# Data Mining
# 資料探勘

## 文字探勘與網頁探勘
## (Text and Web Mining)

1002DM05
MI4
Thu. 9,10 (16:10-18:00) B513

**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/
2012-05-03

# 課程大綱（Syllabus)

| 週次 | 日期 | 內容（Subject/Topics） |
|---|---|---|
| 1 | 101/02/16 | 資料探勘導論 (Introduction to Data Mining) |
| 2 | 101/02/23 | 關連分析 (Association Analysis) |
| 3 | 101/03/01 | 分類與預測 (Classification and Prediction) |
| 4 | 101/03/08 | 分群分析 (Cluster Analysis) |
| 5 | 101/03/15 | 個案分析與實作一 (分群分析)：<br>Banking Segmentation (Cluster Analysis – KMeans) |
| 6 | 101/03/22 | 個案分析與實作二 (關連分析)：<br>Web Site Usage Associations ( Association Analysis) |
| 7 | 101/03/29 | 期中報告 (Midterm Presentation) |
| 8 | 101/04/05 | 教學行政觀摩日 (--No Class--) |

# 課程大綱 (Syllabus)

週次　日期　內容（Subject/Topics）

9　101/04/12　個案分析與實作三 (決策樹、模型評估)：Enrollment Management Case Study (Decision Tree, Model Evaluation)

10　101/04/19　期中考試週

11　101/04/26　個案分析與實作四 (迴歸分析、類神經網路)：Credit Risk Case Study (Regression Analysis, Artificial Neural Network)

12　101/05/03　文字探勘與網頁探勘 (Text and Web Mining)

13　101/05/10　社會網路分析、意見分析 (Social Network Analysis, Opinion Mining)

14　101/05/17　期末專題報告 (Term Project Presentation)

15　101/05/24　畢業考試週

# 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

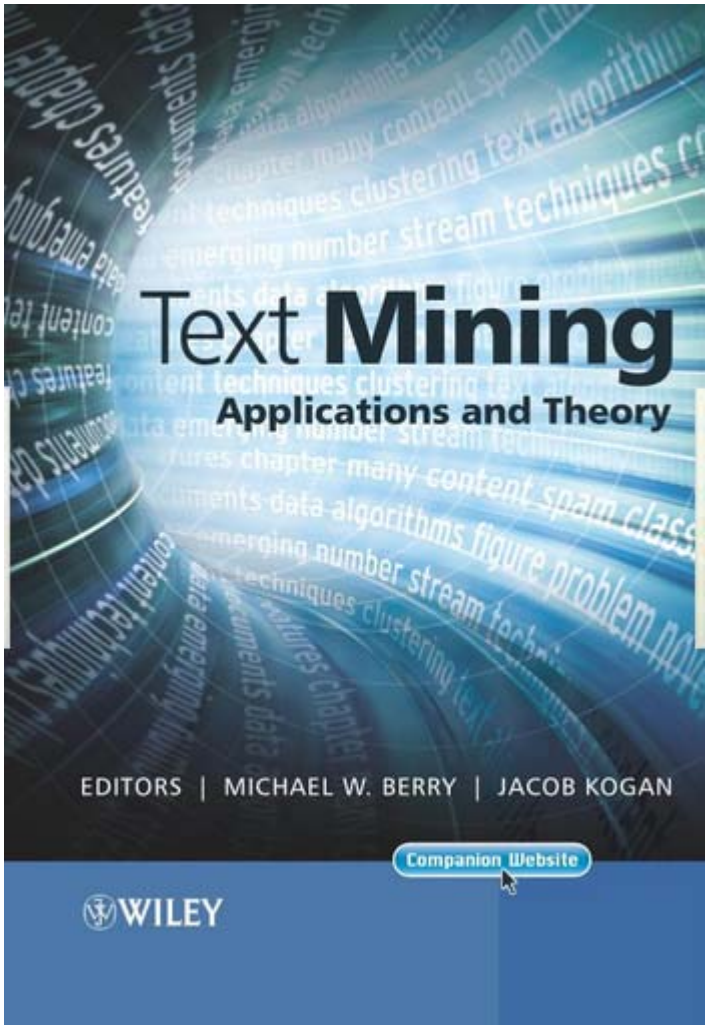
# Learning Objectives

- 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

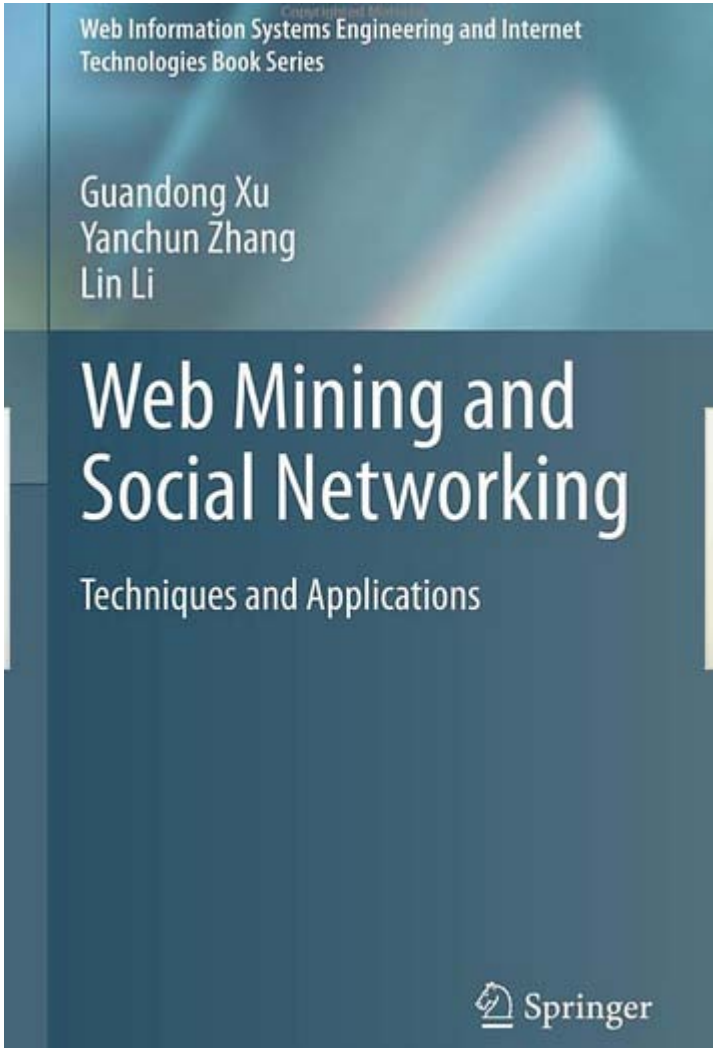
- Text Mining: Applications and Theory
- Web Mining and Social Networking
- Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites
- Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data
- Search Engines – Information Retrieval in Practice

# Text Mining



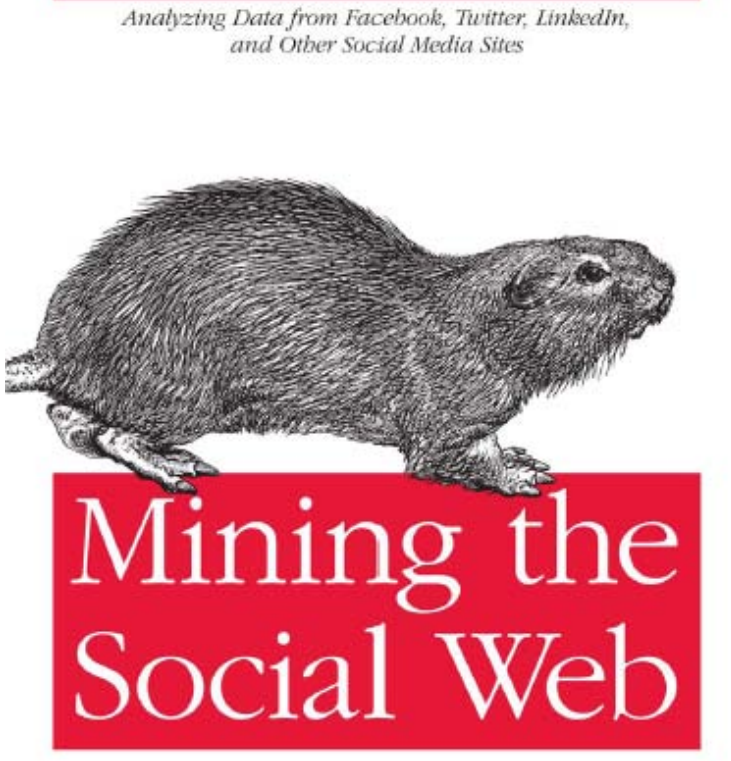

http://www.amazon.com/Text-Mining-Applications-Michael-Berry/dp/0470749822/

# Web Mining and Social Networking



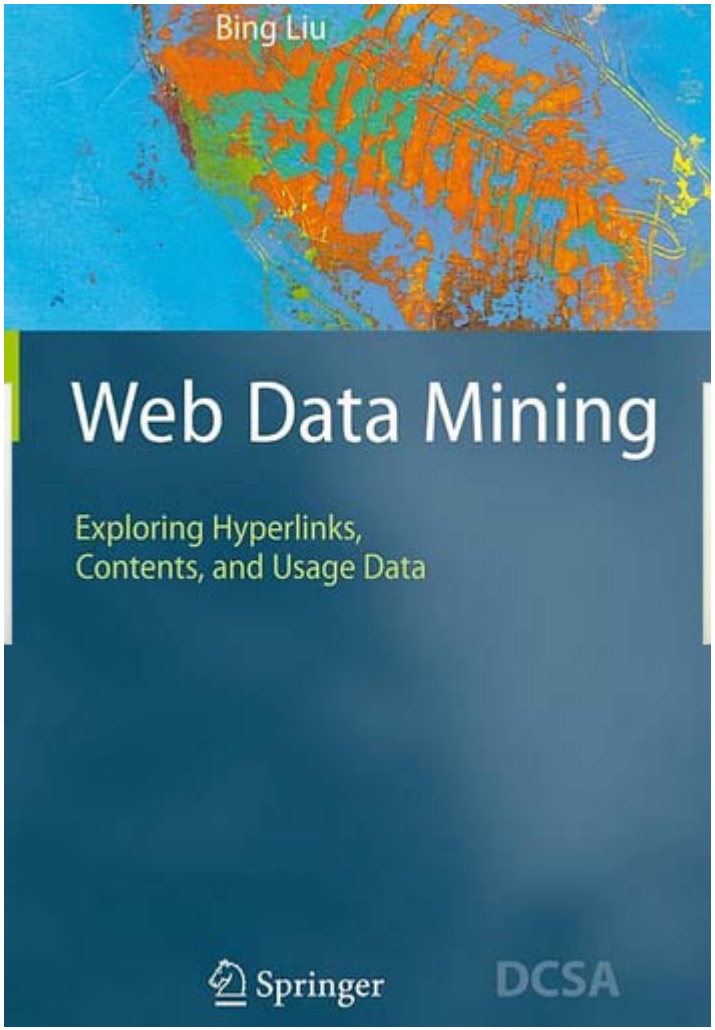

http://www.amazon.com/Web-Mining-Social-Networking-Applications/dp/1441977341

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



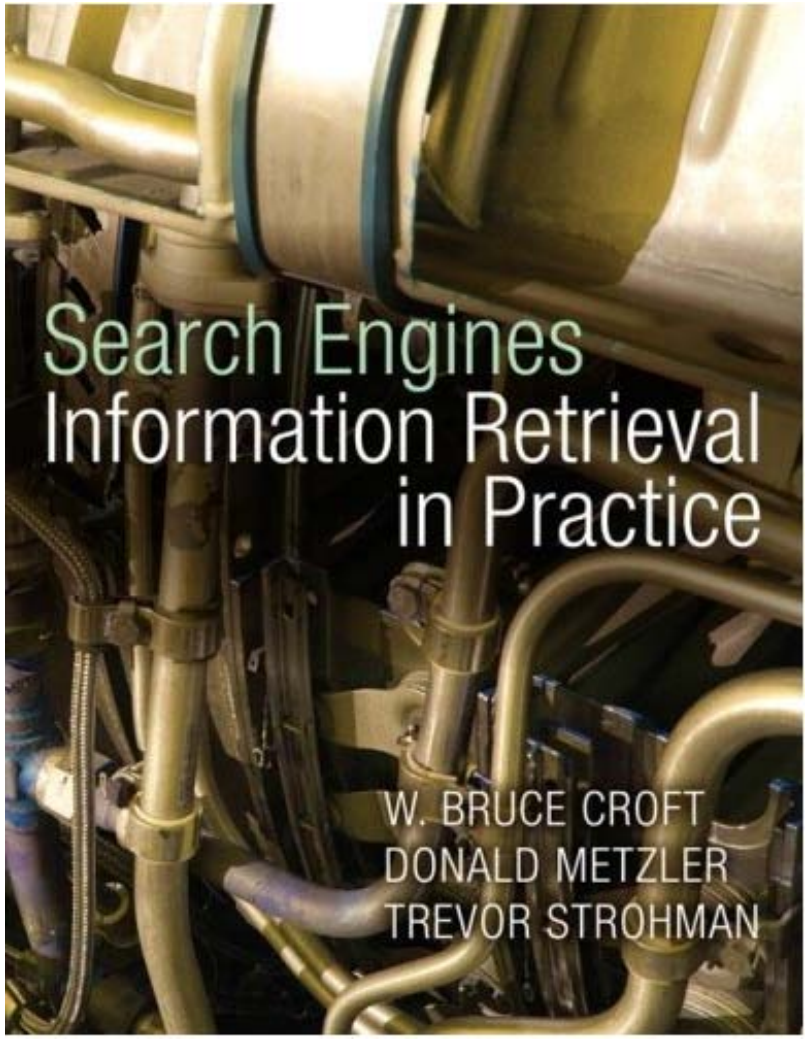

http://www.amazon.com/Mining-Social-Web-Analyzing-Facebook/dp/1449388345

# 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.

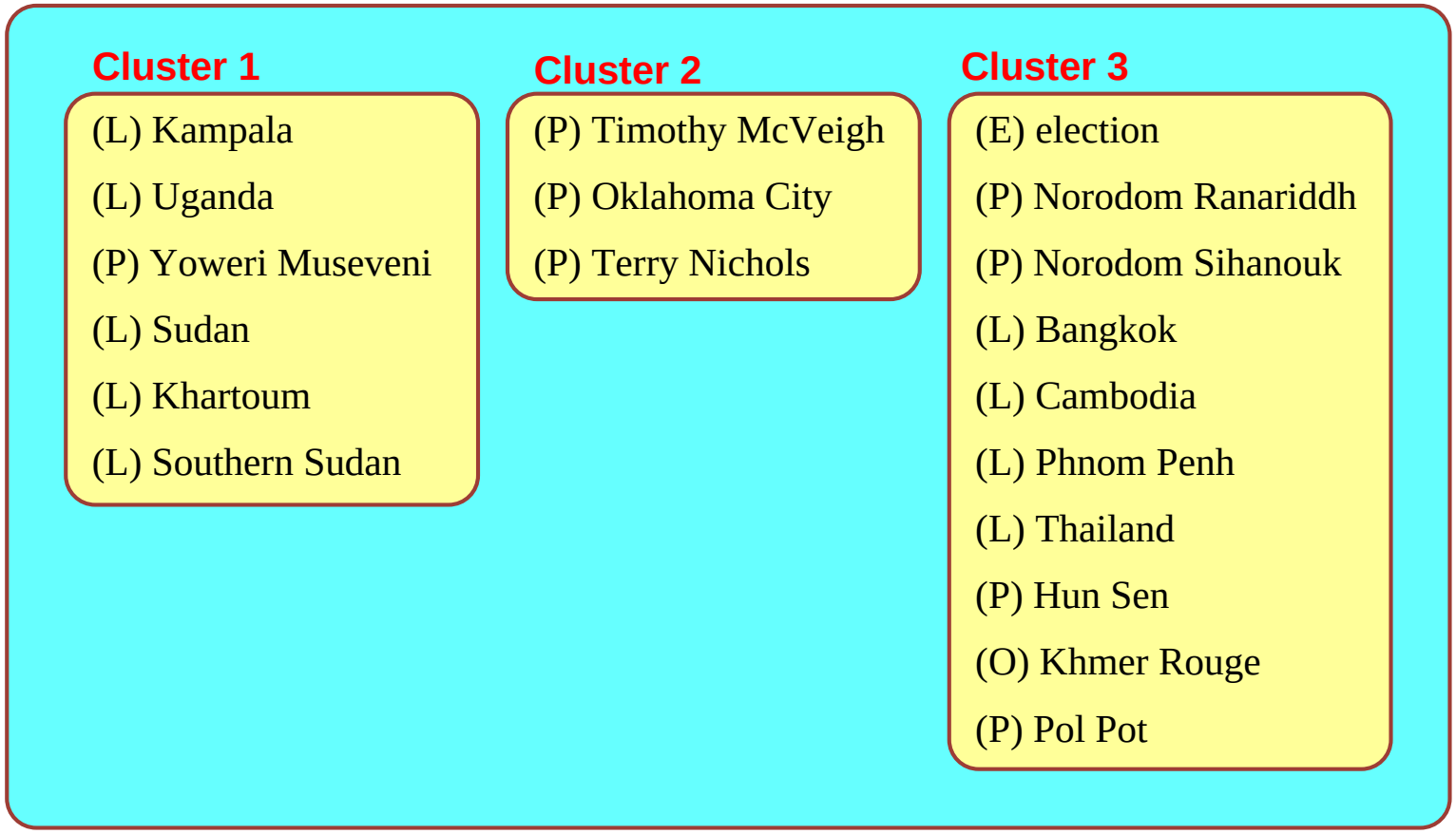
http://en.wikipedia.org/wiki/Text_mining

# 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…




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

# Text Mining Application Area

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

# Text Mining Terminology

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

# 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)

# 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?

# 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

# 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

# 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

# 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

# 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

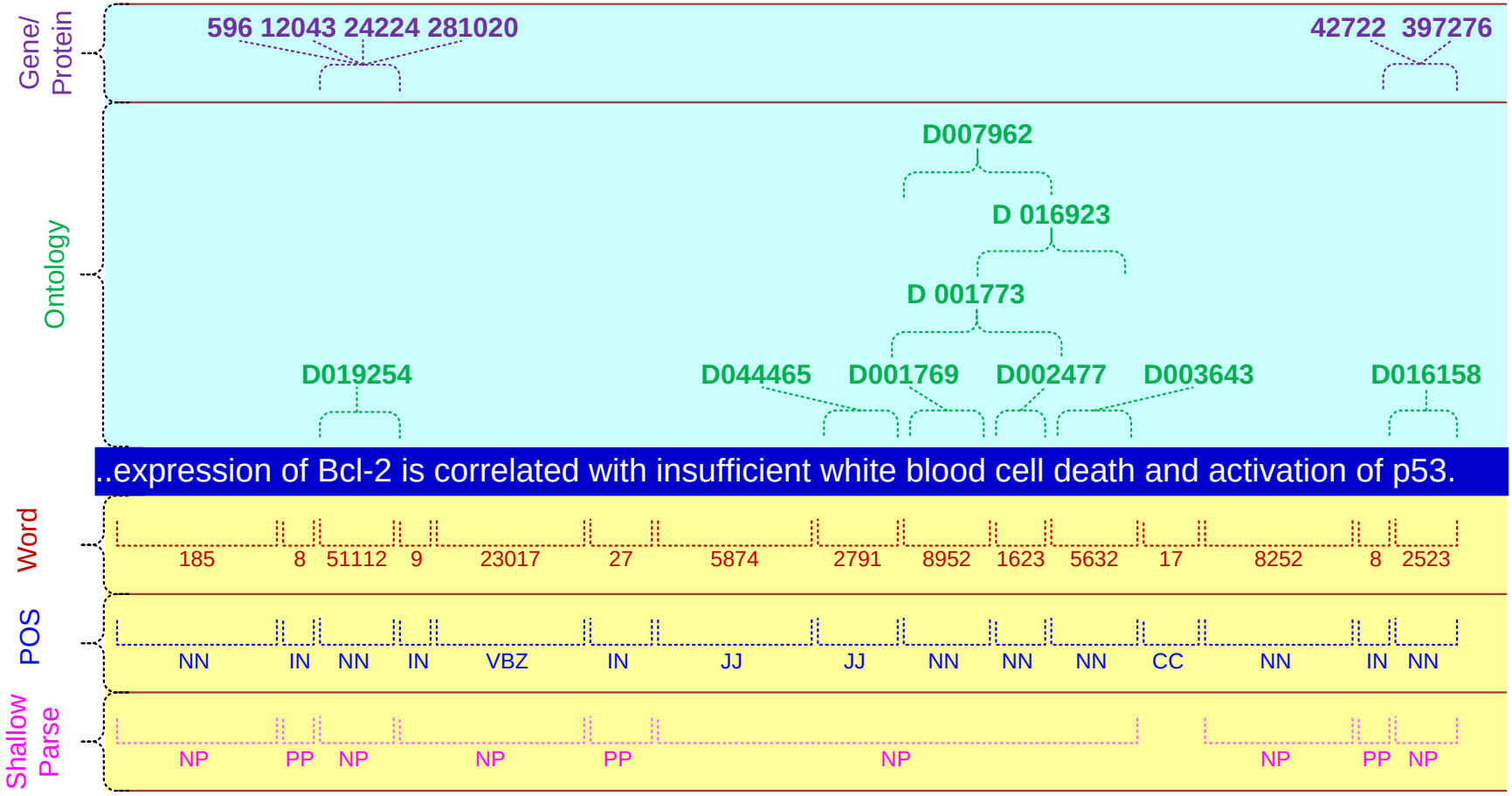
| Category | Example Cues |
|---|---|
| Quantity | Verb count, noun-phrase count, ... |
| Complexity | Avg. no of clauses, sentence length, … |
| Uncertainty | Modifiers, modal verbs, ... |
| Nonimmediacy | Passive voice, objectification, ... |
| Expressivity | Emotiveness |
| Diversity | Lexical diversity, 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

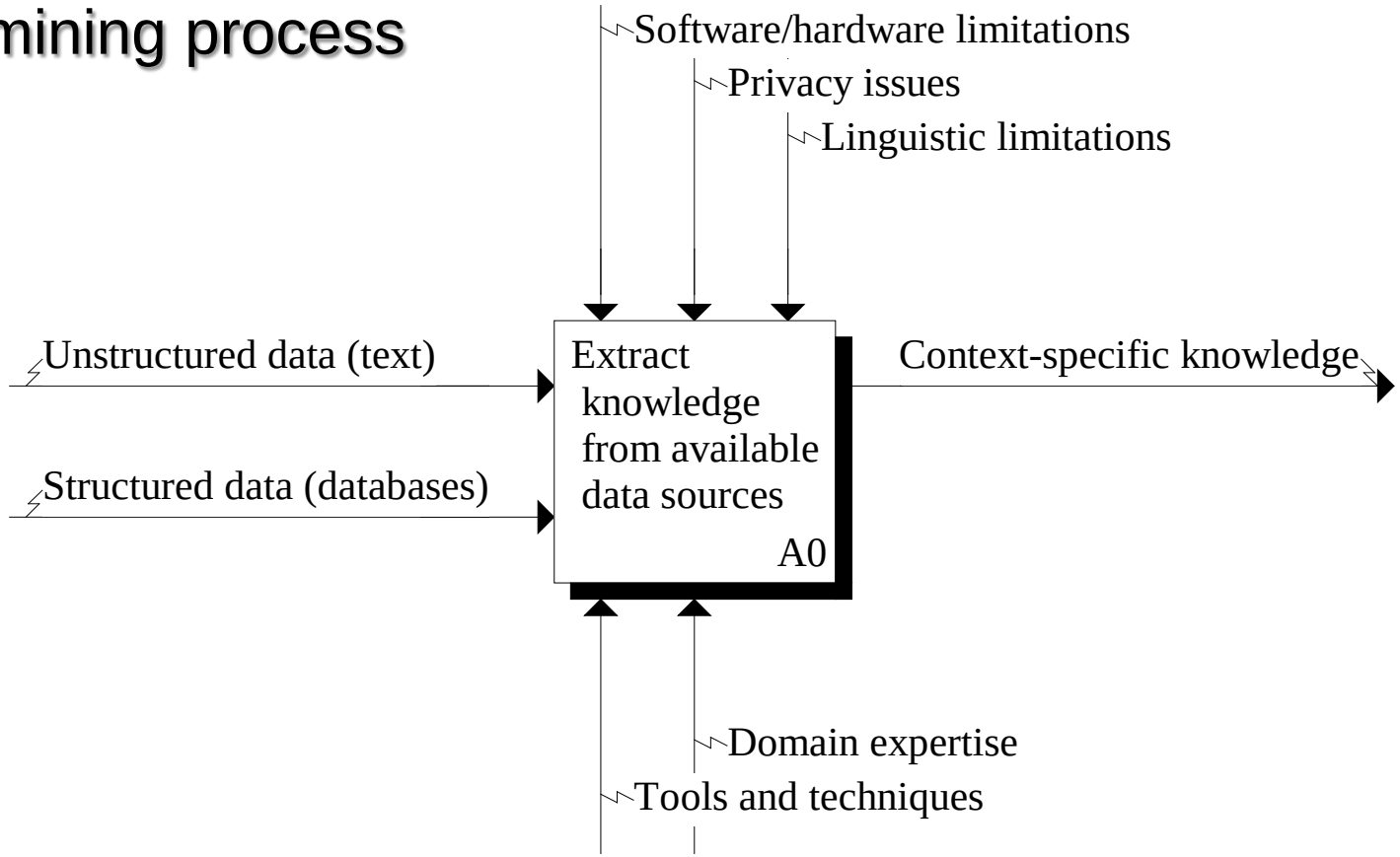
- Application Case: Mining for Lies
  - 371 usable statements are generated
  - 31 features are used
  - Different feature selection methods used
  - 10-fold cross validation is used
  - Results (overall % accuracy)
    - Logistic regression 67.28
    - Decision trees 71.60
    - Neural networks 73.46

# Text Mining Applications
## (gene/protein interaction identification)



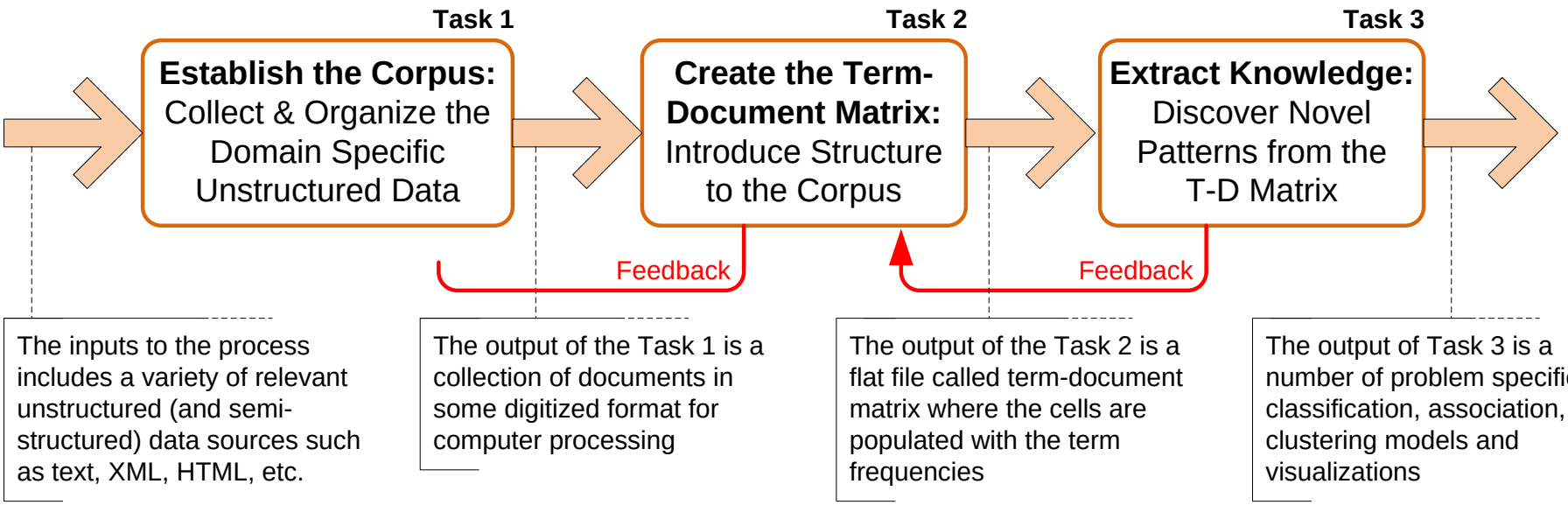
# Text Mining Process

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

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

# 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

| Terms / Documents | 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

# 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 (…)

# Text Mining Application
## (research trend identification in literature)

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

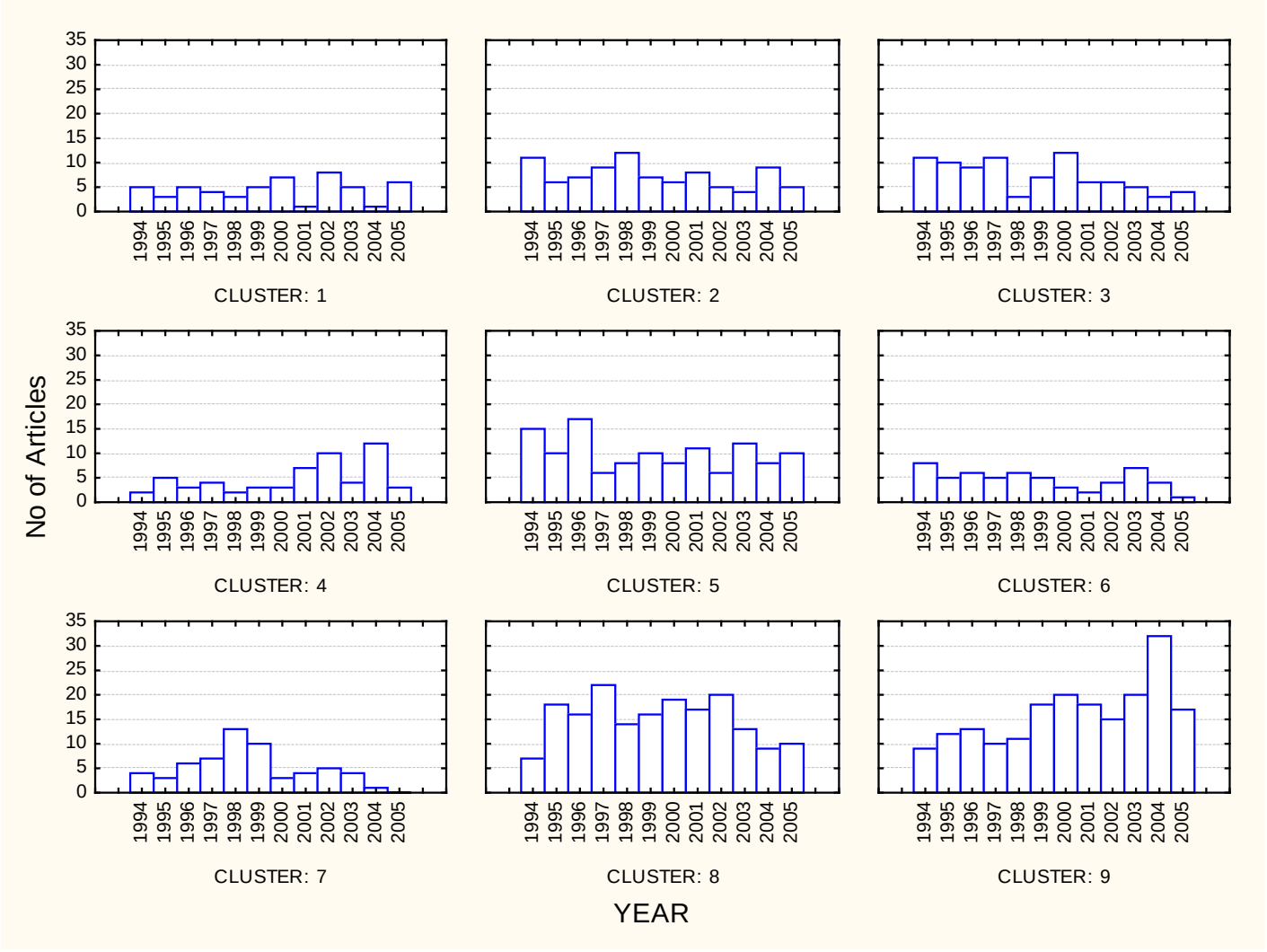
  - Covers 12-year period (1994-2005)
  - 901 papers are included in the study
  - Only the paper abstracts are used
  - 9 clusters are generated for further analysis

# 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 |
| … | … | … | … | … | … | … | … |

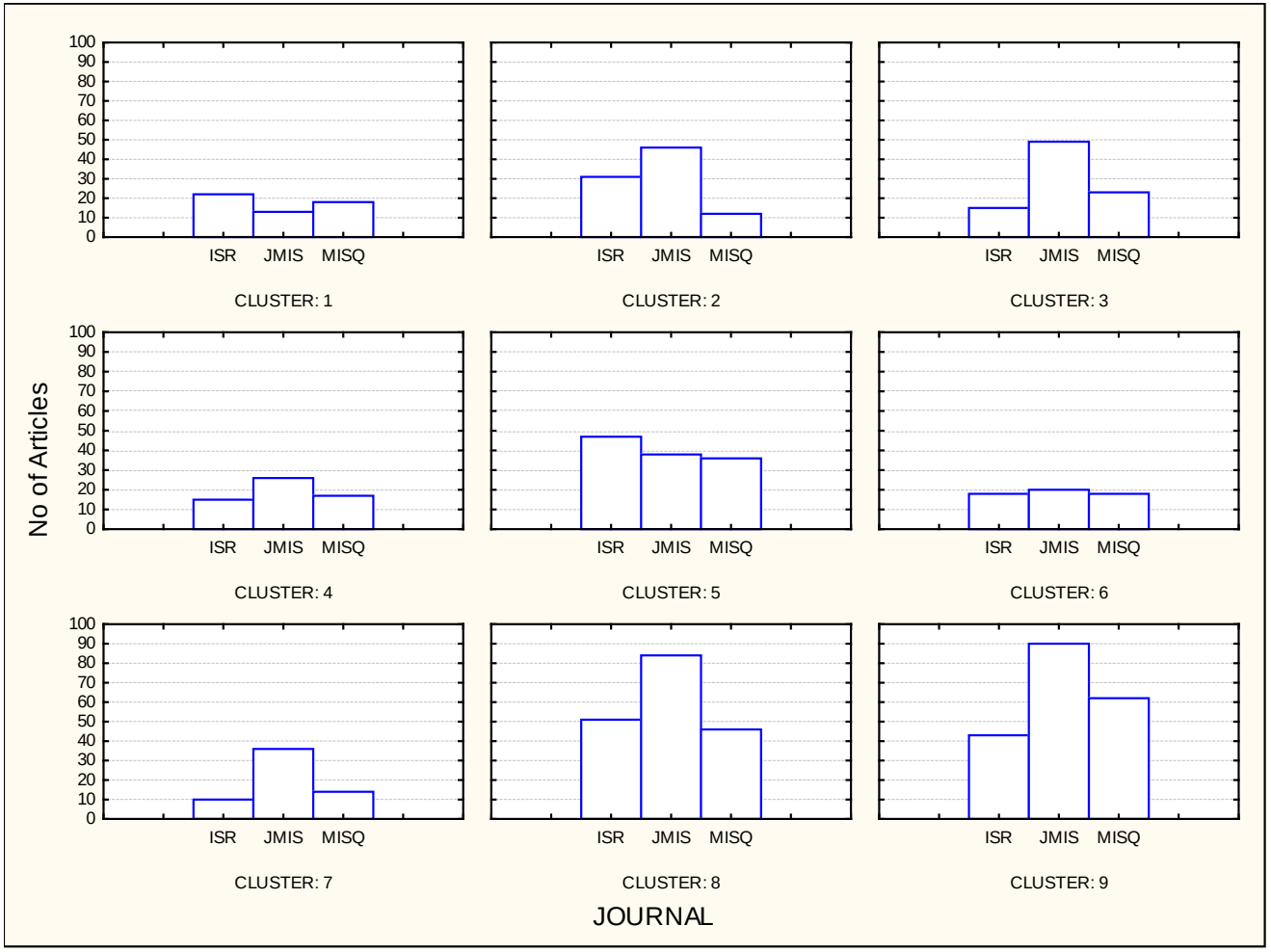
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

# Text Mining Application
## (research trend identification in literature)




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

# Text Mining Application
## (research trend identification in literature)




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

# Text Mining Tools

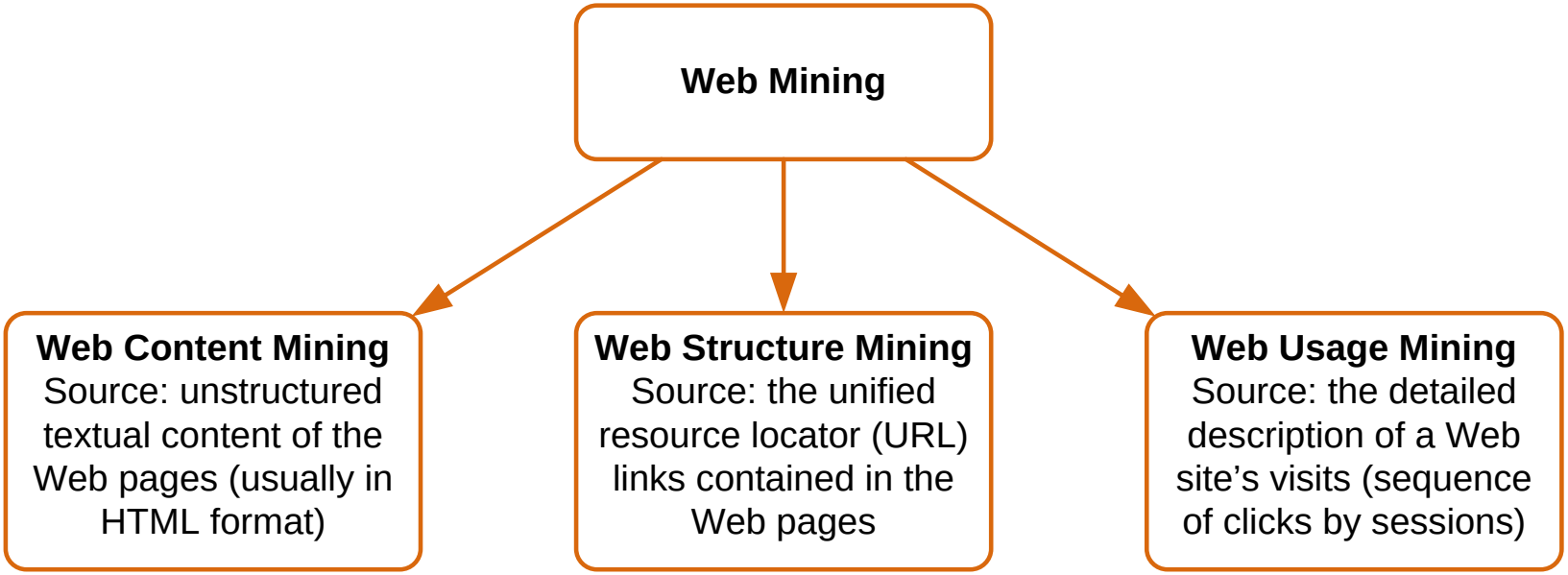
- Commercial Software Tools
  - SPSS PASW Text Miner
  - SAS Enterprise Miner
  - Statistica Data Miner
  - ClearForest, …
- Free Software Tools
  - RapidMiner
  - GATE
  - Spy-EM, …

# 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!

# 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

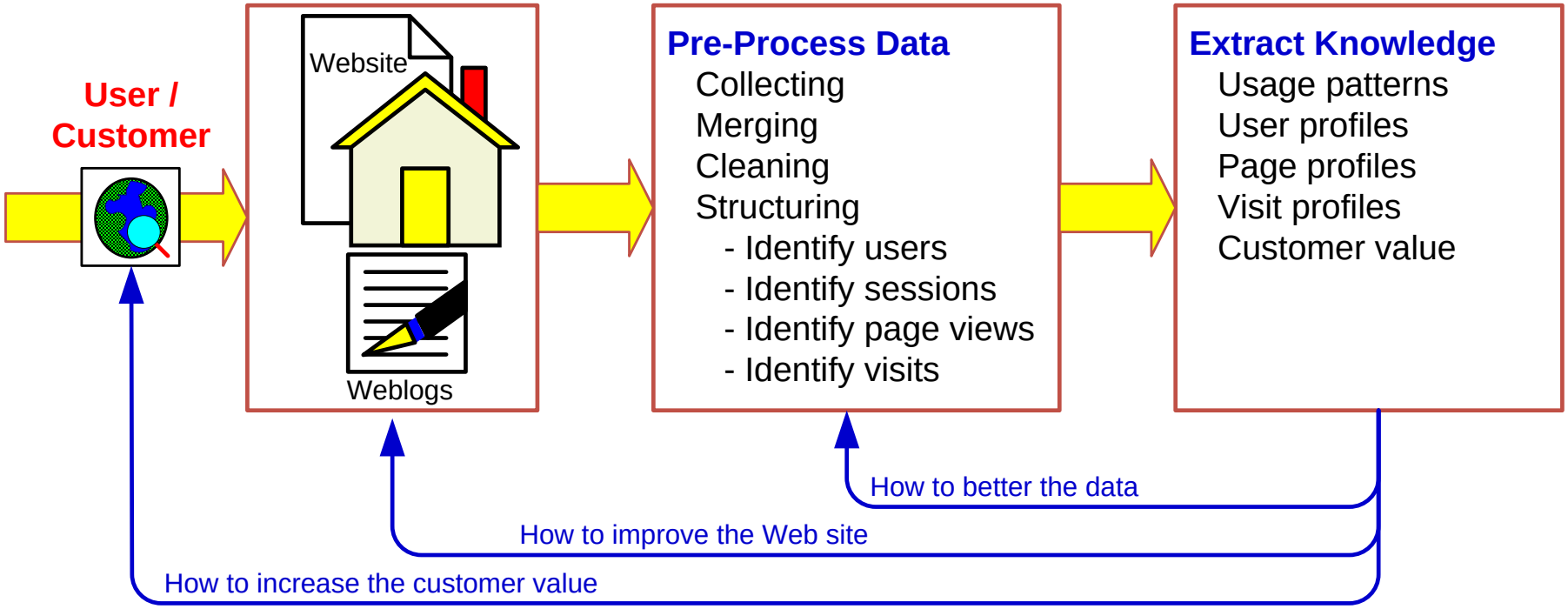
# Web Usage Mining

- 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

# 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…

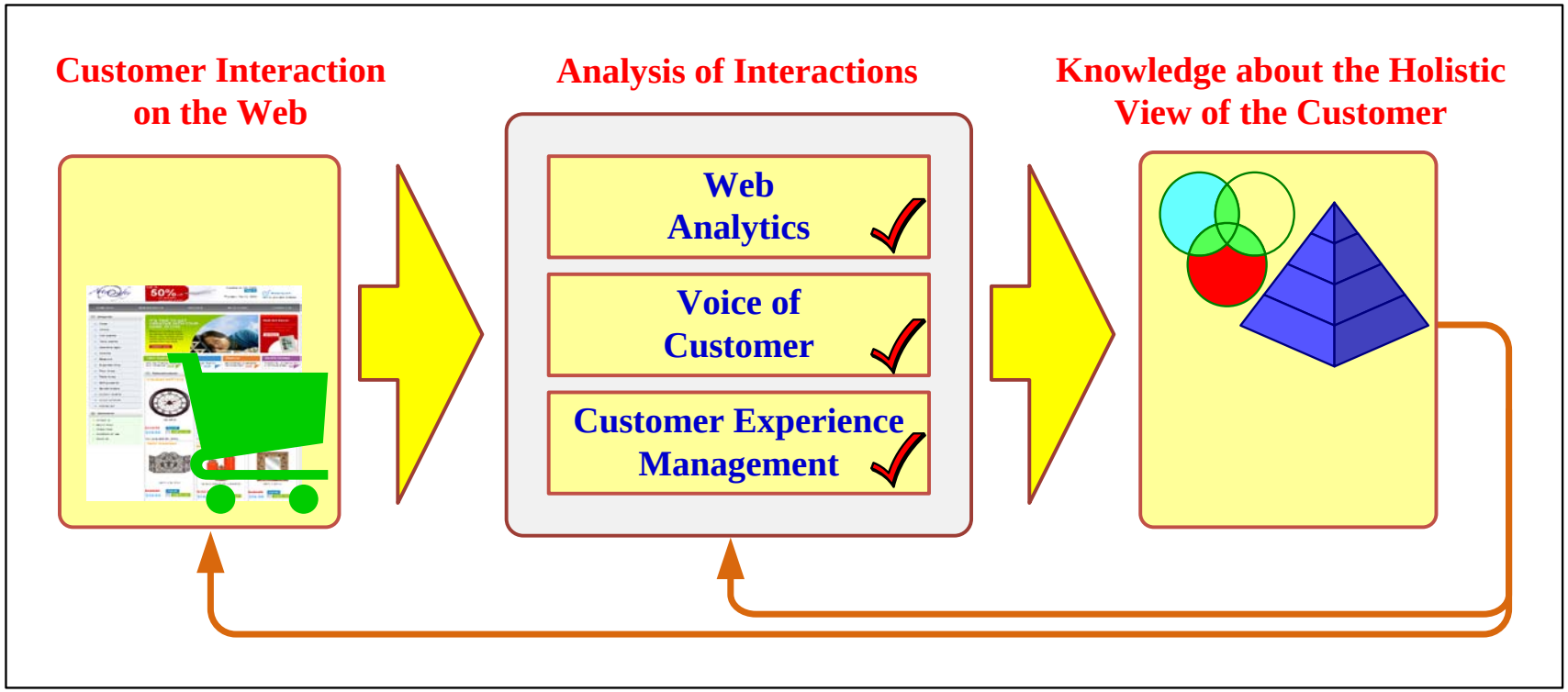
# Web Usage Mining
## (clickstream analysis)




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

# Summary

- Text Mining
- Web Mining

# References


- Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, Ninth Edition, 2011, Pearson.
- 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/
- Text Mining, http://en.wikipedia.org/wiki/Text_mining