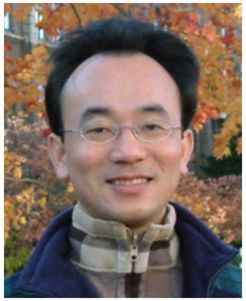
商業智慧實務
Practices of Business Intelligence

Tamkang University

文字探勘與網路探勘
(Text and Web Mining)

1022BI07
MI4
Wed, 9,10 (16:10-18:00) (B113)



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/

2014-04-30

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	103/02/19	商業智慧導論 (Introduction to Business Intelligence)
2	103/02/26	管理決策支援系統與商業智慧 (Management Decision Support System and Business Intelligence)
3	103/03/05	企業績效管理 (Business Performance Management)
4	103/03/12	資料倉儲 (Data Warehousing)
5	103/03/19	商業智慧的資料探勘 (Data Mining for Business Intelligence)
6	103/03/26	商業智慧的資料探勘 (Data Mining for Business Intelligence)
7	103/04/02	教學行政觀摩日 (Off-campus study)
8	103/04/09	資料科學與巨量資料分析 (Data Science and Big Data Analytics)

課程大綱 (Syllabus)

週次	日期	內容（Subject/Topics）
9	103/04/16	期中報告 (Midterm Project Presentation)
10	103/04/23	期中考試週 (Midterm Exam)
11	103/04/30	文字探勘與網路探勘 (Text and Web Mining)
12	103/05/07	意見探勘與情感分析 (Opinion Mining and Sentiment Analysis)
13	103/05/14	社會網路分析 (Social Network Analysis)
14	103/05/21	期末報告 (Final Project Presentation)
15	103/05/28	畢業考試週 (Final Exam)

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

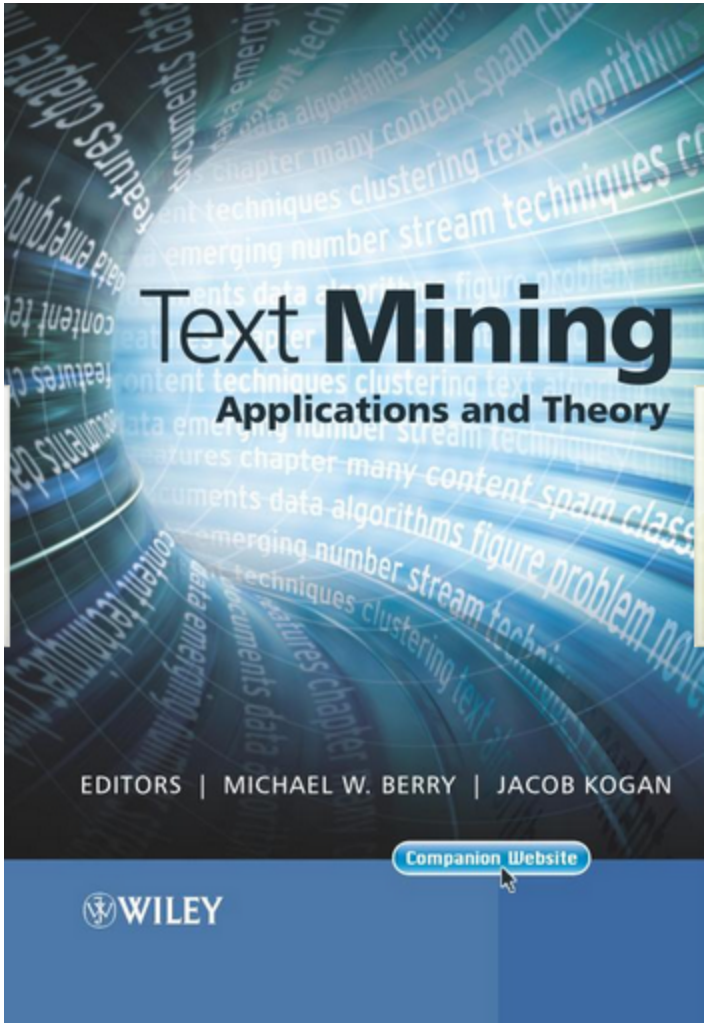
Learning Objectives

- 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

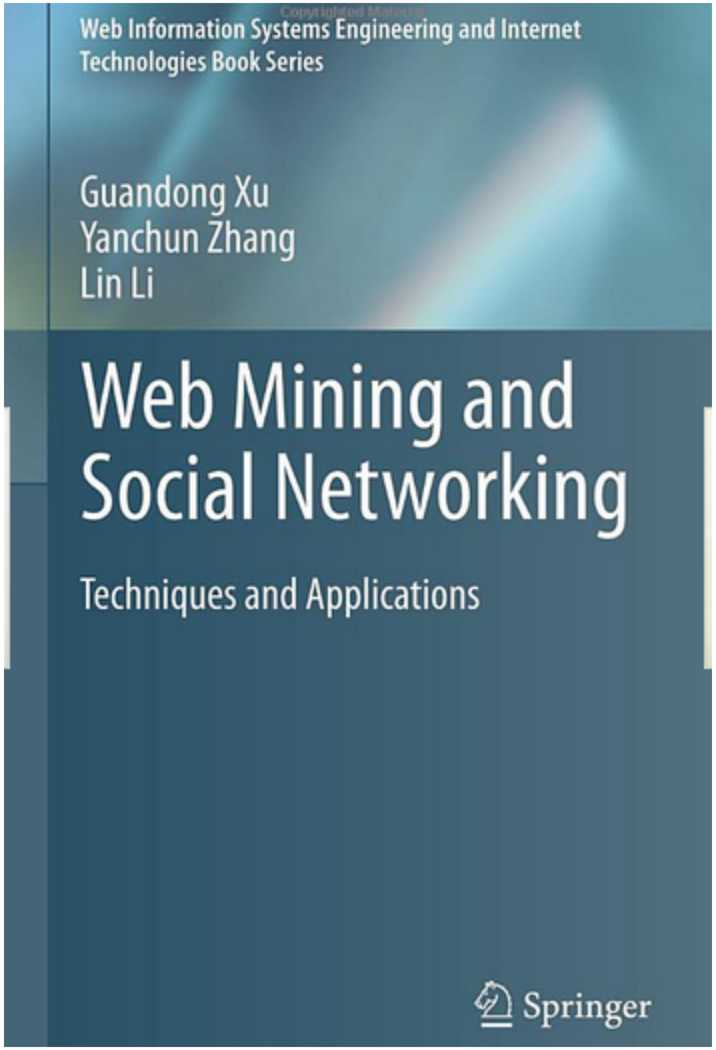
- Text Mining: Applications and Theory
- Web Mining and Social Networking
- Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites
- Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data
- Search Engines – Information Retrieval in Practice

Text Mining



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

Web Mining and Social Networking



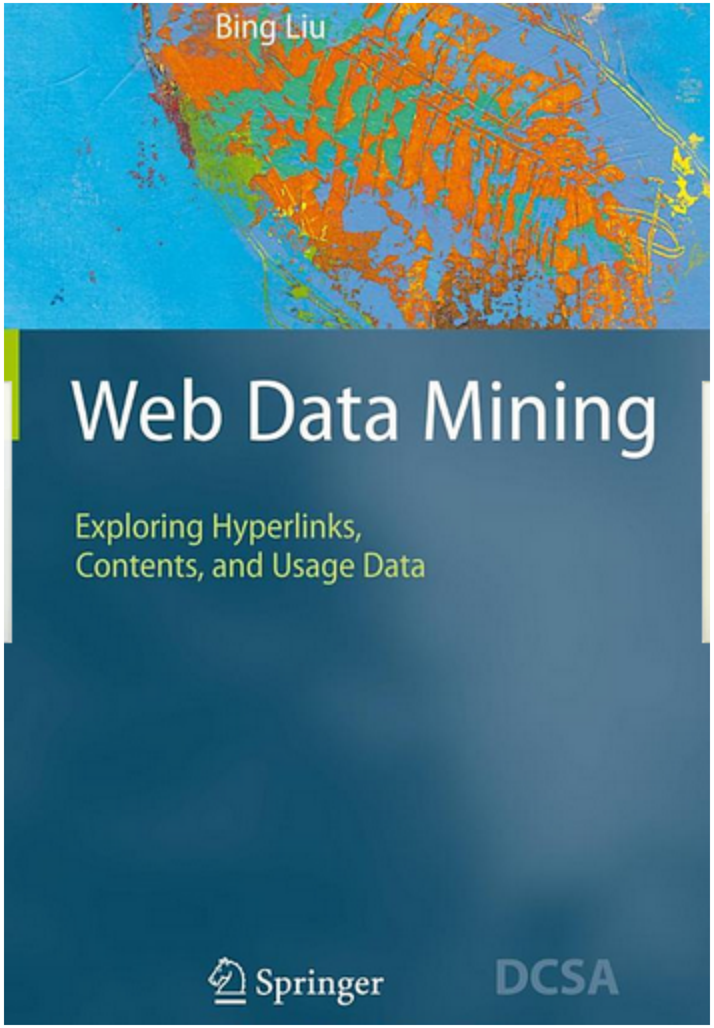
http://www.amazon.com/Web-Mining-Social-Networking-Applications/dp/1441977341

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

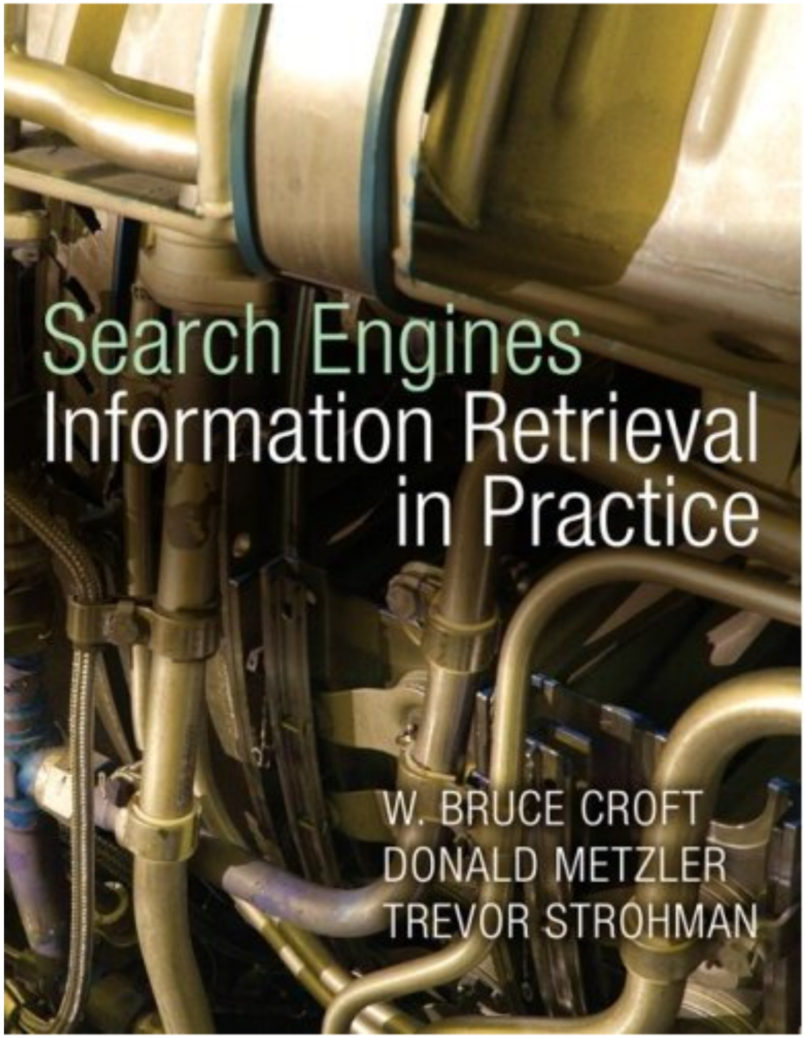
Web Data Mining:

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

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

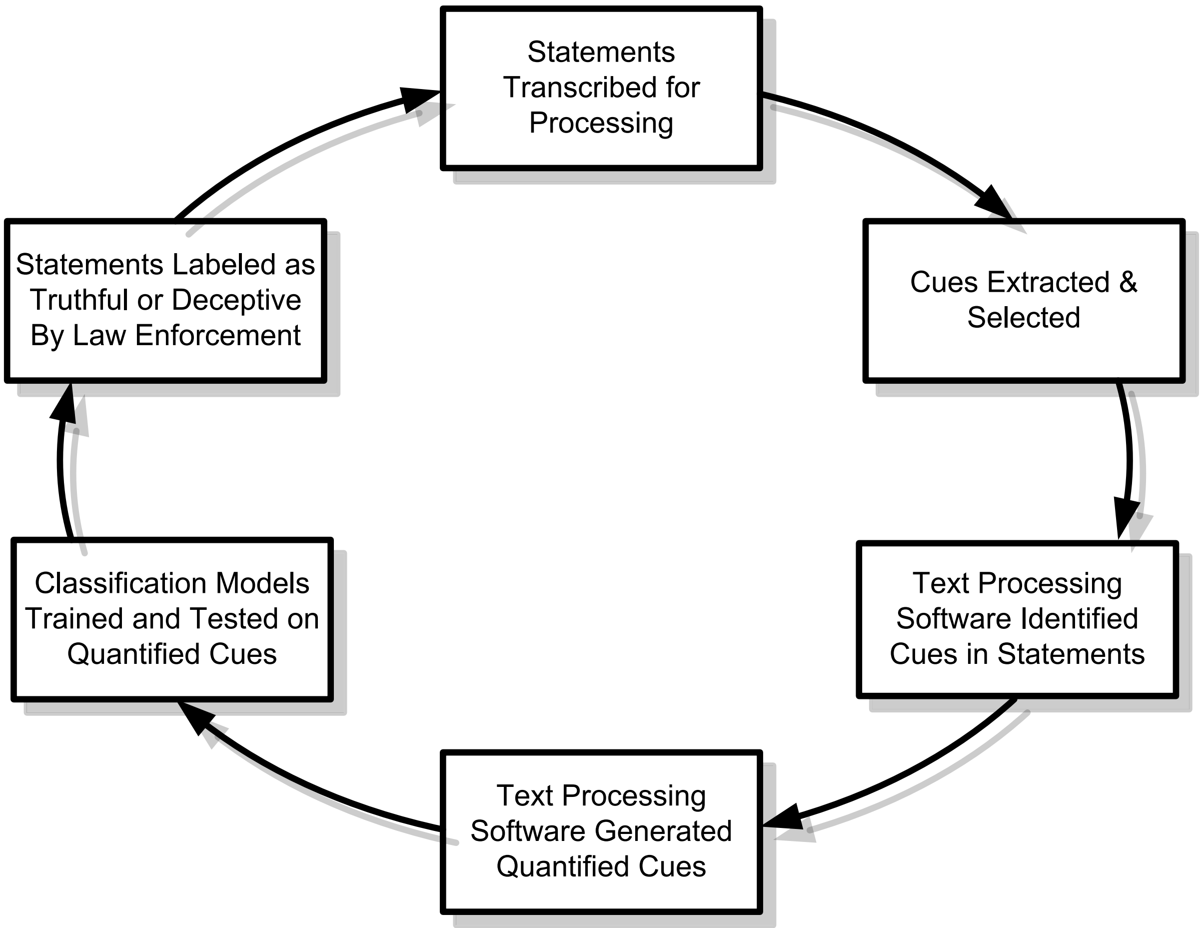
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

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



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

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.

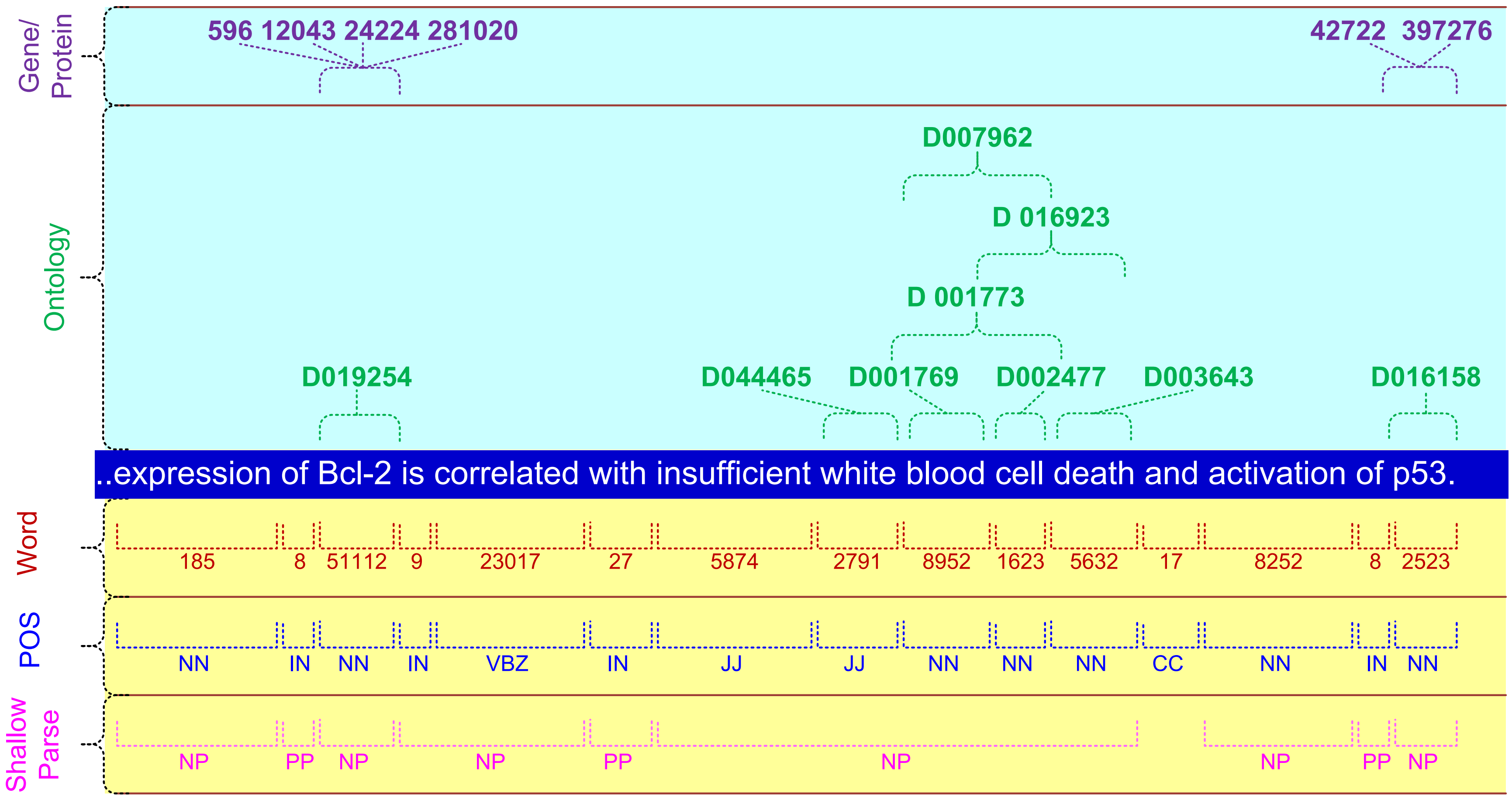
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

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

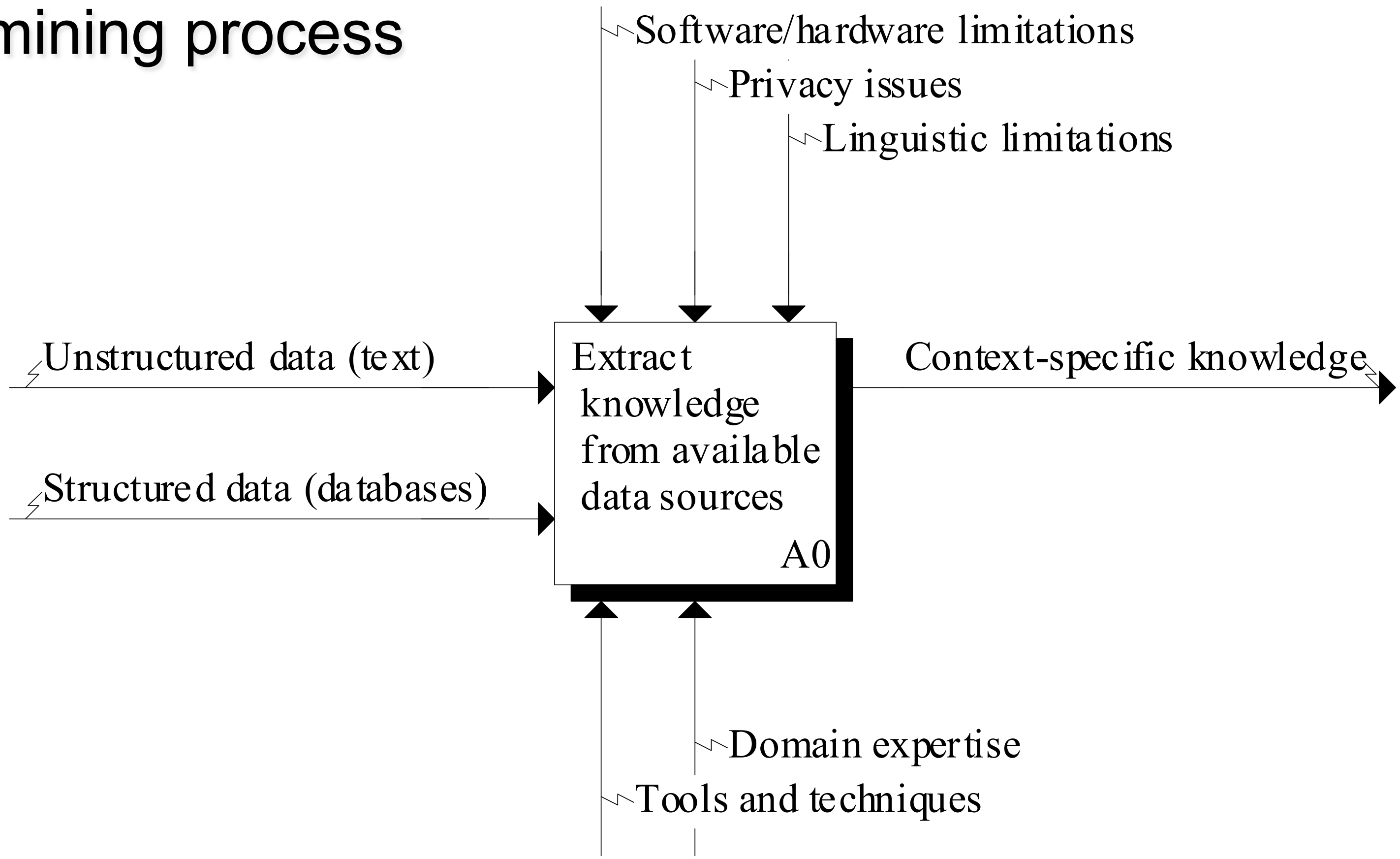
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Text Mining Applications
(gene/protein interaction identification)

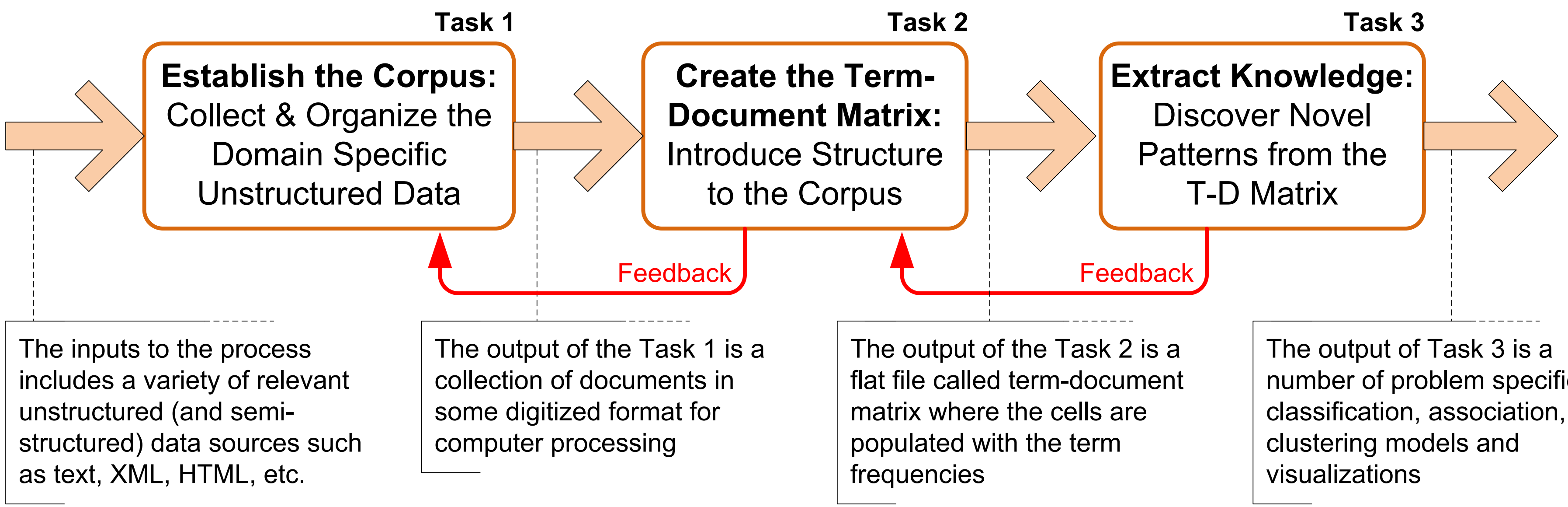


Text Mining Process

Context diagram for the text mining process



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)

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

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

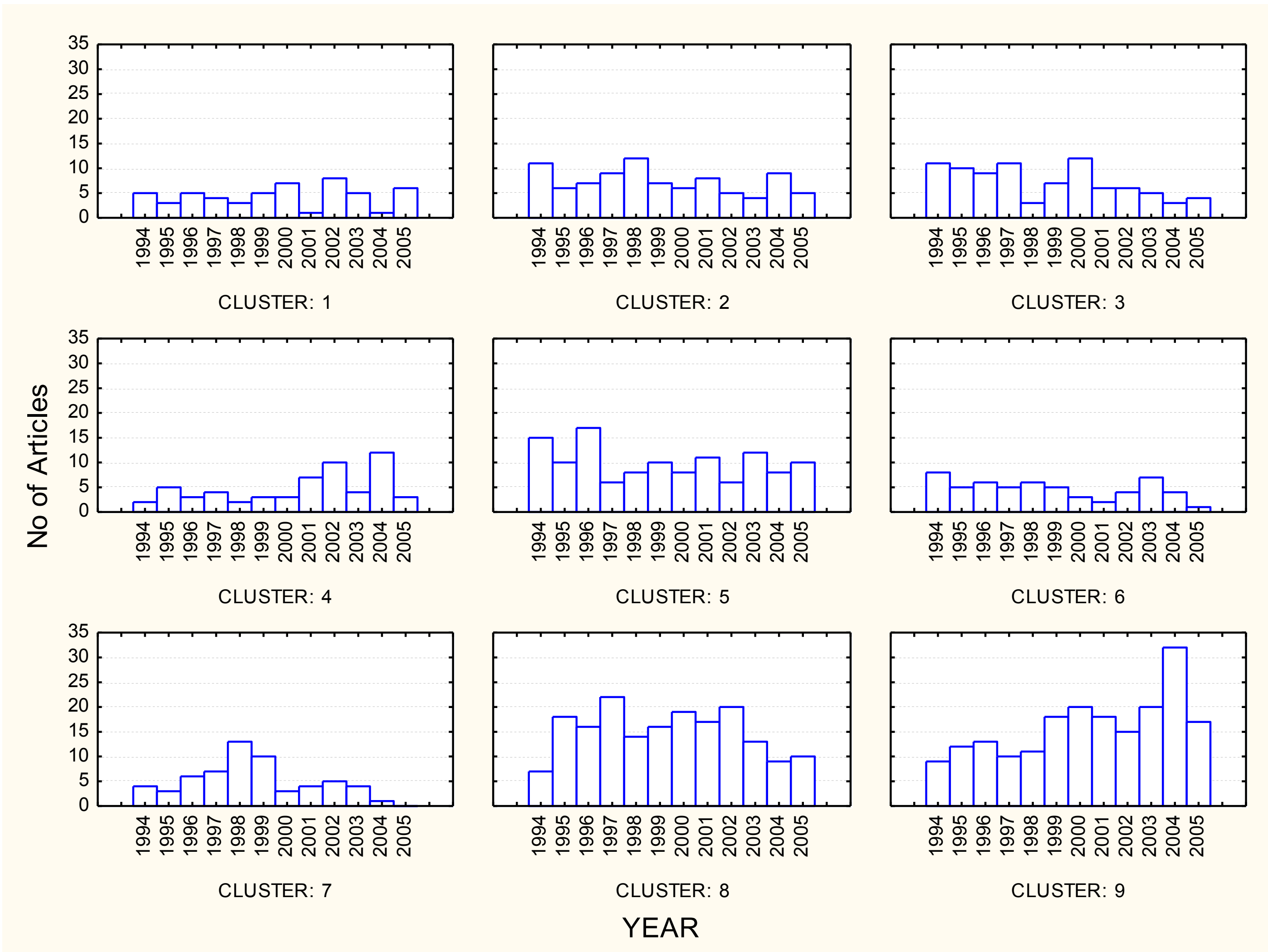
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

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
…	…	…	…	…	…	…	…

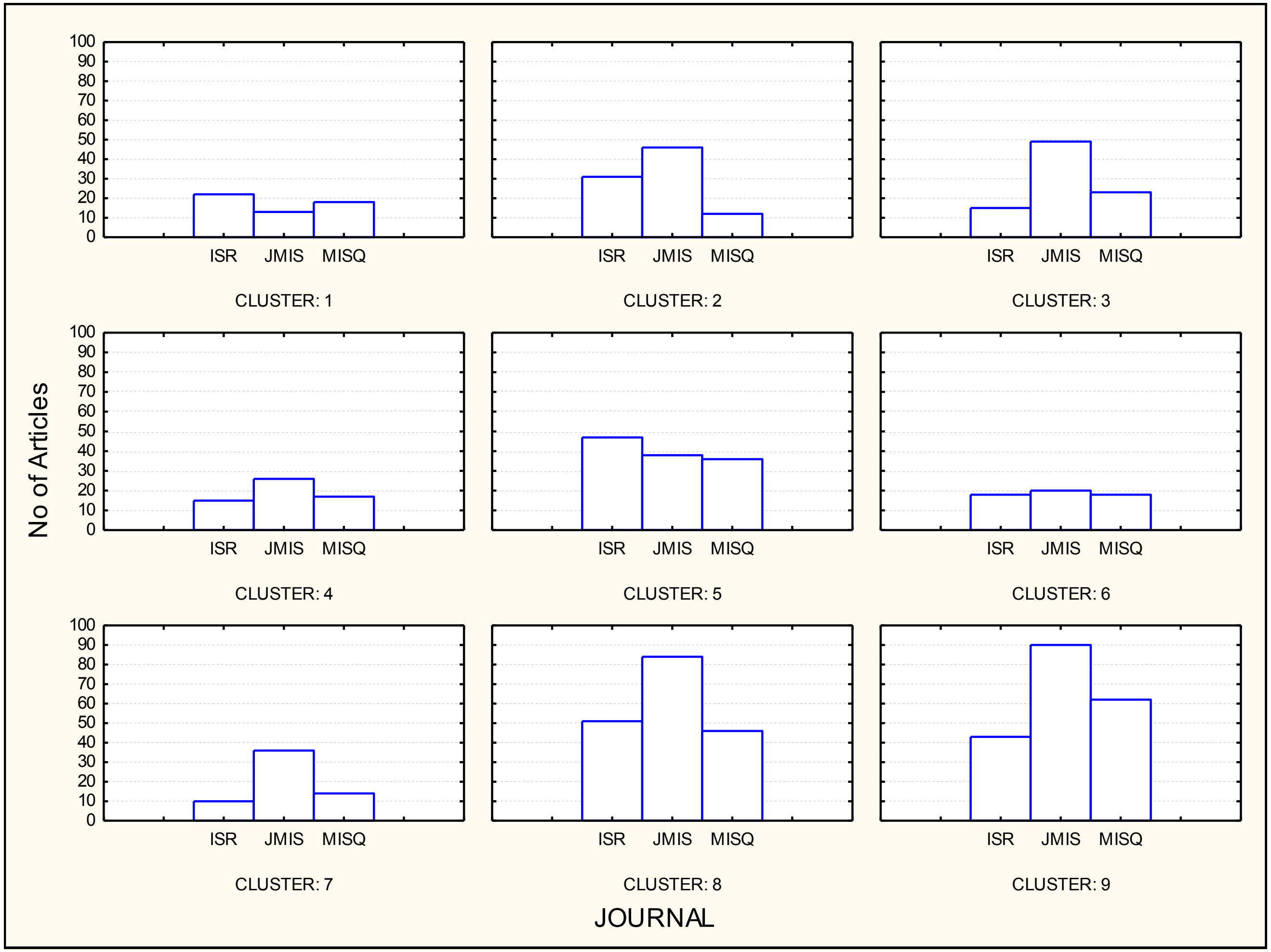
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

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, …

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

SAS Text Analytics

SAS® Text Analytics Software Demo

https://www.youtube.com/watch?v=l1rYdrRCZJ4

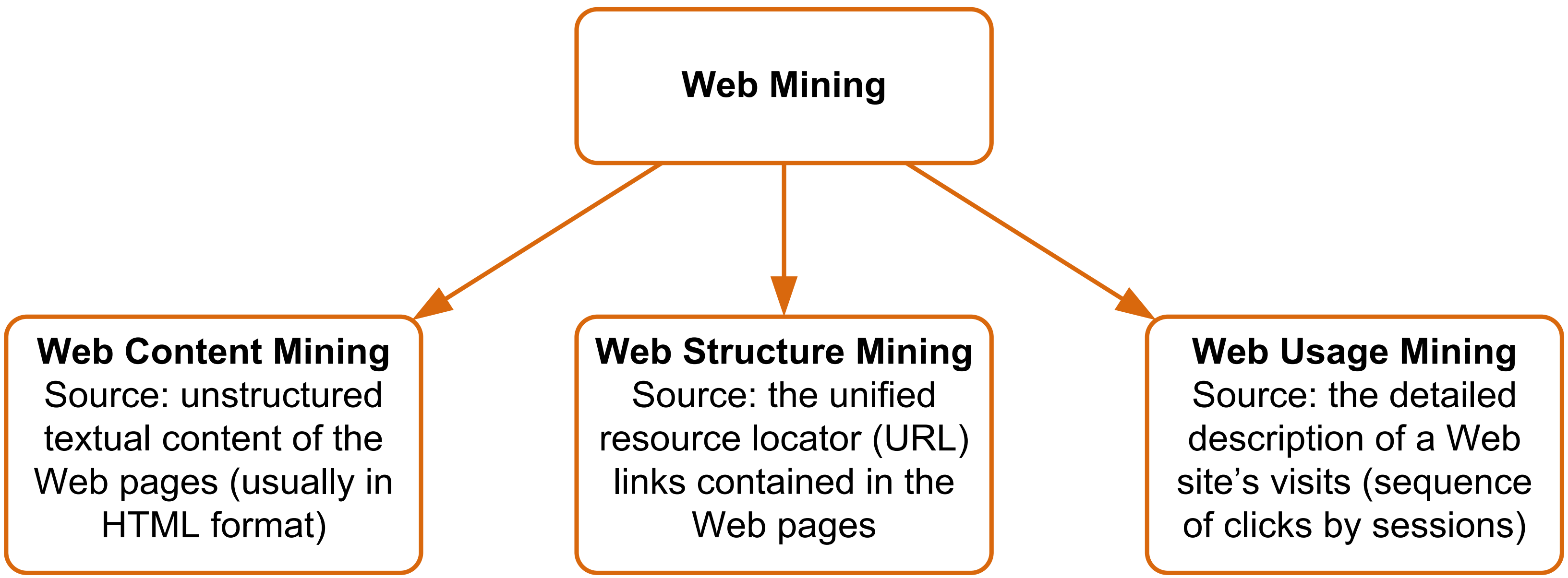
Web Mining Overview

- 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

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

Web Usage Mining

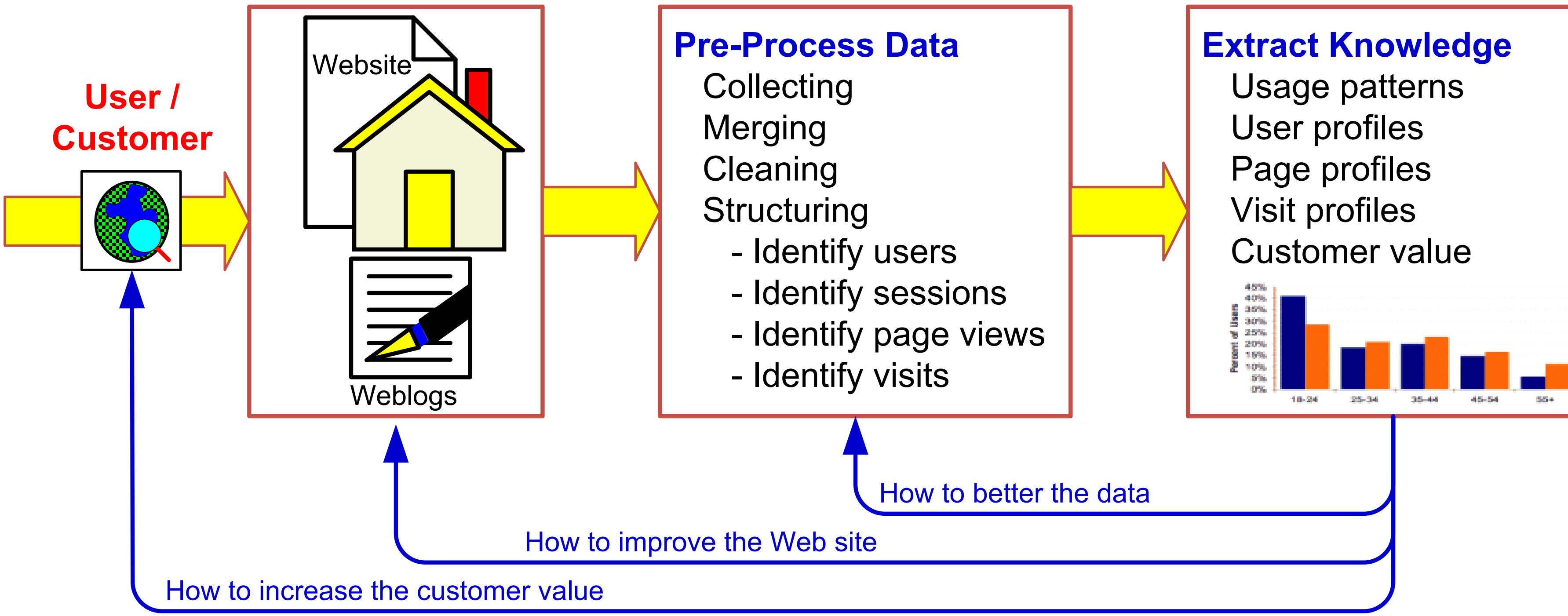
- Extraction of information from data generated through Web page visits and transactions…
  - data stored in server access logs, referrer logs, agent logs, and client-side cookies
  - user characteristics and usage profiles
  - metadata, such as page attributes, content attributes, and usage data
- Clickstream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

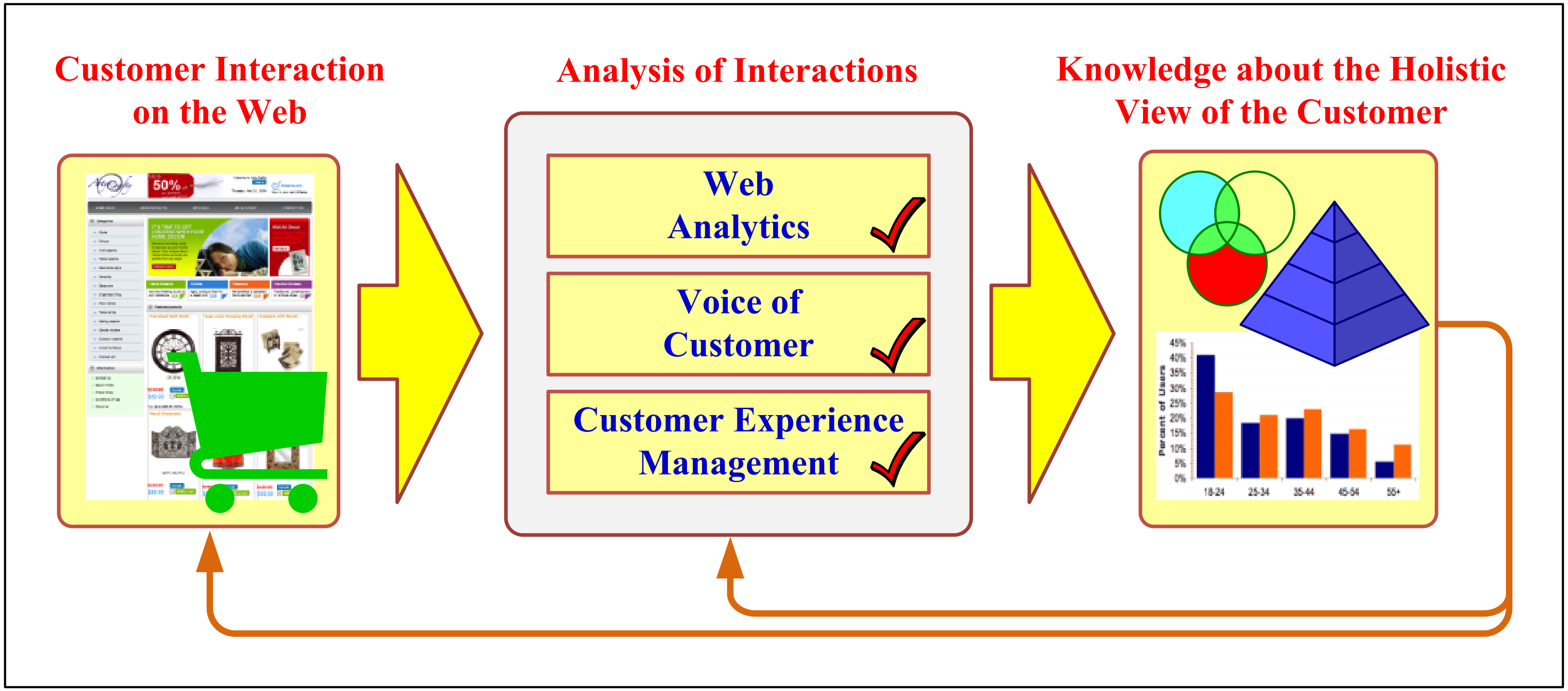
Web Usage Mining
(clickstream analysis)



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

Web Mining Success Stories

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



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

Evaluation of Text Mining and Web Mining

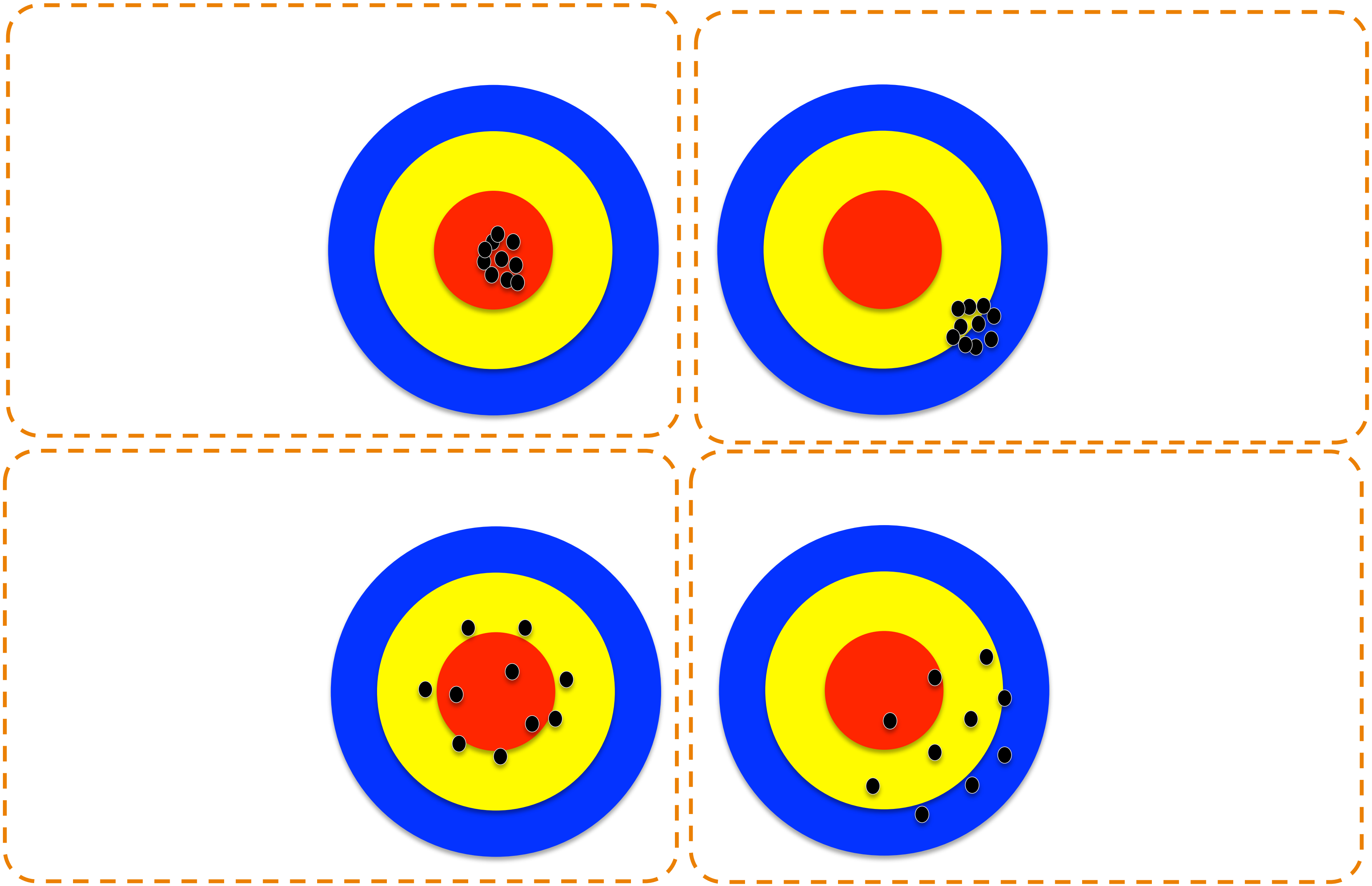
- Evaluation of Information Retrieval
- Evaluation of Classification Model (Prediction)
  - Accuracy
  - Precision
  - Recall
  - F-score

Accuracy

Validity

Precision

Reliability



Accuracy vs. Precision

A

High Accuracy
High Precision

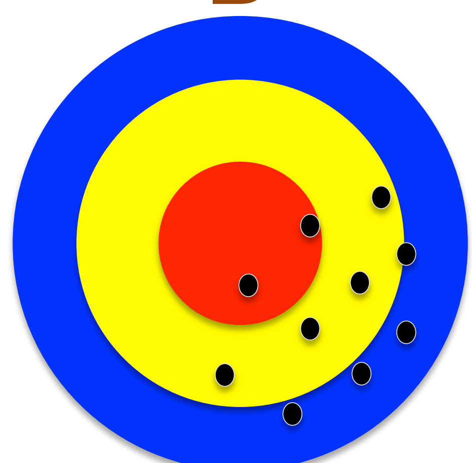
B

Low Accuracy
High Precision

C

High Accuracy
Low Precision

D



Low Accuracy
Low Precision

Accuracy vs. Precision

A

High Accuracy
High Precision

High Validity
High Reliability

B

Low Accuracy
High Precision

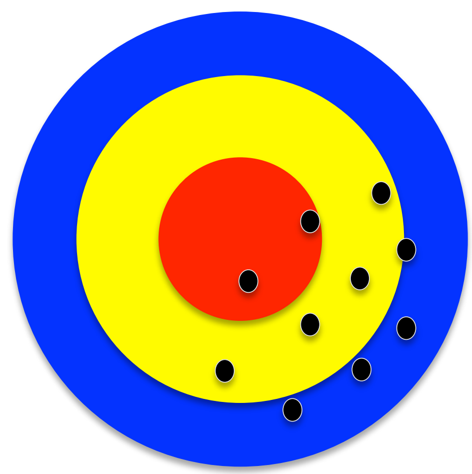
Low Validity
High Reliability

C

High Accuracy
Low Precision

High Validity
Low Reliability

D



Low Accuracy
Low Precision

Low Validity
Low Reliability

Accuracy of Classification Models

- In classification problems, the primary source for accuracy estimation is the confusion matrix

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

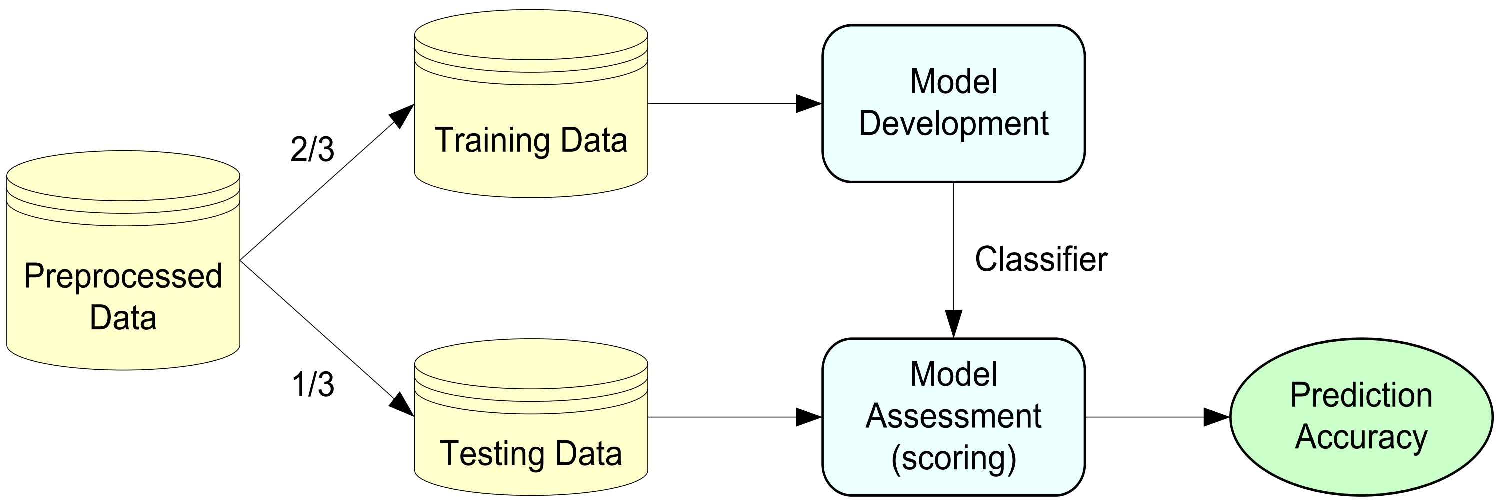
$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Estimation Methodologies for Classification

- Simple split (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets training (~70%) and testing (30%)



  - For ANN, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])

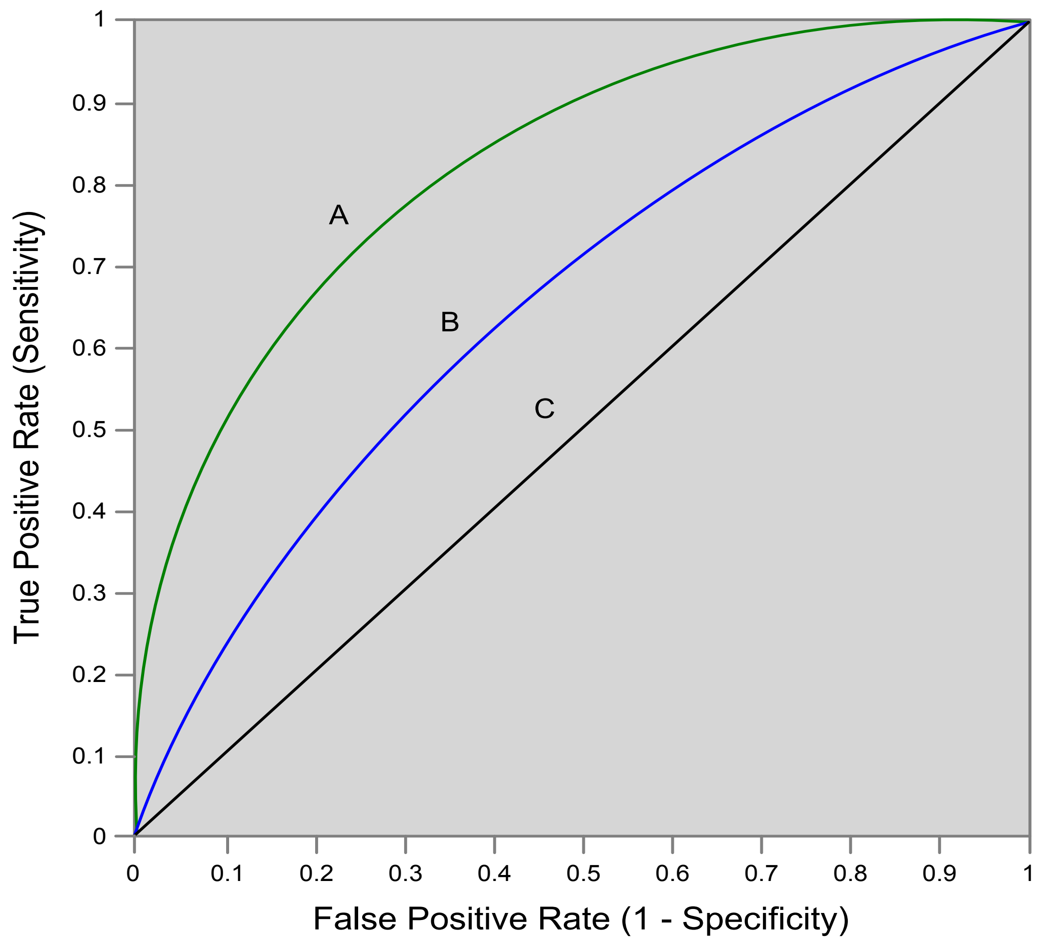
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Estimation Methodologies for Classification

- *k*-Fold Cross Validation (rotation estimation)
  - Split the data into *k* mutually exclusive subsets
  - Use each subset as testing while using the rest of the subsets as training
  - Repeat the experimentation for *k* times
  - Aggregate the test results for true estimation of prediction accuracy training
- Other estimation methodologies
  - Leave-one-out, bootstrapping, jackknifing
  - Area under the ROC curve

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

Estimation Methodologies for Classification – ROC Curve



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

		True Class **(actual value)**		
		Positive	Negative	**total**
Predictive Class (prediction outcome)	Positive	True Positive (TP)	False Positive (FP)	P’
	Negative	False Negative (FN)	True Negative (TN)	N’
total		P	N	

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

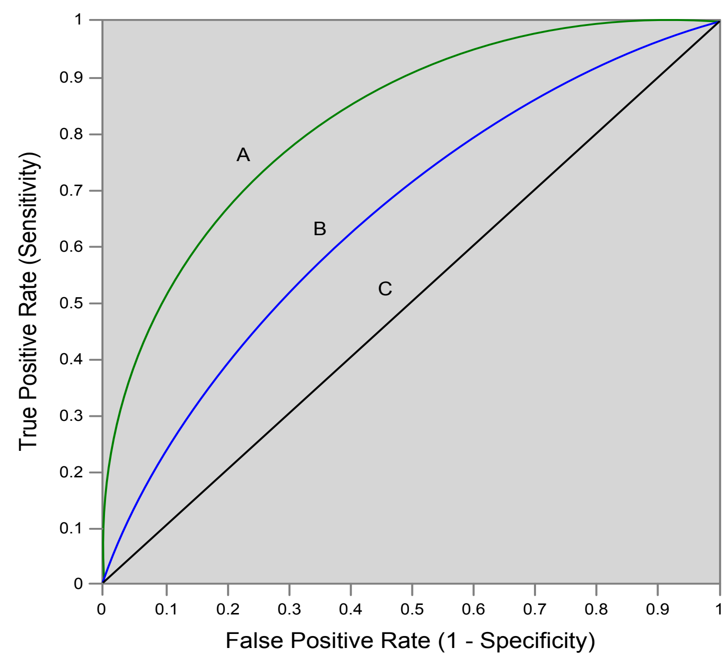
$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$True\ Positive\ Rate\ (\text{Sensitivity}) = \frac{TP}{TP + FN}$$

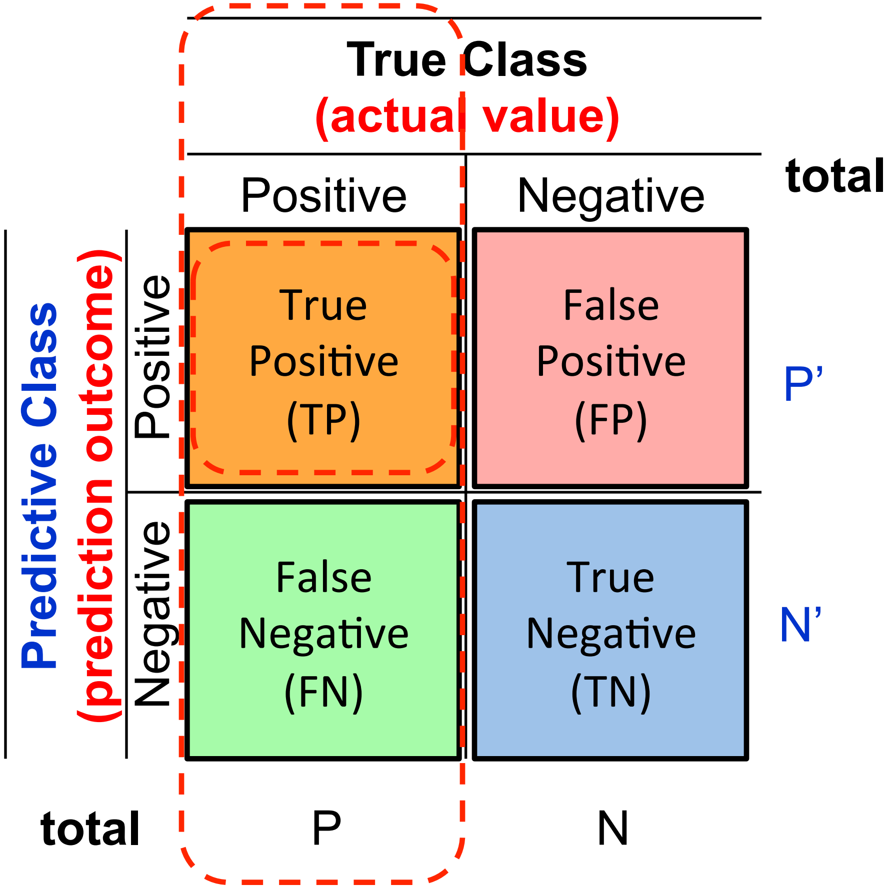
$$True\ Negative\ Rate\ (\text{Specificity}) = \frac{TN}{TN + FP}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

$$False\ Positive\ Rate\ (1 - \text{Specificity}) = \frac{FP}{FP + TN}$$



Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

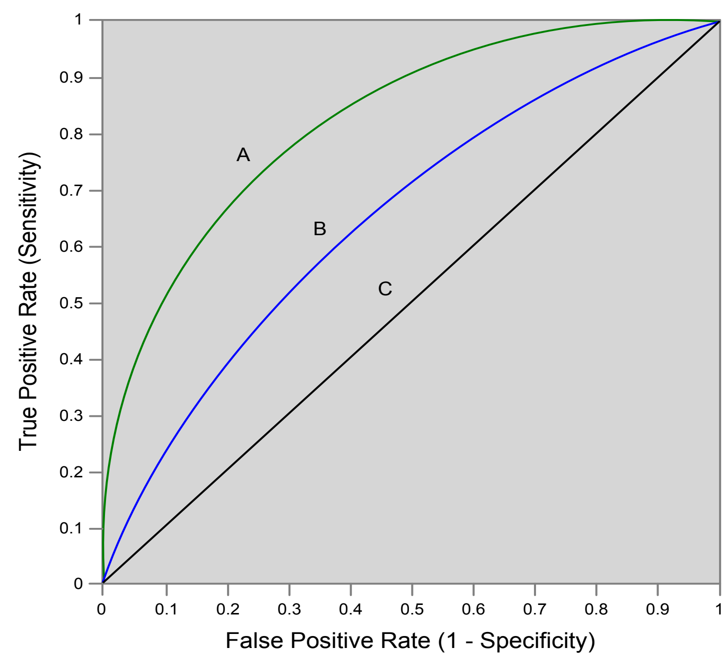


$$True\ Positive\ Rate\ (\text{Sensitivity}) = \frac{TP}{TP + FN}$$

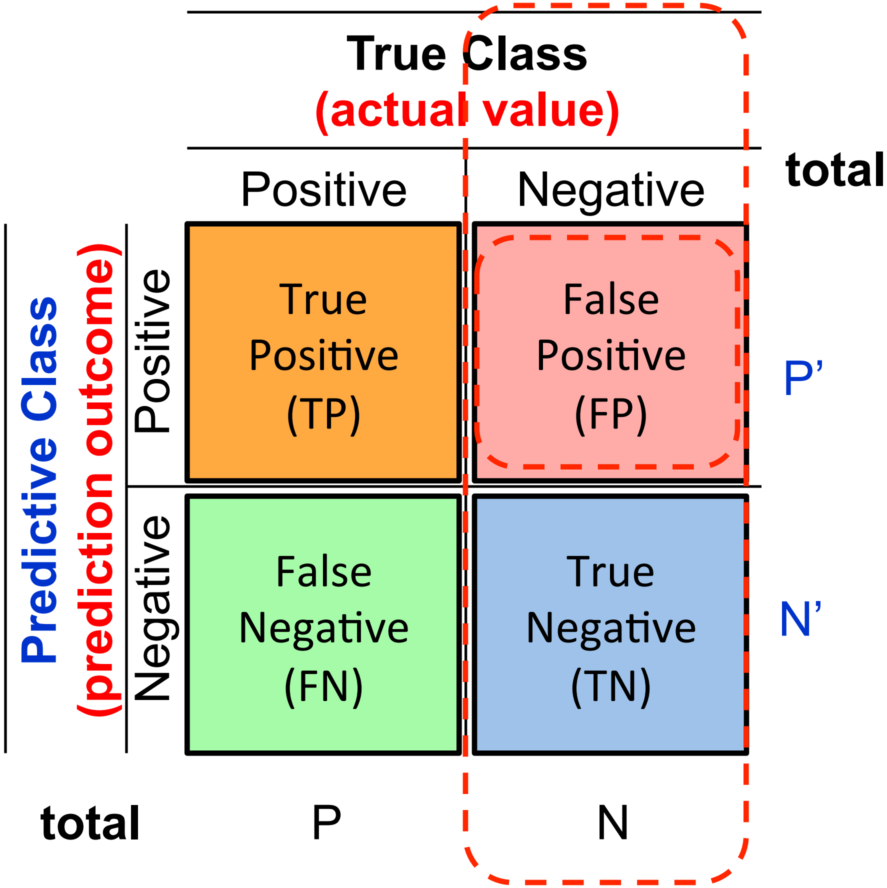
Sensitivity
= True Positive Rate
= Recall
= Hit rate

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$



Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic



$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

Specificity

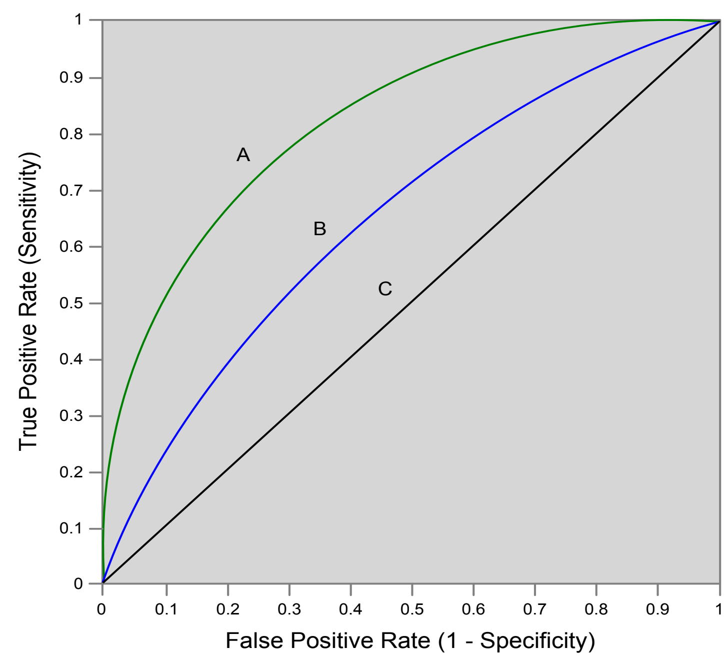
= True Negative Rate

= TN / N

= TN / (TP + TN)

$$True\ Negative\ Rate\ (\text{Specificity}) = \frac{TN}{TN + FP}$$

$$False\ Positive\ Rate\ (1\text{-Specificity}) = \frac{FP}{FP + TN}$$



Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

			True Class (actual value)		
			Positive	Negative	total
Predictive Class (prediction outcome)		Positive	True Positive (TP)	False Positive (FP)	P'
		Negative	False Negative (FN)	True Negative (TN)	N'
total			P	N	

Precision
= Positive Predictive Value (PPV)

$$Precision = \frac{TP}{TP + FP}$$

Recall
= True Positive Rate (TPR)
= Sensitivity
= Hit Rate

$$Recall = \frac{TP}{TP + FN}$$

F1 score (F-score)(F-measure)

is the harmonic mean of precision and recall

= 2TP / (P + P')

= 2TP / (2TP + FP + FN)

$$F = 2 * \frac{precision * recall}{precision + recall}$$

Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

A

63 (TP)	28 (FP)	91
37 (FN)	72 (TN)	109
100	100	200

Recall
= True Positive Rate (TPR)
= Sensitivity
= Hit Rate

Specificity
= True Negative Rate
= TN / N
= TN / (TP + TN)

TPR = 0.63

$$Recall = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate\,(\text{Specificity}) = \frac{TN}{TN + FP}$$

FPR = 0.28

$$\text{False}\ Positive\ Rate\,(1\text{-}\text{Specificity}) = \frac{FP}{FP + TN}$$

PPV = 0.69
=63/(63+28)
=63/91

$$Precision = \frac{TP}{TP + FP}$$

Precision
= Positive Predictive Value (PPV)

F1 = 0.66
= 2*(0.63*0.69)/(0.63+0.69)
= (2 * 63) /(100 + 91)
= (0.63 + 0.69) / 2 =1.32 / 2 =0.66

$$F = 2 * \frac{precision * recall}{precision + recall}$$

F1 score (F-score) (F-measure)
is the harmonic mean of precision and recall
= 2TP / (P + P’)
= 2TP / (2TP + FP + FN)

ACC = 0.68
= (63 + 72) / 200
= 135/200 = 67.5

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

A

63 (TP)	28 (FP)	91
37 (FN)	72 (TN)	109
100	100	200

TPR = 0.63

FPR = 0.28

PPV = 0.69
=63/(63+28)
=63/91

F1 = 0.66
= 2*(0.63*0.69)/(0.63+0.69)
= (2 * 63) /(100 + 91)
= (0.63 + 0.69) / 2 =1.32 / 2 =0.66

ACC = 0.68
= (63 + 72) / 200
= 135/200 = 67.5

B

77 (TP)	77 (FP)	154
23 (FN)	23 (TN)	46
100	100	200

TPR = 0.77
FPR = 0.77
PPV = 0.50
F1 = 0.61
ACC = 0.50

Recall
= True Positive Rate (TPR)
= <u>Sensitivity</u>
= Hit Rate

$$Recall = \frac{TP}{TP + FN}$$

Precision
= Positive Predictive Value (PPV)

$$Precision = \frac{TP}{TP + FP}$$

C

24 (TP)	88 (FP)	112
76 (FN)	12 (TN)	88
100	100	200

TPR = 0.24
FPR = 0.88
PPV = 0.21
F1 = 0.22
ACC = 0.18

C'

76 (TP)	12 (FP)	88
24 (FN)	88 (TN)	112
100	100	200

TPR = 0.76
FPR = 0.12
PPV = 0.86
F1 = 0.81
ACC = 0.82

Recall
= True Positive Rate (TPR)
= Sensitivity
= Hit Rate

$$Recall = \frac{TP}{TP + FN}$$

Precision
= Positive Predictive Value (PPV)

$$Precision = \frac{TP}{TP + FP}$$

Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

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