking



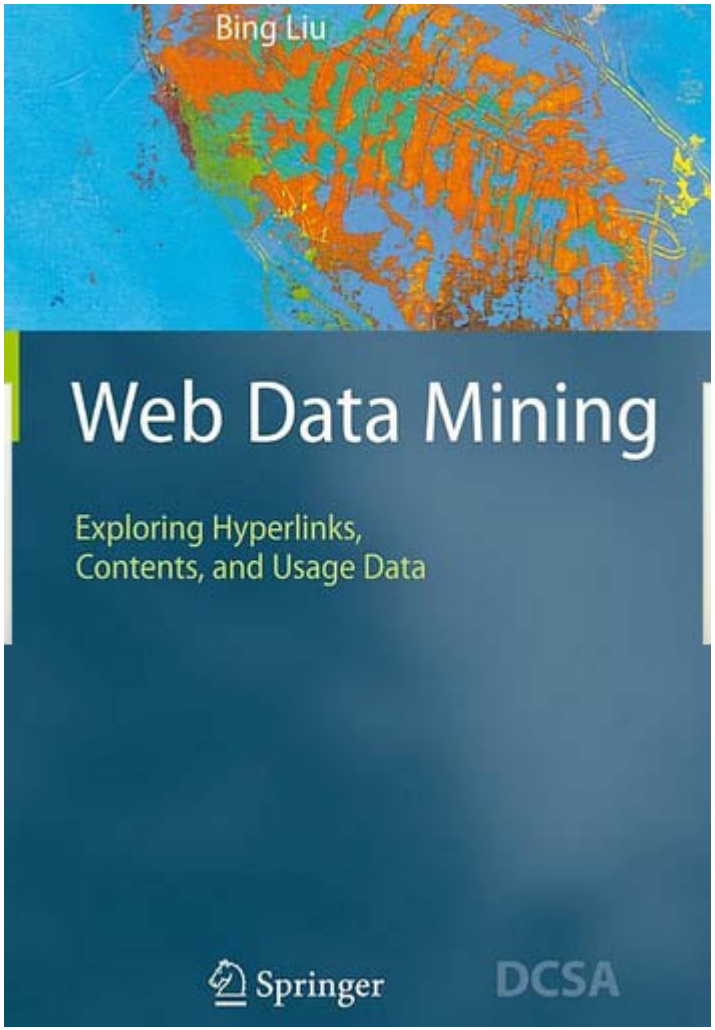
http://www.amazon.com/Web-Mining-Social-Networking-Applications/dp/1441977341

Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites



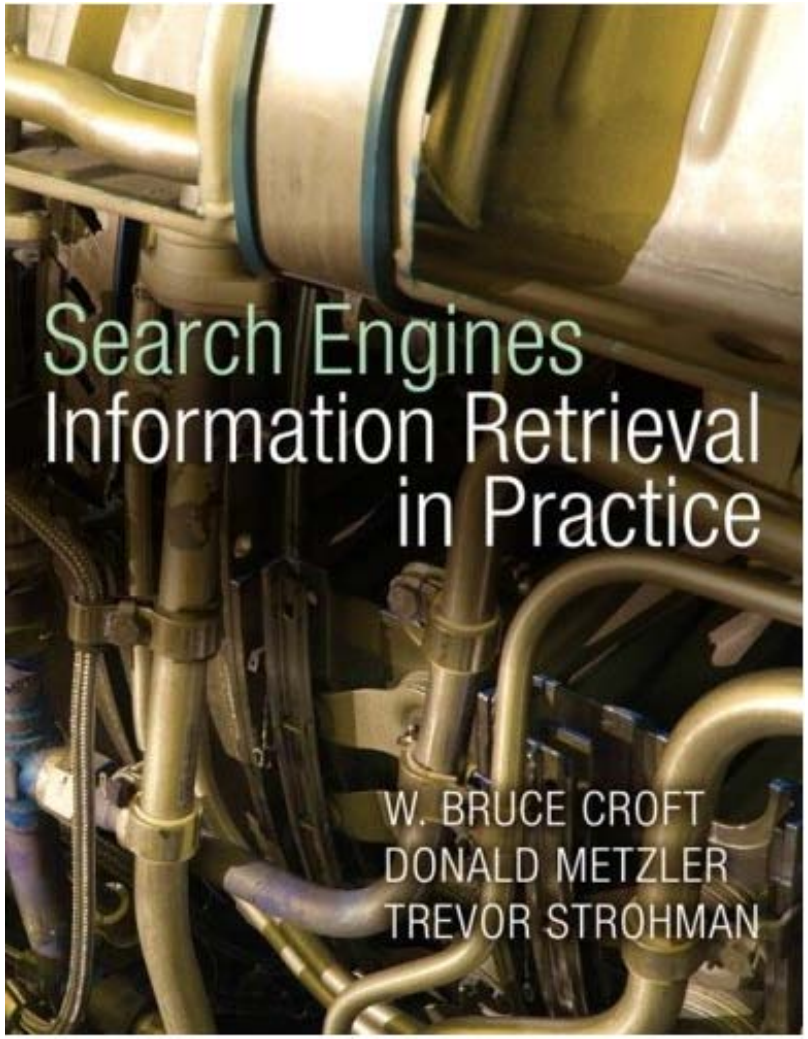
http://www.amazon.com/Mining-Social-Web-Analyzing-Facebook/dp/1449388345

Web Data Mining:
Exploring Hyperlinks, Contents, and Usage Data



http://www.amazon.com/Web-Data-Mining-Data-Centric-Applications/dp/3540378812

Search Engines: Information Retrieval in Practice



http://www.amazon.com/Search-Engines-Information-Retrieval-Practice/dp/0136072240

Text Mining

- Text mining (text data mining)
  - the process of deriving high-quality information from text
- Typical text mining tasks
  - text categorization
  - text clustering
  - concept/entity extraction
  - production of granular taxonomies
  - sentiment analysis
  - document summarization
  - entity relation modeling
    - i.e., learning relations between named entities.

http://en.wikipedia.org/wiki/Text_mining

Web Mining

- Web mining
  - discover useful information or knowledge from the **Web hyperlink structure, page content, and usage data.**
- Three types of web mining tasks
  - Web structure mining
  - Web content mining
  - Web usage mining

Source: Bing Liu (2009) Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data

Processing Text

- Converting documents to *index terms*
- Why?
  - Matching the exact string of characters typed by the user is too restrictive
    - i.e., it doesn't work very well in terms of effectiveness
  - Not all words are of equal value in a search
  - Sometimes not clear where words begin and end
    - Not even clear what a word is in some languages
      - e.g., Chinese, Korean

Text Statistics

- Huge variety of words used in text but
- Many statistical characteristics of word occurrences are predictable
  - e.g., distribution of word counts
- Retrieval models and ranking algorithms depend heavily on statistical properties of words
  - e.g., important words occur often in documents but are not high frequency in collection

Tokenizing

- Forming words from sequence of characters
- Surprisingly complex in English, can be harder in other languages
- Early IR systems:
  - any sequence of alphanumeric characters of length 3 or more
  - terminated by a space or other special character
  - upper-case changed to lower-case

Tokenizing

- Example:
  - “Bigcorp's 2007 bi-annual report showed profits rose 10%.” becomes
  - “bigcorp 2007 annual report showed profits rose”
- Too simple for search applications or even large-scale experiments
- Why? Too much information lost
  - Small decisions in tokenizing can have major impact on effectiveness of some queries

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Tokenizing Problems

- Small words can be important in some queries, usually in combinations
  - xp, ma, pm, ben e king, el paso, master p, gm, j lo, world war II
- Both hyphenated and non-hyphenated forms of many words are common
  - Sometimes hyphen is not needed
    - e-bay, wal-mart, active-x, cd-rom, t-shirts
  - At other times, hyphens should be considered either as part of the word or a word separator
    - winston-salem, mazda rx-7, e-cards, pre-diabetes, t-mobile, spanish-speaking

Tokenizing Problems

- Special characters are an important part of tags, URLs, code in documents
- Capitalized words can have different meaning from lower case words
  - Bush, Apple
- Apostrophes can be a part of a word, a part of a possessive, or just a mistake
  - rosie o'donnell, can't, don't, 80's, 1890's, men's straw hats, master's degree, england's ten largest cities, shriner's

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Tokenizing Problems

- Numbers can be important, including decimals
  - nokia 3250, top 10 courses, united 93, quicktime 6.5 pro, 92.3 the beat, 288358
- Periods can occur in numbers, abbreviations, URLs, ends of sentences, and other situations
  - I.B.M., Ph.D., cs.umass.edu, F.E.A.R.
- Note: tokenizing steps for queries must be identical to steps for documents

Tokenizing Process

- First step is to use parser to identify appropriate parts of document to tokenize
- Defer complex decisions to other components
  - word is any sequence of alphanumeric characters, terminated by a space or special character, with everything converted to lower-case
  - everything indexed
  - example: 92.3 → 92 3 but search finds documents with 92 and 3 adjacent
  - incorporate some rules to reduce dependence on query transformation components

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Tokenizing Process

- Not that different than simple tokenizing process used in past
- Examples of rules used with TREC
  - Apostrophes in words ignored
    - o'connor → oconnor bob's → bobs
  - Periods in abbreviations ignored
    - I.B.M. → ibm Ph.D. → ph d

Stopping

- Function words (determiners, prepositions) have little meaning on their own
- High occurrence frequencies
- Treated as *stopwords* (i.e. removed)
  - reduce index space, improve response time, improve effectiveness
- Can be important in combinations
  - e.g., “to be or not to be”

Stopping

- Stopword list can be created from high-frequency words or based on a standard list
- Lists are customized for applications, domains, and even parts of documents
  - e.g., “click” is a good stopword for anchor text
- Best policy is to index all words in documents, make decisions about which words to use at query time

Stemming

- Many morphological variations of words
  - *inflectional* (plurals, tenses)
  - *derivational* (making verbs nouns etc.)
- In most cases, these have the same or very similar meanings
- Stemmers attempt to reduce morphological variations of words to a common stem
  - usually involves removing suffixes
- Can be done at indexing time or as part of query processing (like stopwords)

Stemming

- Generally a small but significant effectiveness improvement
  - can be crucial for some languages
  - e.g., 5-10% improvement for English, up to 50% in Arabic

kitab	*a book*
kitabi	*my book*
al**kitab**	*the book*
kitabuki	*your book* (f)
kitabuka	*your book (m)*
kitabuhu	*his book*
kat**ab**a	*to write*
ma**ktab**a	*library, bookstore*
ma**ktab**	*office*

Words with the Arabic root **ktb**

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Stemming

- Two basic types
  - Dictionary-based: uses lists of related words
  - Algorithmic: uses program to determine related words
- Algorithmic stemmers
  - *suffix-s:* remove 's' endings assuming plural
    - e.g., cats → cat, lakes → lake, wiis → wii
    - Many *false negatives*: supplies → supplie
    - Some *false positives*: ups → up

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Porter Stemmer

- Algorithmic stemmer used in IR experiments since the 70s
- Consists of a series of rules designed to the longest possible suffix at each step
- Effective in TREC
- Produces *stems* not *words*
- Makes a number of errors and difficult to modify

Porter Stemmer

- Example step (1 of 5)

Step 1a:

- Replace ***sses*** by ***ss*** (e.g., stresses → stress).
- Delete ***s*** if the preceding word part contains a vowel not immediately before the ***s*** (e.g., gaps → gap but gas → gas).
- Replace ***ied*** or ***ies*** by ***i*** if preceded by more than one letter, otherwise by ***ie*** (e.g., ties → tie, cries → cri).
- If suffix is ***us*** or ***ss*** do nothing (e.g., stress → stress).

Step 1b:

- Replace ***eed***, ***eedly*** by ***ee*** if it is in the part of the word after the first non-vowel following a vowel (e.g., agreed → agree, feed → feed).
- Delete ***ed***, ***edly***, ***ing***, ***ingly*** if the preceding word part contains a vowel, and then if the word ends in ***at***, ***bl***, or ***iz*** add ***e*** (e.g., fished → fish, pirating → pirate), or if the word ends with a double letter that is not ***ll***, ***ss*** or ***zz***, remove the last letter (e.g., falling→ fall, dripping → drip), or if the word is short, add ***e*** (e.g., hoping → hope).
- Whew!

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Porter Stemmer

False positives	*False negatives*
organization/organ	european/europe
generalization/generic	cylinder/cylindrical
numerical/numerous	matrices/matrix
policy/police	urgency/urgent
university/universe	create/creation
addition/additive	analysis/analyses
negligible/negligent	useful/usefully
execute/executive	noise/noisy
past/paste	decompose/decomposition
ignore/ignorant	sparse/sparsity
special/specialized	resolve/resolution
head/heading	triangle/triangular

- Porter2 stemmer addresses some of these issues
- Approach has been used with other languages

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Krovetz Stemmer

- Hybrid algorithmic-dictionary
  - Word checked in dictionary
    - If present, either left alone or replaced with “exception”
    - If not present, word is checked for suffixes that could be removed
    - After removal, dictionary is checked again
- Produces words not stems
- Comparable effectiveness
- Lower false positive rate, somewhat higher false negative

Stemmer Comparison

Original text:

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Porter stemmer:

document describ market strategi carri compani agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale market share stimul demand price cut volum sale

Krovetz stemmer:

document describe marketing strategy carry company agriculture chemical report prediction market share chemical report market statistic agrochemic pesticide herbicide fungicide insecticide fertilizer predict sale stimulate demand price cut volume sale

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Phrases

- Many queries are 2-3 word phrases
- Phrases are
  - More precise than single words
    - e.g., documents containing “black sea” vs. two words “black” and “sea”
  - Less ambiguous
    - e.g., “big apple” vs. “apple”
- Can be difficult for ranking
    - e.g., Given query “fishing supplies”, how do we score documents with
      - exact phrase many times, exact phrase just once, individual words in same sentence, same paragraph, whole document, variations on words?

Phrases

- Text processing issue – how are phrases recognized?
- Three possible approaches:
  - Identify syntactic phrases using a *part-of-speech* (POS) tagger
  - Use word *n-grams*
  - Store word positions in indexes and use *proximity operators* in queries

POS Tagging

- POS taggers use statistical models of text to predict syntactic tags of words
  - Example tags:
    - NN (singular noun), NNS (plural noun), VB (verb), VBD (verb, past tense), VBN (verb, past participle), IN (preposition), JJ (adjective), CC (conjunction, e.g., “and”, “or”), PRP (pronoun), and MD (modal auxiliary, e.g., “can”, “will”).
- Phrases can then be defined as simple noun groups, for example

Pos Tagging Example

Original text:

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Brill tagger:

Document/NN will/MD describe/VB marketing/NN strategies/NNS carried/VBD out/IN by/IN U.S./NNP companies/NNS for/IN their/PRP agricultural/JJ chemicals/NNS ,/, report/NN predictions/NNS for/IN market/NN share/NN of/IN such/JJ chemicals/NNS ,/, or/CC report/NN market/NN statistics/NNS for/IN agrochemicals/NNS ,/, pesticide/NN ,/, herbicide/NN ,/, fungicide/NN ,/, insecticide/NN ,/, fertilizer/NN ,/, predicted/VBN sales/NNS ,/, market/NN share/NN ,/, stimulate/VB demand/NN ,/, price/NN cut/NN ,/, volume/NN of/IN sales/NNS ./.

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Example Noun Phrases

TREC data		Patent data	
Frequency	*Phrase*	*Frequency*	*Phrase*
65824	united states	975362	present invention
61327	article type	191625	u.s. pat
33864	los angeles	147352	preferred embodiment
18062	hong kong	95097	carbon atoms
17788	north korea	87903	group consisting
17308	new york	81809	room temperature
15513	san diego	78458	seq id
15009	orange county	75850	brief description
12869	prime minister	66407	prior art
12799	first time	59828	perspective view
12067	soviet union	58724	first embodiment
10811	russian federation	56715	reaction mixture
9912	united nations	54619	detailed description
8127	southern california	54117	ethyl acetate
7640	south korea	52195	example 1
7620	end recording	52003	block diagram
7524	european union	46299	second embodiment
7436	south africa	41694	accompanying drawings
7362	san francisco	40554	output signal
7086	news conference	37911	first end
6792	city council	35827	second end
6348	middle east	34881	appended claims
6157	peace process	33947	distal end
5955	human rights	32338	cross-sectional view
5837	white house	30193	outer surface

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Word N-Grams

- POS tagging too slow for large collections
- Simpler definition – phrase is any sequence of *n* words – known as *n-grams*
  - *bigram*: 2 word sequence, *trigram*: 3 word sequence, *unigram*: single words
  - N-grams also used at character level for applications such as OCR
- N-grams typically formed from *overlapping* sequences of words
  - i.e. move n-word “window” one word at a time in document

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

N-Grams

- Frequent n-grams are more likely to be meaningful phrases
- N-grams form a Zipf distribution
  - Better fit than words alone
- Could index all n-grams up to specified length
  - Much faster than POS tagging
  - Uses a lot of storage
    - e.g., document containing 1,000 words would contain 3,990 instances of word n-grams of length $2 \leq n \leq 5$

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Google N-Grams

- Web search engines index n-grams
- Google sample:

Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

- Most frequent trigram in English is “all rights reserved”
  - In Chinese, “limited liability corporation”

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Document Structure and Markup

- Some parts of documents are more important than others
- Document parser recognizes structure using markup, such as HTML tags
  - Headers, anchor text, bolded text all likely to be important
  - Metadata can also be important
  - Links used for *link analysis*

Example Web Page

Tropical fish

From Wikipedia, the free encyclopedia

Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species. Fishkeepers often use the term *tropical fish* to refer only those requiring fresh water, with saltwater tropical fish referred to as *marine fish*.

Tropical fish are popular aquarium fish , due to their often bright coloration. In freshwater fish, this coloration typically derives from iridescence, while salt water fish are generally pigmented.

Example Web Page

```
<html>
<head>
<meta name="keywords" content="Tropical fish, Airstone, Albinism, Algae eater,
Aquarium, Aquarium fish feeder, Aquarium furniture, Aquascaping, Bath treatment
(fishkeeping),Berlin Method, Biotope" />
…
<title>Tropical fish - Wikipedia, the free encyclopedia</title>
</head>
<body>
 …
<h1 class="firstHeading">Tropical fish</h1>
…
<p><b>Tropical fish</b> include <a href="/wiki/Fish" title="Fish">fish</a> found in <a
href="/wiki/Tropics" title="Tropics">tropical</a> environments around the world,
including both <a href="/wiki/Fresh_water" title="Fresh water">freshwater</a> and <a
href="/wiki/Sea_water" title="Sea water">salt water</a> species. <a
href="/wiki/Fishkeeping" title="Fishkeeping">Fishkeepers</a> often use the term
<i>tropical fish</i> to refer only those requiring fresh water, with saltwater tropical fish
referred to as <i><a href="/wiki/List_of_marine_aquarium_fish_species" title="List of
marine aquarium fish species">marine fish</a></i>.</p>
<p>Tropical fish are popular <a href="/wiki/Aquarium" title="Aquarium">aquarium</a>
fish , due to their often bright coloration. In freshwater fish, this coloration typically
derives from <a href="/wiki/Iridescence" title="Iridescence">iridescence</a>, while salt
water fish are generally <a href="/wiki/Pigment" title="Pigment">pigmented</a>.</p>
…
</body></html>
```

Link Analysis

- Links are a key component of the Web
- Important for navigation, but also for search
  - e.g., Example website
  - “Example website” is the anchor text
  - “http://example.com” is the destination link
  - both are used by search engines

Anchor Text

- Used as a description of the content of the *destination page*
  - i.e., collection of anchor text in all links pointing to a page used as an additional text field
- Anchor text tends to be short, descriptive, and similar to query text
- Retrieval experiments have shown that anchor text has significant impact on effectiveness for *some types of queries*
  - i.e., more than PageRank

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

PageRank

- Billions of web pages, some more informative than others
- Links can be viewed as information about the *popularity* (*authority*?) of a web page
  - can be used by ranking algorithm
- *Inlink* count could be used as simple measure
- Link analysis algorithms like PageRank provide more reliable ratings
  - less susceptible to link spam

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Random Surfer Model

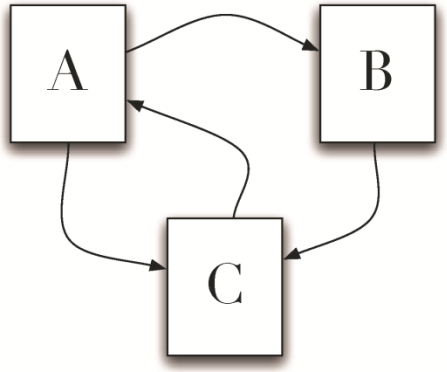
- Browse the Web using the following algorithm:
  - Choose a random number *r* between 0 and 1
  - If $r < \lambda$:
    - Go to a random page
  - If $r \geq \lambda$:
    - Click a link at random on the current page
  - Start again
- PageRank of a page is the probability that the "random surfer" will be looking at that page
  - links from popular pages will increase PageRank of pages they point to

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Dangling Links

- Random jump prevents getting stuck on pages that
  - do not have links
  - contains only links that no longer point to other pages
  - have links forming a loop
- Links that point to the first two types of pages are called *dangling links*
  - may also be links to pages that have not yet been crawled

PageRank



- PageRank (*PR*) of page C = *PR*(A)/2 + *PR*(B)/1
- More generally,

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

  - where B_u is the set of pages that point to u, and L_v is the number of outgoing links from page v (not counting duplicate links)

PageRank

- Don’t know PageRank values at start
- Assume equal values (1/3 in this case), then iterate:
  - first iteration: $PR(C) = 0.33/2 + 0.33 = 0.5$, $PR(A) = 0.33$, and $PR(B) = 0.17$
  - second: $PR(C) = 0.33/2 + 0.17 = 0.33$, $PR(A) = 0.5$, $PR(B) = 0.17$
  - third: $PR(C) = 0.42$, $PR(A) = 0.33$, $PR(B) = 0.25$
- Converges to $PR(C) = 0.4$, $PR(A) = 0.4$, and $PR(B) = 0.2$

PageRank

- Taking random page jump into account, 1/3 chance of going to any page when $r < \lambda$
- $PR(\text{C}) = \lambda/3 + (1 - \lambda) \cdot (PR(\text{A})/2 + PR(\text{B})/1)$
- More generally,

$$PR(u) = \frac{\lambda}{N} + (1 - \lambda). \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

  – where N is the number of pages, λ typically 0.15

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

```
1: procedure PageRank(G)
2:     ▷ G is the web graph, consisting of vertices (pages) and edges (links).
3:     (P, L) ← G                                  ▷ Split graph into pages and links
4:     I ← a vector of length |P|                  ▷ The current PageRank estimate
5:     R ← a vector of length |P|   ▷ The resulting better PageRank estimate
6:     for all entries I_i ∈ I do
7:         I_i ← 1/|P|                 ▷ Start with each page being equally likely
8:     end for
9:     while R has not converged do
10:        for all entries R_i ∈ R do
11:            R_i ← λ/|P| ▷ Each page has a λ/|P| chance of random selection
12:        end for
13:        for all pages p ∈ P do
14:            Q ← the set of pages such that (p, q) ∈ L and q ∈ P
15:            if |Q| > 0 then
16:                for all pages q ∈ Q do
17:                    R_q ← R_q + (1 − λ)I_p/|Q|      ▷ Probability I_p of being at
   page p
18:                end for
19:            else
20:                for all pages q ∈ P do
21:                    R_q ← R_q + (1 − λ)I_p/|P|
22:                end for
23:            end if
24:            I ← R                       ▷ Update our current PageRank estimate
25:        end for
26:    end while
27:    return R
28: end procedure
```

A PageRank Implementation

- Preliminaries:
  - 1) Extract links from the source text. You'll also want to extract the URL from each document in a separate file. Now you have all the links (source-destination pairs) and all the source documents
  - 2) Remove all links from the list that do not connect two documents in the corpus. The easiest way to do this is to sort all links by destination, then compare that against the corpus URLs list (also sorted)
  - 3) Create a new file I that contains a (url, pagerank) pair for each URL in the corpus. The initial PageRank value is 1/#D (#D = number of urls)
- At this point there are two interesting files:
  - [L] links (trimmed to contain only corpus links, sorted by source URL)
  - [I] URL/PageRank pairs, initialized to a constant

A PageRank Implementation

- Preliminaries - Link Extraction from .corpus file using Galago

  DocumentSplit -> IndexReaderSplitParser -> TagTokenizer

  split = new DocumentSplit (filename, filetype, new byte[0], new byte[0])

  index = new IndexReaderSplitParser (split)

  tokenizer = new.TagTokenizer ( )

  tokenizer.setProcessor (NullProcessor (Document.class))

  doc = index.nextDocument ( )

  tokenizer.process (doc)

  - doc.identifier contains the file's name
  - doc.tags now contains all tags
  - Links can be extracted by finding all tags with name "a"
  - Links should be processed so that they can be compared with some file name in the corpus

A PageRank Implementation

Iteration:

- Steps:
  1. Make a new output file, R.
  2. Read L and I in parallel (since they're all sorted by URL).
  3. For each unique source URL, determine whether it has any outgoing links:
  4. If not, add its current PageRank value to the sum: T (terminals).
  5. If it does have outgoing links, write (source_url, dest_url, $I_p/|Q|$), where I_p is the current PageRank value, $|Q|$ is the number of outgoing links, and dest_url is a link destination.
     Do this for all outgoing links. Write this to R.
  6. Sort R by destination URL.
  7. Scan R and I at the same time. The new value of R_p is:
     (1 - lambda) / #D (a fraction of the sum of all pages)
     plus: lambda * sum(T) / #D (the total effect from terminal pages),
     plus: lambda * all incoming mass from step 5. ()
  8. Check for convergence
  9. Write new R_p values to a new I file.

A PageRank Implementation

- Convergence check
  - Stopping criteria for this types of PR algorithm typically is of the form ||new - old|| < tau where new and old are the new and old PageRank vectors, respectively.
  - Tau is set depending on how much precision you need. Reasonable values include 0.1 or 0.01. If you want really fast, but inaccurate convergence, then you can use something like tau=1.
  - The setting of tau also depends on N (= number of documents in the collection), since ||new-old|| (for a fixed numerical precision) increases as N increases, so you can alternatively formulate your convergence criteria as ||new – old|| / N < tau.
  - Either the L1 or L2 norm can be used.

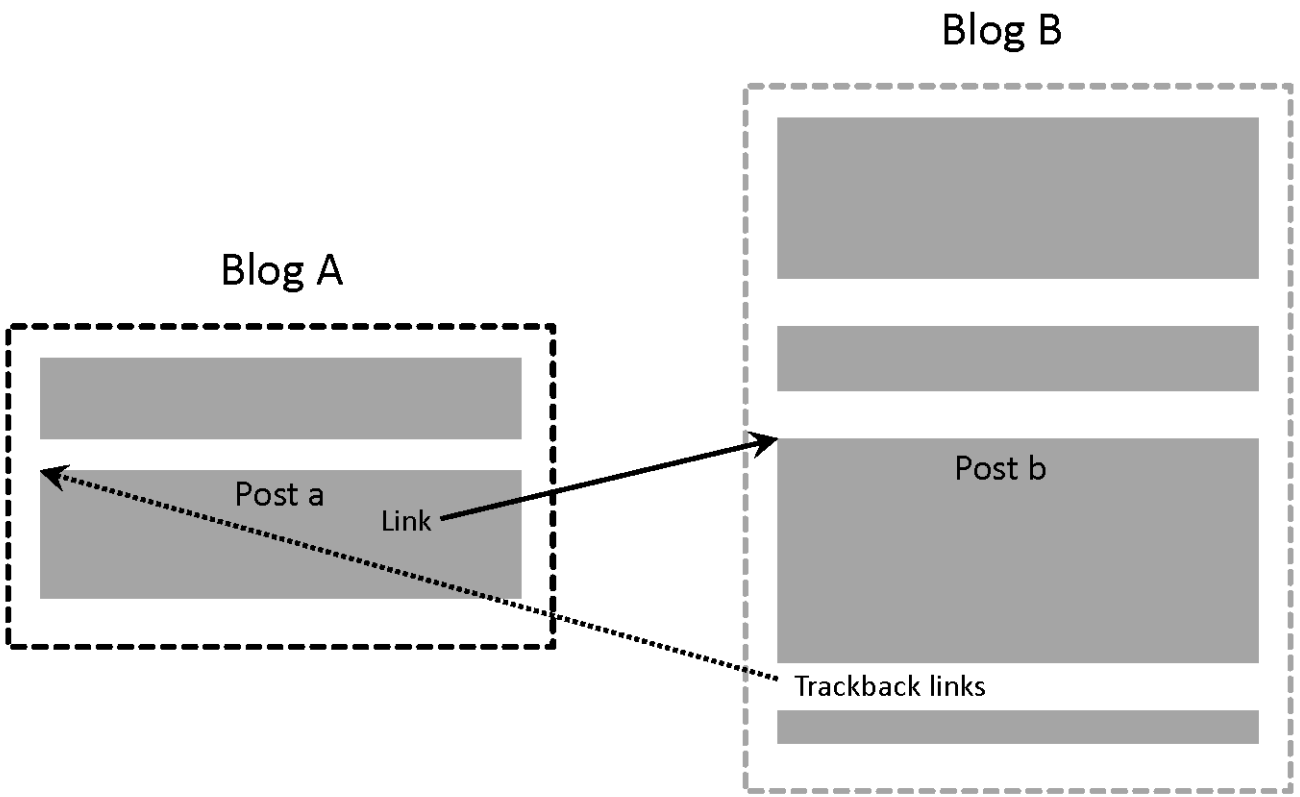
Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Link Quality

- Link quality is affected by spam and other factors
  - e.g., *link farms* to increase PageRank
  - *trackback links* in blogs can create loops
  - links from comments section of popular blogs
    - Blog services modify comment links to contain `rel=nofollow` attribute
    - e.g., “Come visit my <a `rel=nofollow` href="http://www.page.com">web page.”

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Trackback Links



Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Information Extraction (IE)

- Automatically extract structure from text
  - annotate document using tags to identify extracted structure
- *Named entity recognition (NER)*
  - identify words that refer to something of interest in a particular application
  - e.g., people, companies, locations, dates, product names, prices, etc.

Named Entity Recognition (NER)

Fred Smith, who lives at 10 Water Street, Springfield, MA, is a long-time collector of **tropical fish.**

<PersonName><GivenName>Fred</GivenName> <Sn>Smith</Sn> </PersonName>, who lives at <address><Street >10 Water Street</Street>, <City>Springfield</City>, <State>MA</State></address>, is a long-time collector of tropical fish.

- Example showing semantic annotation of text using XML tags
- Information extraction also includes document structure and more complex features such as *relationships* and *events*

Named Entity Recognition

- *Rule-based*
  - Uses *lexicons* (lists of words and phrases) that categorize names
    - e.g., locations, peoples' names, organizations, etc.
  - Rules also used to verify or find new entity names
    - e.g., "<number> <word> street" for addresses
    - "<street address>, <city>" or "in <city>" to verify city names
    - "<street address>, <city>, <state>" to find new cities
    - "<title> <name>" to find new names

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Named Entity Recognition

- Rules either developed manually by trial and error or using machine learning techniques
- S*tatistical*
  - uses a probabilistic model of the words in and around an entity
  - probabilities estimated using *training data* (manually annotated text)
  - Hidden Markov Model (HMM)
  - Conditional Random Field (CRF)

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Named Entity Recognition

- Accurate recognition requires about 1M words of training data (1,500 news stories)
  - may be more expensive than developing rules for some applications
- Both rule-based and statistical can achieve about 90% effectiveness for categories such as names, locations, organizations
  - others, such as product name, can be much worse

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Internationalization

- 2/3 of the Web is in English
- About 50% of Web users do not use English as their primary language
- Many (maybe most) search applications have to deal with multiple languages
  - *monolingual search*: search in one language, but with many possible languages
  - *cross-language search*: search in multiple languages at the same time

Internationalization

- Many aspects of search engines are language-neutral
- Major differences:
  - Text encoding (converting to Unicode)
  - Tokenizing (many languages have no word separators)
  - Stemming
- Cultural differences may also impact interface design and features provided

Chinese “Tokenizing”

1. Original text

旱灾在中国造成的影响

(the impact of droughts in China)

2. Word segmentation

旱灾　在　中国　造成　的　影响

drought　at　china　make　　impact

3. Bigrams

旱灾　灾在　在中　中国　国造

造成　成的　的影　影响

Source: Croft et al. (2008) Search Engines: Information Retrieval in Practice

Summary

- Social Network Analysis
- Link Mining
- Text and Web Mining

References

- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Elsevier
- Michael W. Berry and Jacob Kogan, Text Mining: Applications and Theory, 2010, Wiley
- Guandong Xu, Yanchun Zhang, Lin Li, Web Mining and Social Networking: Techniques and Applications, 2011, Springer
- Matthew A. Russell, Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites, 2011, O'Reilly Media
- Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2009, Springer
- Bruce Croft, Donald Metzler, and Trevor Strohman, Search Engines: Information Retrieval in Practice, 2008, Addison Wesley, http://www.search-engines-book.com/
- Jaideep Srivastava, Nishith Pathak, Sandeep Mane, and Muhammad A. Ahmad, Data Mining for Social Network Analysis, Tutorial at IEEE ICDM 2006, Hong Kong, 2006
- Sentinel Visualizer, http://www.fmsasg.com/SocialNetworkAnalysis/
- Text Mining, http://en.wikipedia.org/wiki/Text_mining