Data Warehousing
資料倉儲

Text Mining and Web Mining

1001DW10
MI4
Tue. 6,7 (13:10-15:00) B427

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.tku.edu.tw/myday/

2011-12-20

Syllabus

週次	日期	內容（Subject/Topics）
1	100/09/06	Introduction to Data Warehousing
2	100/09/13	Data Warehousing, Data Mining, and Business Intelligence
3	100/09/20	Data Preprocessing: Integration and the ETL process
4	100/09/27	Data Warehouse and OLAP Technology
5	100/10/04	Data Warehouse and OLAP Technology
6	100/10/11	Data Cube Computation and Data Generation
7	100/10/18	Data Cube Computation and Data Generation
8	100/10/25	Project Proposal
9	100/11/01	期中考試週

Syllabus

週次	日期	內容（Subject/Topics）
10	100/11/08	Association Analysis
11	100/11/15	Association Analysis
12	100/11/22	Classification and Prediction
13	100/11/29	Classification and Prediction
14	100/12/06	Cluster Analysis
15	100/12/13	Social Network Analysis and Link Mining
16	100/12/20	Text Mining and Web Mining
17	100/12/27	Project Presentation
18	101/01/03	期末考試週

Learning Objectives

- Describe text mining and understand the need for text mining
- Differentiate between text mining, Web mining and data mining
- Understand the different application areas for text mining
- Know the process of carrying out a text mining project
- Understand the different methods to introduce structure to text-based data

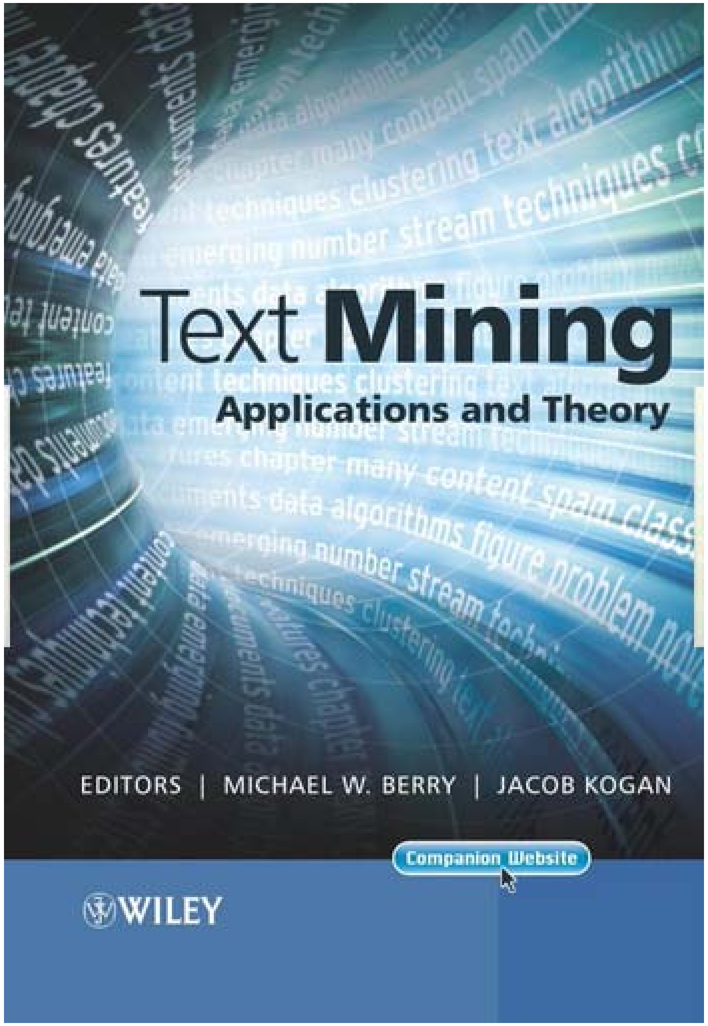
Learning Objectives

- Describe Web mining, its objectives, and its benefits
- Understand the three different branches of Web mining
  - Web content mining
  - Web structure mining
  - Web usage mining
- Understand the applications of these three mining paradigms

Text and Web Mining

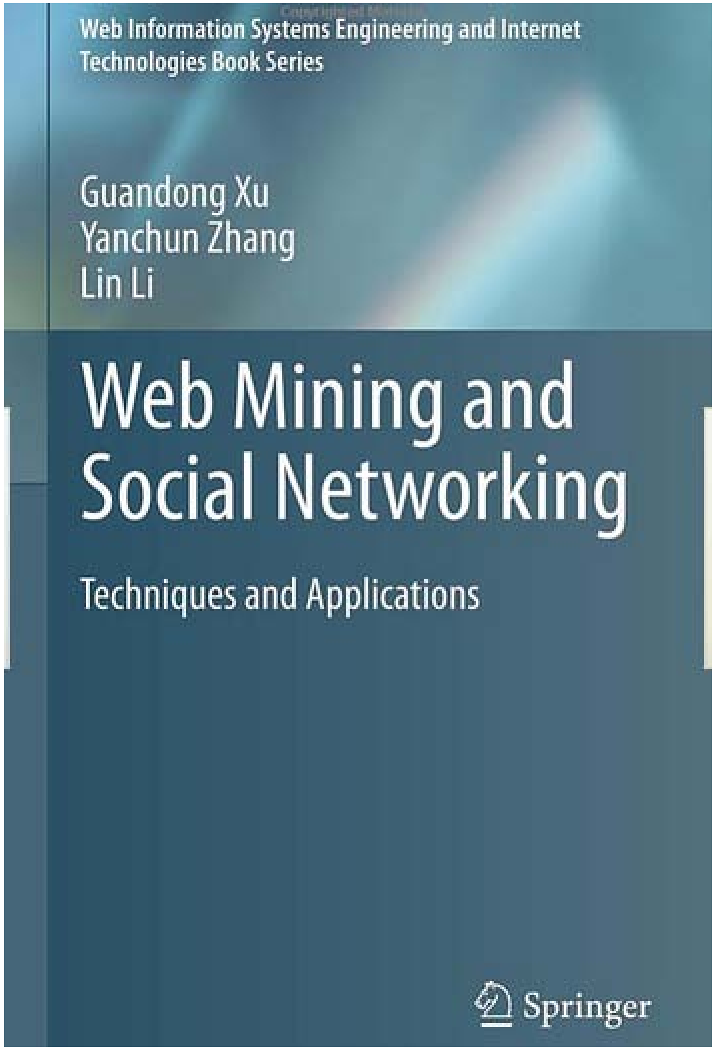
- Text Mining: Applications and Theory
- Web Mining and Social Networking
- Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites
- Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data
- Search Engines – Information Retrieval in Practice

Text Mining



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

Web Mining and Social Networking



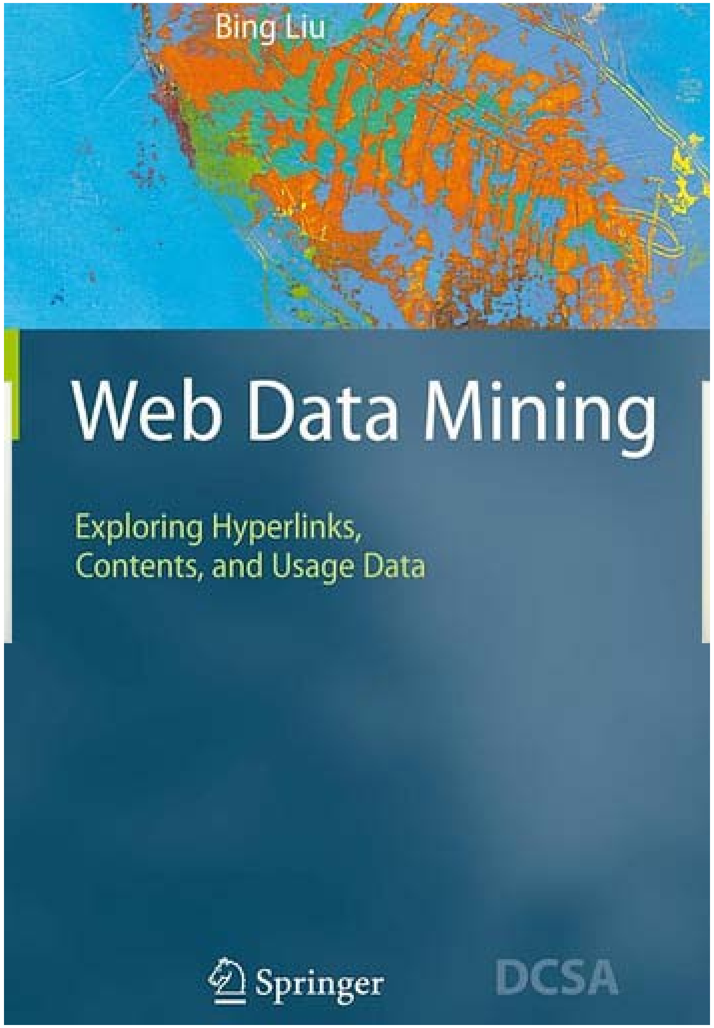
http://www.amazon.com/Web-Mining-Social-Networking-Applications/dp/1441977341

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

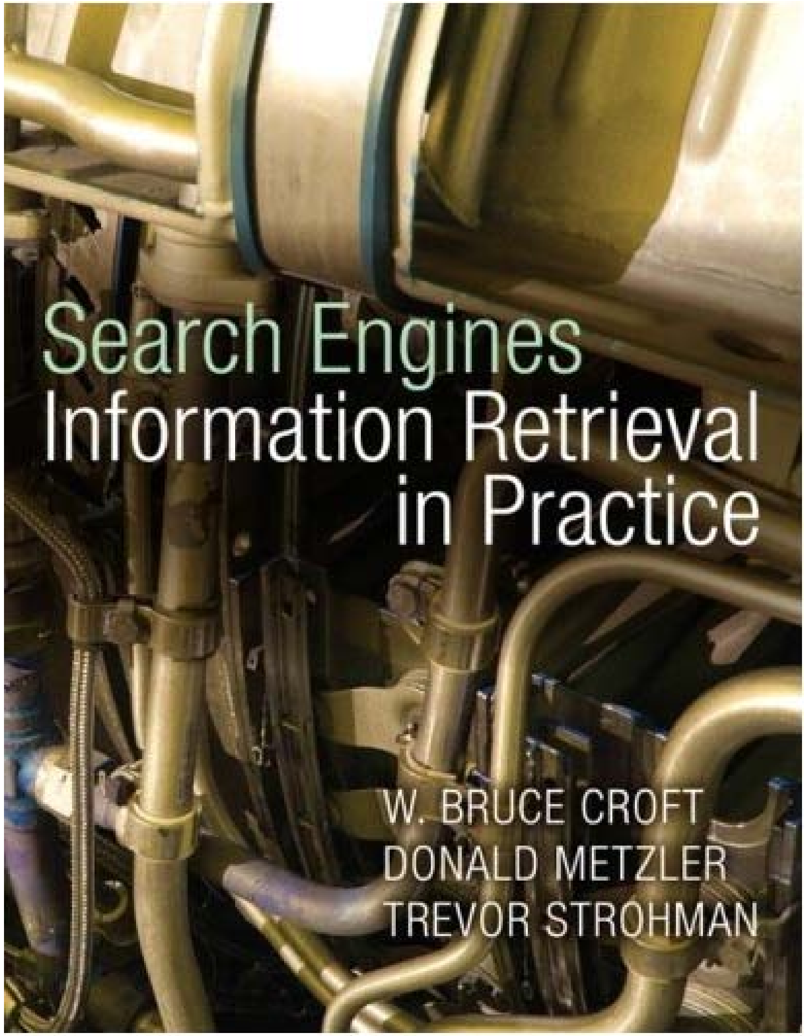
Web Data Mining:

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

Mining Text For Security…

Cluster 1

- (L) Kampala
- (L) Uganda
- (P) Yoweri Museveni
- (L) Sudan
- (L) Khartoum
- (L) Southern Sudan

Cluster 2

- (P) Timothy McVeigh
- (P) Oklahoma City
- (P) Terry Nichols

Cluster 3

- (E) election
- (P) Norodom Ranariddh
- (P) Norodom Sihanouk
- (L) Bangkok
- (L) Cambodia
- (L) Phnom Penh
- (L) Thailand
- (P) Hun Sen
- (O) Khmer Rouge
- (P) Pol Pot

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Concepts

- 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months
- Tapping into these information sources is not an option, but a need to stay competitive
- Answer: text mining
  - A semi-automated process of extracting knowledge from unstructured data sources
  - a.k.a. text data mining or knowledge discovery in textual databases

Data Mining versus Text Mining

- Both seek for novel and useful patterns
- Both are semi-automated processes
- Difference is the nature of the data:
  - Structured versus unstructured data
  - Structured data: in databases
  - Unstructured data: Word documents, PDF files, text excerpts, XML files, and so on
- Text mining – first, impose structure to the data, then mine the structured data

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Concepts

- Benefits of text mining are obvious especially in text-rich data environments
  - e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.
- Electronic communization records (e.g., Email)
  - Spam filtering
  - Email prioritization and categorization
  - Automatic response generation

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Application Area

- Information extraction
- Topic tracking
- Summarization
- Categorization
- Clustering
- Concept linking
- Question answering

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Terminology

- Unstructured or semistructured data
- Corpus (and corpora)
- Terms
- Concepts
- Stemming
- Stop words (and include words)
- Synonyms (and polysemes)
- Tokenizing

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Terminology

- Term dictionary
- Word frequency
- Part-of-speech tagging (POS)
- Morphology
- Term-by-document matrix (TDM)
  - Occurrence matrix
- Singular Value Decomposition (SVD)
  - Latent Semantic Indexing (LSI)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining for Patent Analysis

- What is a patent?
  - "exclusive rights granted by a country to an inventor for a limited period of time in exchange for a disclosure of an invention"
- How do we do patent analysis (PA)?
- Why do we need to do PA?
  - What are the benefits?
  - What are the challenges?
- How does text mining help in PA?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Natural Language Processing (NLP)

- Structuring a collection of text
  - Old approach: bag-of-words
  - New approach: natural language processing
- NLP is …
  - a very important concept in text mining
  - a subfield of artificial intelligence and computational linguistics
  - the studies of "understanding" the natural human language
- Syntax versus semantics based text mining

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Natural Language Processing (NLP)

- What is “Understanding” ?
  - Human understands, what about computers?
  - Natural language is vague, context driven
  - True understanding requires extensive knowledge of a topic

  - Can/will computers ever understand natural language the same/accurate way we do?

Natural Language Processing (NLP)

- Challenges in NLP
  - Part-of-speech tagging
  - Text segmentation
  - Word sense disambiguation
  - Syntax ambiguity
  - Imperfect or irregular input
  - Speech acts
- Dream of AI community
  - to have algorithms that are capable of automatically reading and obtaining knowledge from text

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Natural Language Processing (NLP)

- WordNet
  - A laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets
  - A major resource for NLP
  - Need automation to be completed
- Sentiment Analysis
  - A technique used to detect favorable and unfavorable opinions toward specific products and services
  - CRM application

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

NLP Task Categories

- Information retrieval   (IR)
- Information extraction (IE)
- Named-entity recognition (NER)
- Question answering (QA)
- Automatic summarization
- Natural language generation and understanding (NLU)
- Machine translation (ML)
- Foreign language reading and writing
- Speech recognition
- Text proofing
- Optical character recognition (OCR)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Applications

- Marketing applications
  - Enables better CRM
- Security applications
  - ECHELON, OASIS
  - Deception detection (…)
- Medicine and biology
  - Literature-based gene identification (…)
- Academic applications
  - Research stream analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Applications

- Application Case: Mining for Lies
- Deception detection
  - A difficult problem
  - If detection is limited to only text, then the problem is even more difficult
- The study
  - analyzed text based testimonies of person of interests at military bases
  - used only text-based features (cues)

Text Mining Applications

- Application Case: Mining for Lies

Text Mining Applications

- Application Case: Mining for Lies

Category	Example Cues
Quantity	Verb count, noun-phrase count, ...
Complexity	Avg. no of clauses, sentence length, …
Uncertainty	Modifiers, modal verbs, ...
Nonimmediacy	Passive voice, objectification, ...
Expressivity	Emotiveness
Diversity	Lexical diversity, redundancy, ...
Informality	Typographical error ratio
Specificity	Spatiotemporal, perceptual information …
Affect	Positive affect, negative affect, etc.

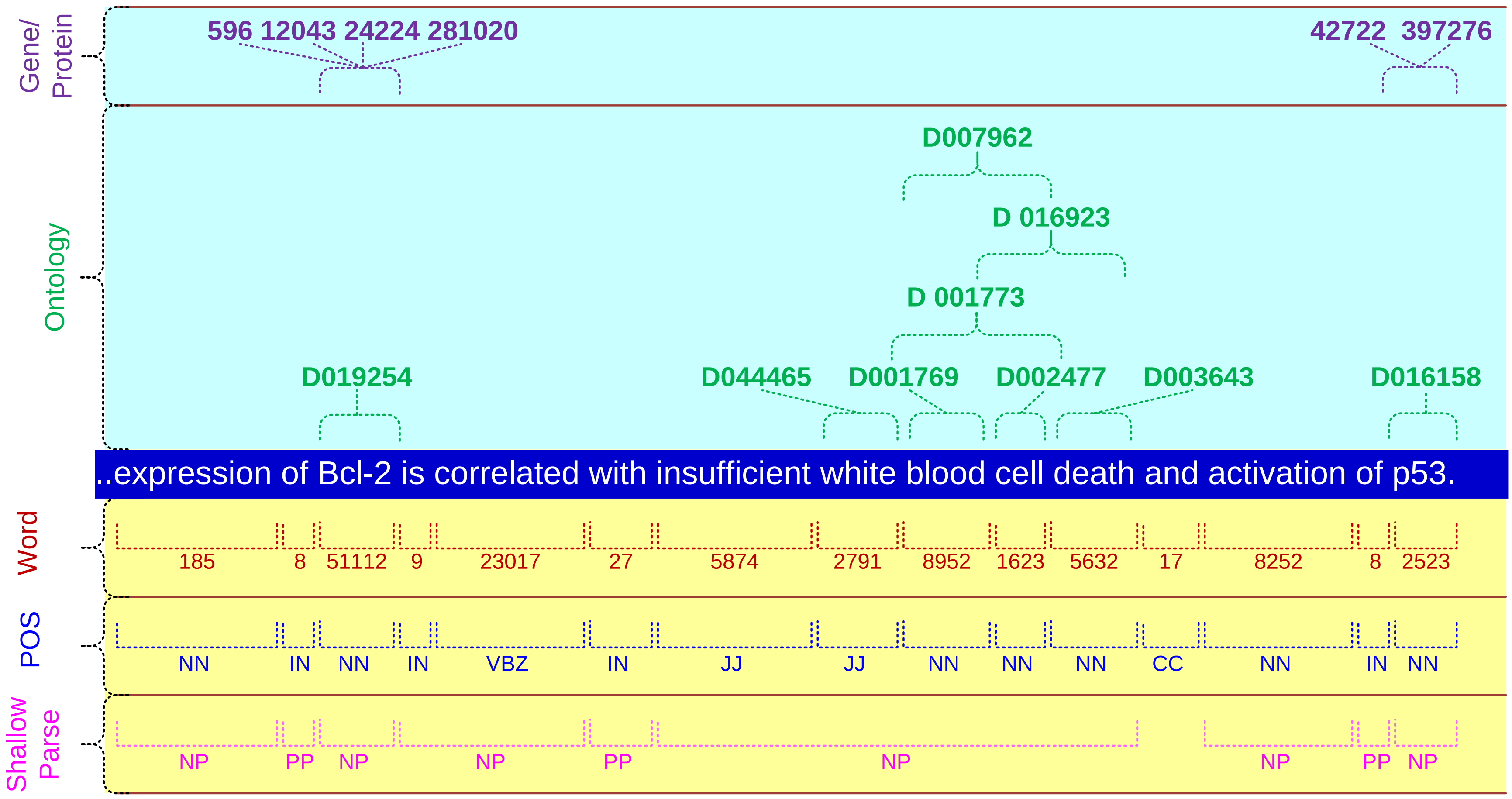
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Applications

- Application Case: Mining for Lies
  - 371 usable statements are generated
  - 31 features are used
  - Different feature selection methods used
  - 10-fold cross validation is used
  - Results (overall % accuracy)
    - Logistic regression 67.28
    - Decision trees 71.60
    - Neural networks 73.46

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

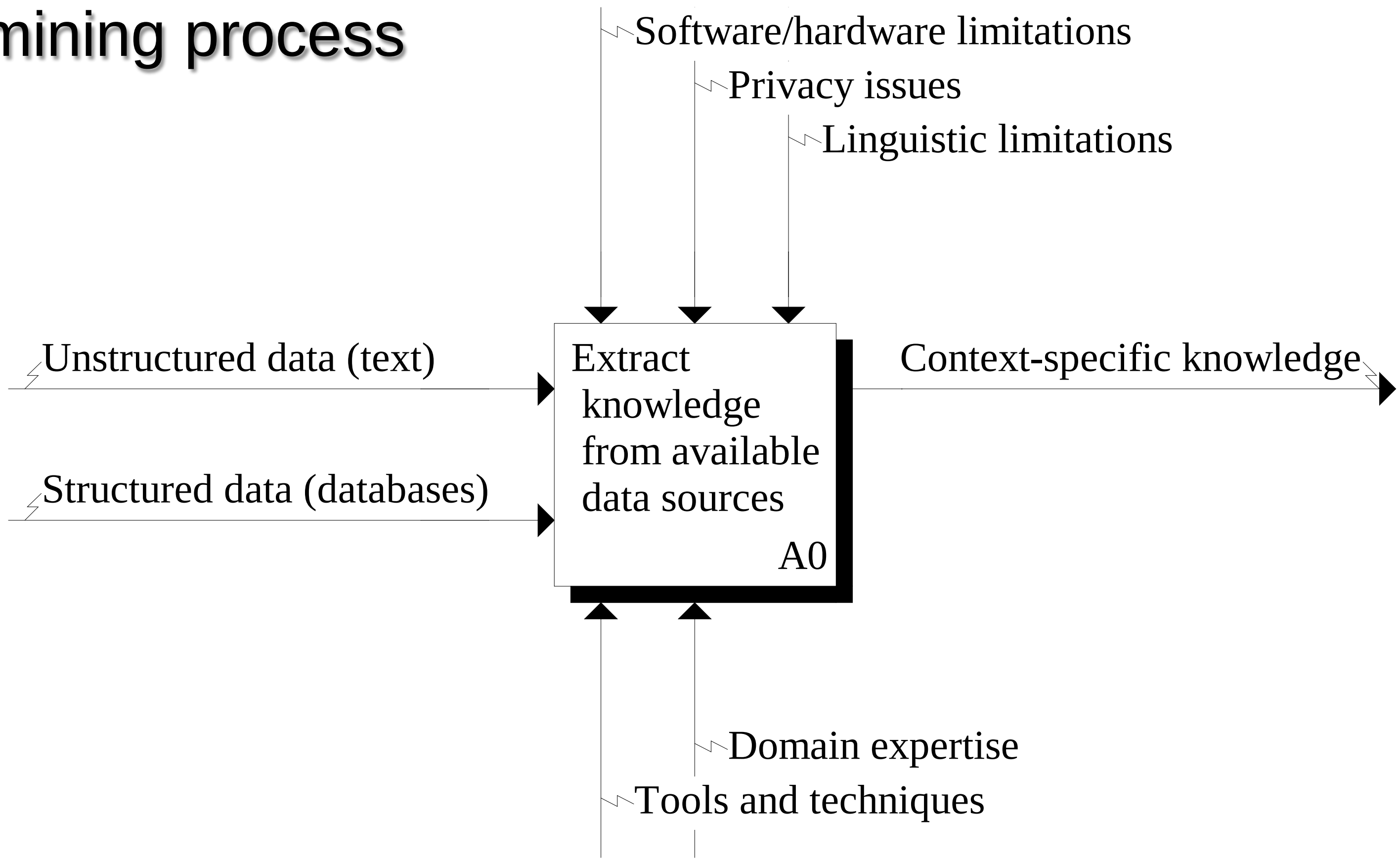
Text Mining Applications
(gene/protein interaction identification)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

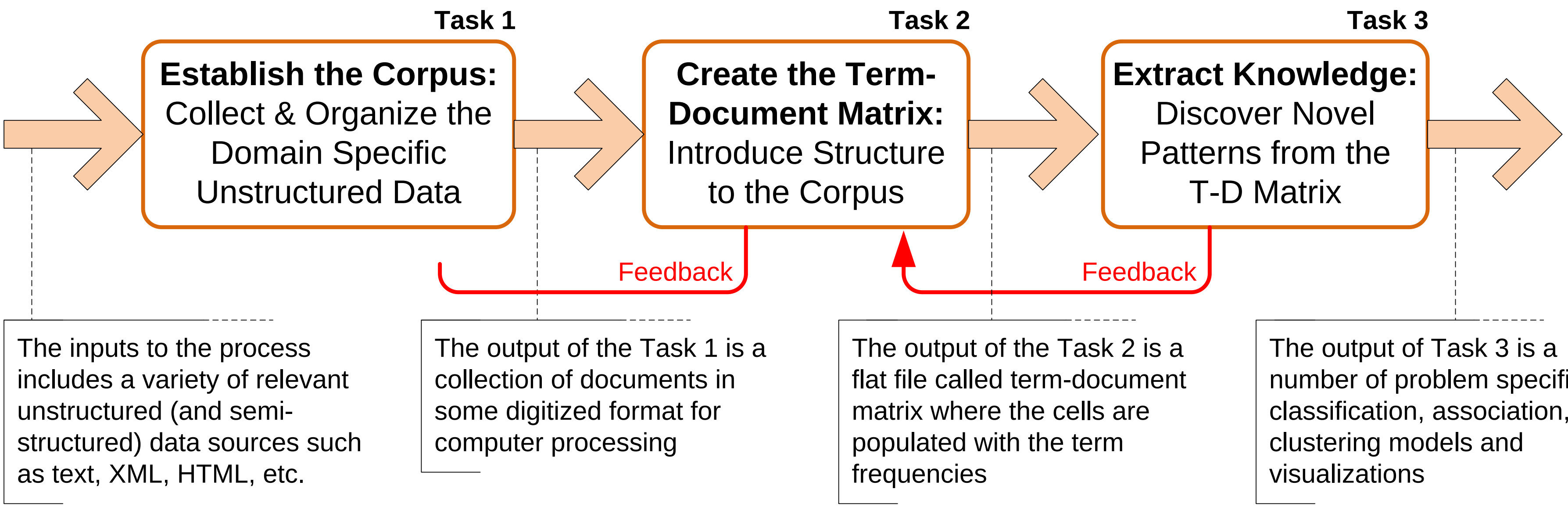
Text Mining Process

Context diagram for the text mining process



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Process



The three-step text mining process

Text Mining Process

- Step 1: Establish the corpus
  - Collect all relevant unstructured data (e.g., textual documents, XML files, emails, Web pages, short notes, voice recordings…)
  - Digitize, standardize the collection (e.g., all in ASCII text files)
  - Place the collection in a common place (e.g., in a flat file, or in a directory as separate files)

Text Mining Process

- Step 2: Create the Term–by–Document Matrix

Documents \ Terms	investment risk	project management	software engineering	development	SAP	...
Document 1	1			1		
Document 2		1				
Document 3			3		1	
Document 4		1				
Document 5			2	1		
Document 6	1			1		
...						

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Process

- Step 2: Create the Term–by–Document Matrix (TDM), cont.
  - – Should all terms be included?
    - Stop words, include words
    - Synonyms, homonyms
    - Stemming
  - – What is the best representation of the indices (values in cells)?
    - Row counts; binary frequencies; log frequencies;
    - Inverse document frequency

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Process

- Step 2: Create the Term–by–Document Matrix (TDM), cont.
  - TDM is a sparse matrix. How can we reduce the dimensionality of the TDM?
    - Manual - a domain expert goes through it
    - Eliminate terms with very few occurrences in very few documents (?)
    - Transform the matrix using singular value decomposition (SVD)
    - SVD is similar to principle component analysis

Text Mining Process

- Step 3: Extract patterns/knowledge
  - Classification (text categorization)
  - Clustering (natural groupings of text)
    - Improve search recall
    - Improve search precision
    - Scatter/gather
    - Query-specific clustering
  - Association
  - Trend Analysis (…)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Application
(research trend identification in literature)

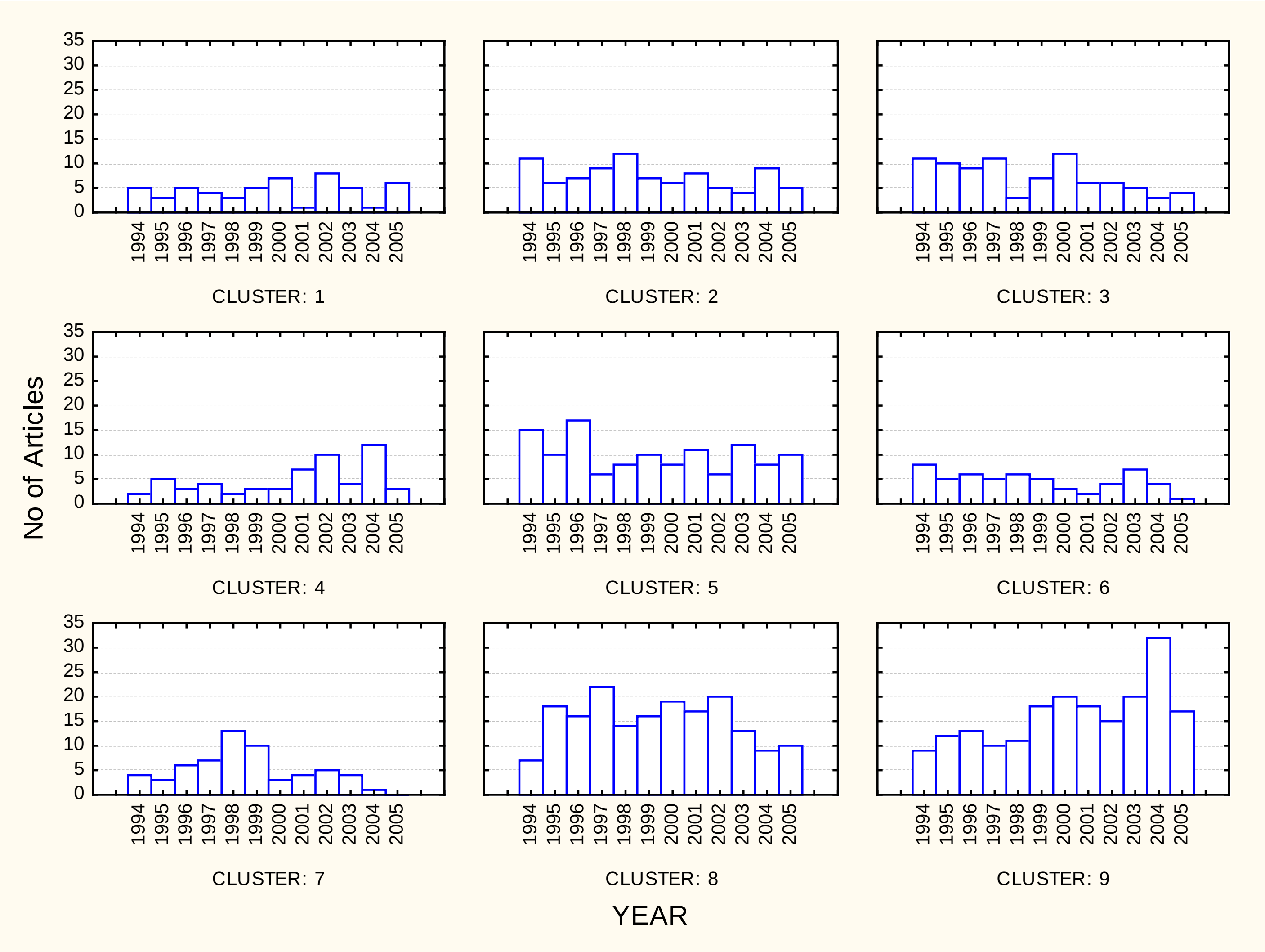
- Mining the published IS literature
  - MIS Quarterly (MISQ)
  - Journal of MIS (JMIS)
  - Information Systems Research (ISR)

  - Covers 12-year period (1994-2005)
  - 901 papers are included in the study
  - Only the paper abstracts are used
  - 9 clusters are generated for further analysis

Text Mining Application
(research trend identification in literature)

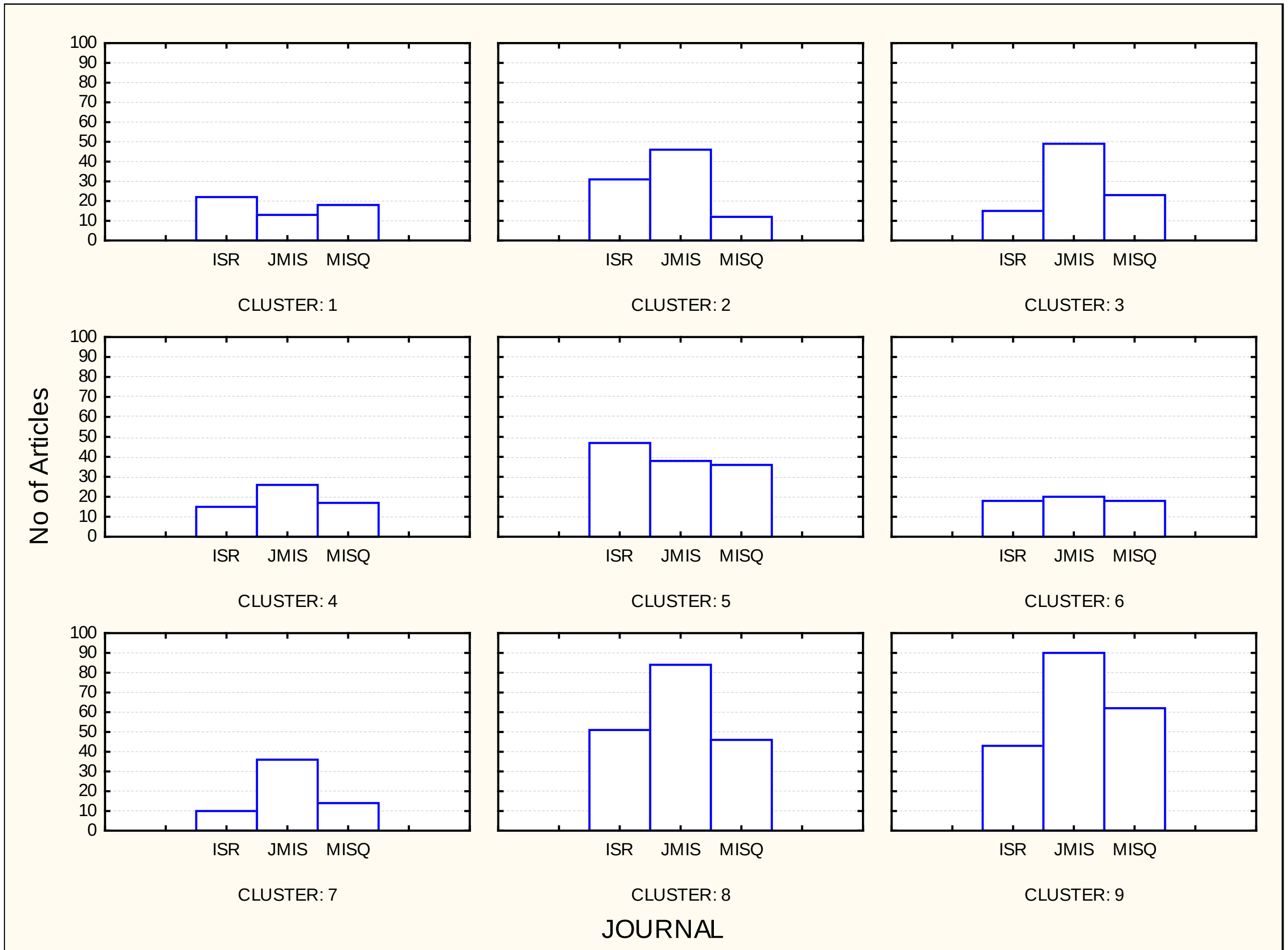
Journal	Year	Author(s)	Title	Vol/No	Pages	Keywords	Abstract
MISQ	2005	A. Malhotra, S. Gosain and O. A. El Sawy	Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation	29/1	145-187	knowledge management supply chain absorptive capacity interorganizational information systems configuration approaches	The need for continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganizational partner ships for sharing
ISR	1999	D. Robey and M. C. Boudreau	Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications	2-Oct	167-185	organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication mis implementation culture systems	Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory
JMIS	2001	R. Aron and E. K. Clemons	Achieving the optimal balance between investment in quality and investment in self-promotion for information products	18/2	65-88	information products internet advertising product positioning signaling signaling games	When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extent to which the advertised features of
…	…	…	…	…	…	…	…

Text Mining Application
(research trend identification in literature)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Application
(research trend identification in literature)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Tools

- Commercial Software Tools
  - SPSS PASW Text Miner
  - SAS Enterprise Miner
  - Statistica Data Miner
  - ClearForest, …
- Free Software Tools
  - RapidMiner
  - GATE
  - Spy-EM, …

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

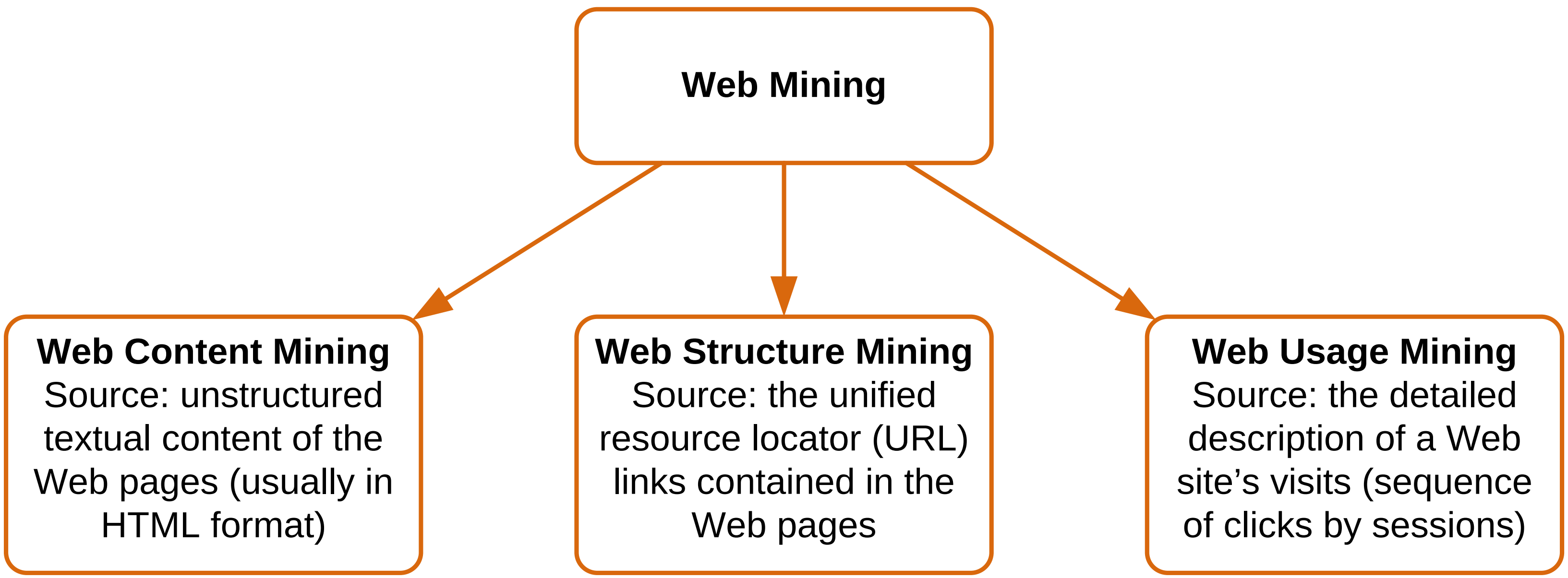
Web Mining Overview

- Web is the largest repository of data
- Data is in HTML, XML, text format
- Challenges (of processing Web data)
  - The Web is too big for effective data mining
  - The Web is too complex
  - The Web is too dynamic
  - The Web is not specific to a domain
  - The Web has everything
- Opportunities and challenges are great!

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Web Mining

- Web mining (or Web data mining) is the process of discovering intrinsic relationships from Web data (textual, linkage, or usage)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Web Content/Structure Mining

- Mining of the textual content on the Web
- Data collection via Web crawlers

- Web pages include hyperlinks
  - Authoritative pages
  - Hubs
  - hyperlink-induced topic search (HITS) alg

Web Usage Mining

- Extraction of information from data generated through Web page visits and transactions…
  - data stored in server access logs, referrer logs, agent logs, and client-side cookies
  - user characteristics and usage profiles
  - metadata, such as page attributes, content attributes, and usage data
- Clickstream data
- Clickstream analysis

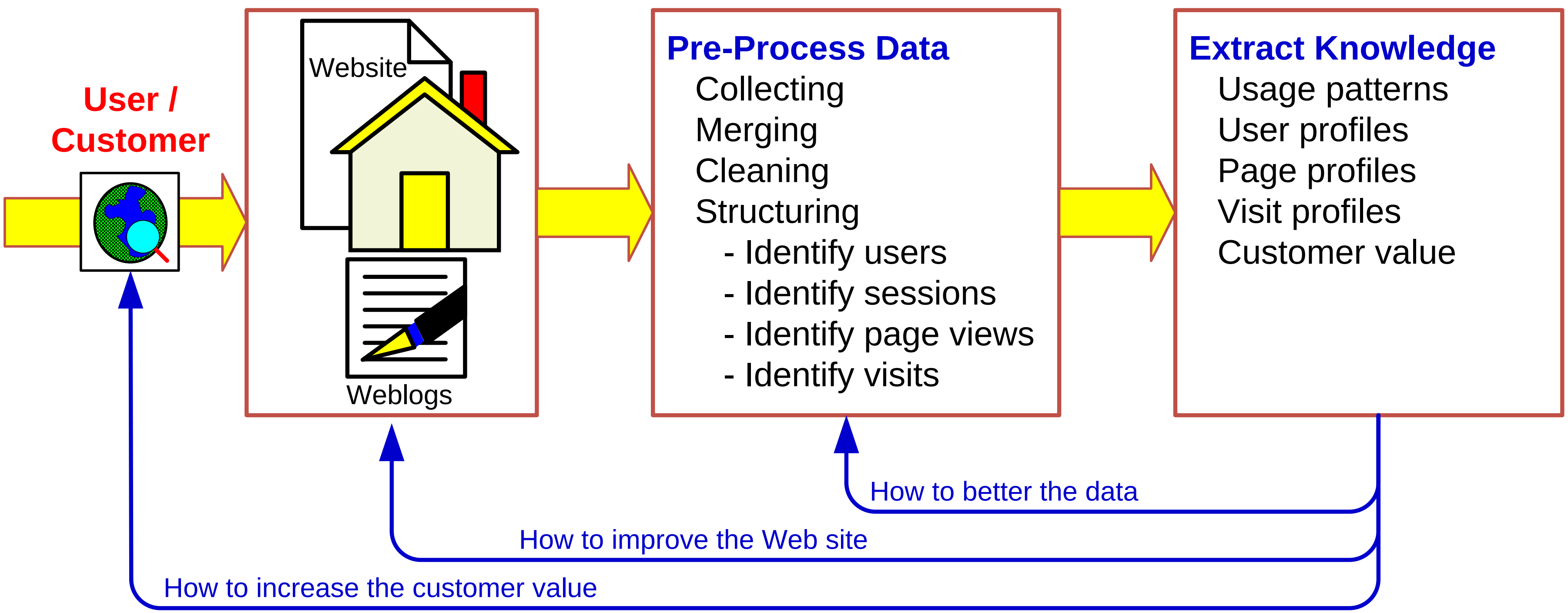
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Web Usage Mining

- Web usage mining applications
  - Determine the lifetime value of clients
  - Design cross-marketing strategies across products.
  - Evaluate promotional campaigns
  - Target electronic ads and coupons at user groups based on user access patterns
  - Predict user behavior based on previously learned rules and users' profiles
  - Present dynamic information to users based on their interests and profiles…

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

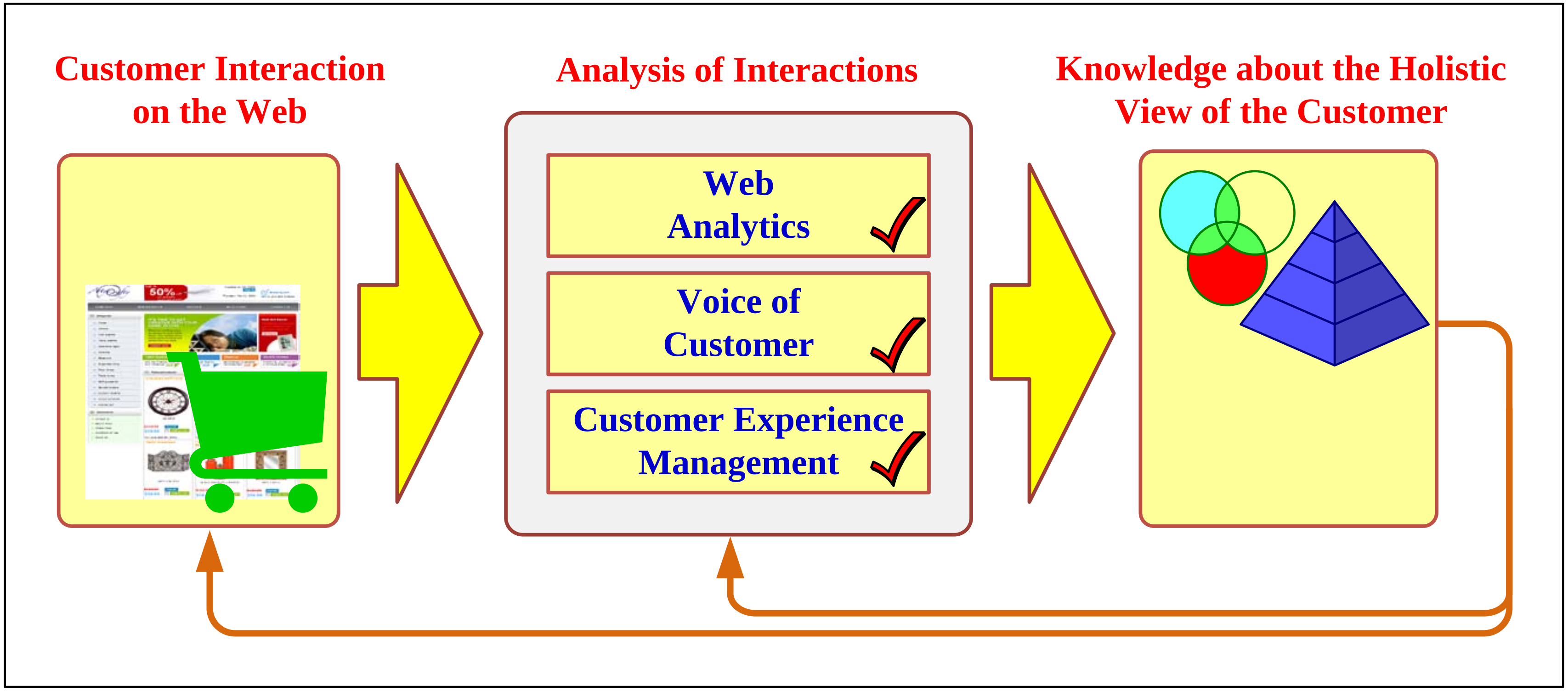
Web Usage Mining
(clickstream analysis)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Web Mining Success Stories

- Amazon.com, Ask.com, Scholastic.com, …
- Website Optimization Ecosystem



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Web Mining Tools

Product Name	URL
Angoss Knowledge WebMiner	angoss.com
ClickTracks	clicktracks.com
LiveStats from DeepMetrix	deepmetrix.com
Megaputer WebAnalyst	megaputer.com
MicroStrategy Web Traffic Analysis	microstrategy.com
SAS Web Analytics	sas.com
SPSS Web Mining for Clementine	spss.com
WebTrends	webtrends.com
XML Miner	scientio.com

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

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

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

Text Statistics

- Huge variety of words used in text <u>but</u>
- 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

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

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

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

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

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

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

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

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

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

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

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)

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

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*
kita**b**i	*my book*
al**kit**a**b**	*the book*
kita**b**uki	*your book* (f)
kita**b**uka	*your book (m)*
kita**b**uhu	*his book*
ka**t**a**b**a	*to write*
ma**kt**a**b**a	*library, bookstore*
ma**kt**a**b**	*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

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

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

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

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

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

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 $2 \leq n \leq 5$

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.

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

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>
```

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

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

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

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

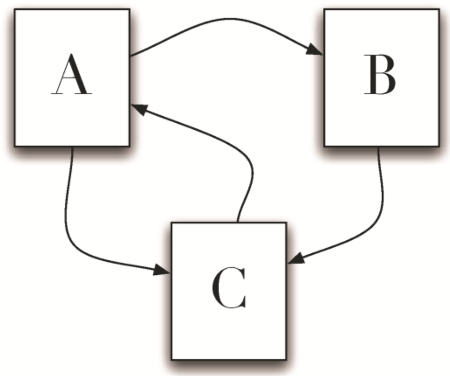
- Browse the Web using the following algorithm:
  - Choose a random number *r* between 0 and 1
  - If $r < \lambda$:
    - Go to a random page
  - If $r \geq \lambda$:
    - Click a link at random on the current page
  - Start again
- PageRank of a page is the probability that the "random surfer" will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

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

PageRank

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

```
procedure PageRank(G)
        ▷ G is the web graph, consisting of vertices (pages) and edges (links).
    (P, L) ← G                              ▷ Split graph into pages and links
    I ← a vector of length |P|              ▷ The current PageRank estimate
    R ← a vector of length |P|     ▷ The resulting better PageRank estimate
    for all entries I_i ∈ I do
        I_i ← 1/|P|                  ▷ Start with each page being equally likely
    end for
    while R has not converged do
        for all entries R_i ∈ R do
            R_i ← λ/|P| ▷ Each page has a λ/|P| chance of random selection
        end for
        for all pages p ∈ P do
            Q ← the set of pages such that (p, q) ∈ L and q ∈ P
            if |Q| > 0 then
                for all pages q ∈ Q do
                    R_q ← R_q + (1 − λ)I_p/|Q|      ▷ Probability I_p of being at page p
                end for
            else
                for all pages q ∈ P do
                    R_q ← R_q + (1 − λ)I_p/|P|
                end for
            end if
            I ← R                    ▷ Update our current PageRank estimate
        end for
    end while
    return R
end procedure
```

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

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

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

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

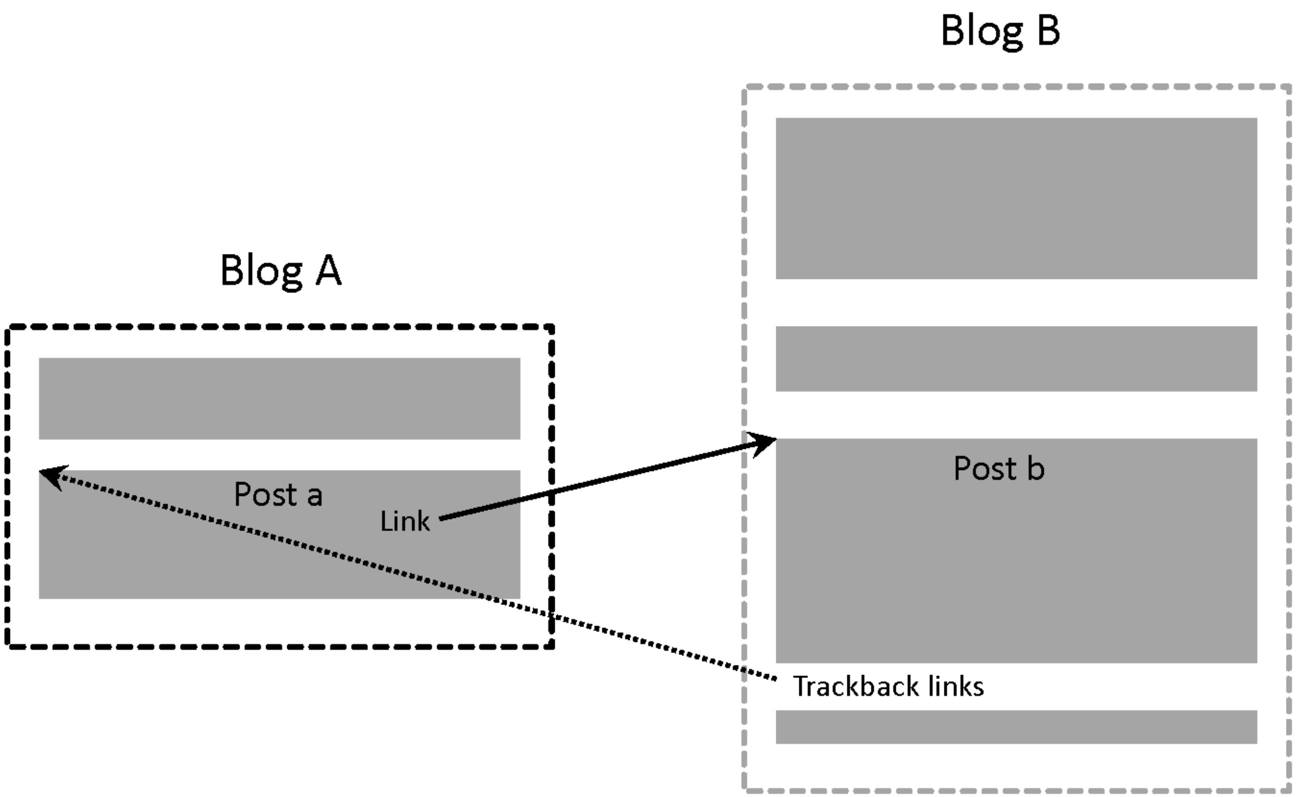
- Convergence check
  - Stopping criteria for this types of PR algorithm typically is of the form ||new - old|| < tau where new and old are the new and old PageRank vectors, respectively.
  - Tau is set depending on how much precision you need. Reasonable values include 0.1 or 0.01. If you want really 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.

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.

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

Named Entity Recognition (NER)

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

<p ><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.</p>

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

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

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)

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

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

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

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

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

- Text Mining
- Web Mining

References

- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Elsevier
- Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, Ninth Edition, 2011, Pearson.
- 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