Web Mining
(網路探勘)

Introduction to Web Mining
(網路探勘導論)

1011WM01
TLMXM1A
Wed 8,9 (15:10-17:00) U705

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/
2012-09-12

淡江大學101學年度第1學期
課程教學計畫表
(2012.09-2013.01)

- 課程名稱：Web Mining (網路探勘)
- 授課教師：戴敏育 (Min-Yuh Day)
- 開課系級：資管一碩士班Ａ(TLMXM1A)
- 開課資料：選修 單學期 2 學分 (2 Credits, Elective)
- 上課時間：週三 8, 9 (Wed 15:10-17:00)
- 上課教室：U705

課程簡介

- 本課程介紹網路探勘的基礎概念及技術。
- 課程內容包括
  - 網路探勘導論、
  - 關聯規則和序列模式、
  - 監督式學習、
  - 非監督式學習、
  - 部分監督式學習、
  - 資訊檢索與網路搜尋、
  - 社會網路分析、
  - 網路爬行、
  - 結構化資料擷取、
  - 資訊整合、
  - 意見探勘與情感分析、
  - 網路使用挖掘。

Course Introduction

- This course introduces the fundamental concepts and technology of web mining.
- Topics include
  - Introduction to Web Mining,
  - Association Rules and Sequential Patterns,
  - Supervised Learning,
  - Unsupervised Learning,
  - Partially Supervised Learning,
  - Information Retrieval and Web Search,
  - Social Network Analysis,
  - Web Crawling,
  - Structured Data Extraction,
  - Information Integration,
  - Opinion Mining and Sentiment Analysis, and
  - Web Usage Mining.

課程目標

- 瞭解及應用網路探勘基本概念與技術。
- 進行網路探勘相關之資訊管理研究。

Objective

- Students will be able to understand and apply the fundamental concepts and technology of web mining.
- Students will be able to conduct information systems research in the context of web mining.

課程大綱 (Syllabus)

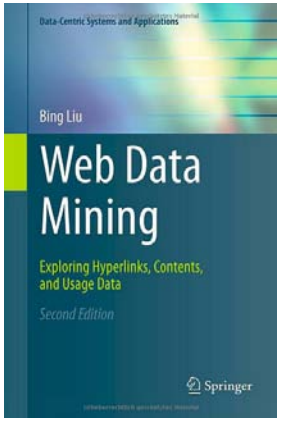
週次	日期	內容（Subject/Topics）
1	101/09/12	Introduction to Web Mining (網路探勘導論)
2	101/09/19	Association Rules and Sequential Patterns (關聯規則和序列模式)
3	101/09/26	Supervised Learning (監督式學習)
4	101/10/03	Unsupervised Learning (非監督式學習)
5	101/10/10	國慶紀念日(放假一天)
6	101/10/17	Paper Reading and Discussion (論文研讀與討論)
7	101/10/24	Partially Supervised Learning (部分監督式學習)
8	101/10/31	Information Retrieval and Web Search (資訊檢索與網路搜尋)
9	101/11/07	Social Network Analysis (社會網路分析)

課程大綱 (Syllabus)

週次	日期	內容（Subject/Topics）
10	101/11/14	Midterm Presentation (期中報告)
11	101/11/21	Web Crawling (網路爬行)
12	101/11/28	Structured Data Extraction (結構化資料擷取)
13	101/12/05	Information Integration (資訊整合)
14	101/12/12	Opinion Mining and Sentiment Analysis (意見探勘與情感分析)
15	101/12/19	Paper Reading and Discussion (論文研讀與討論)
16	101/12/26	Web Usage Mining (網路使用挖掘)
17	102/01/02	Project Presentation 1 (期末報告1)
18	102/01/09	Project Presentation 2 (期末報告2)

教材課本與參考書籍

- 教材課本 (Textbook)
  - Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition.
  - http://www.cs.uic.edu/~liub/WebMiningBook.html



- 參考書籍 (References)
  - Related Papers.

學期成績計算方式

- 平時評量： 50.0 %
- 其他 (課堂參與及報告討論表現)：50.0 %

Introduction to Web Mining

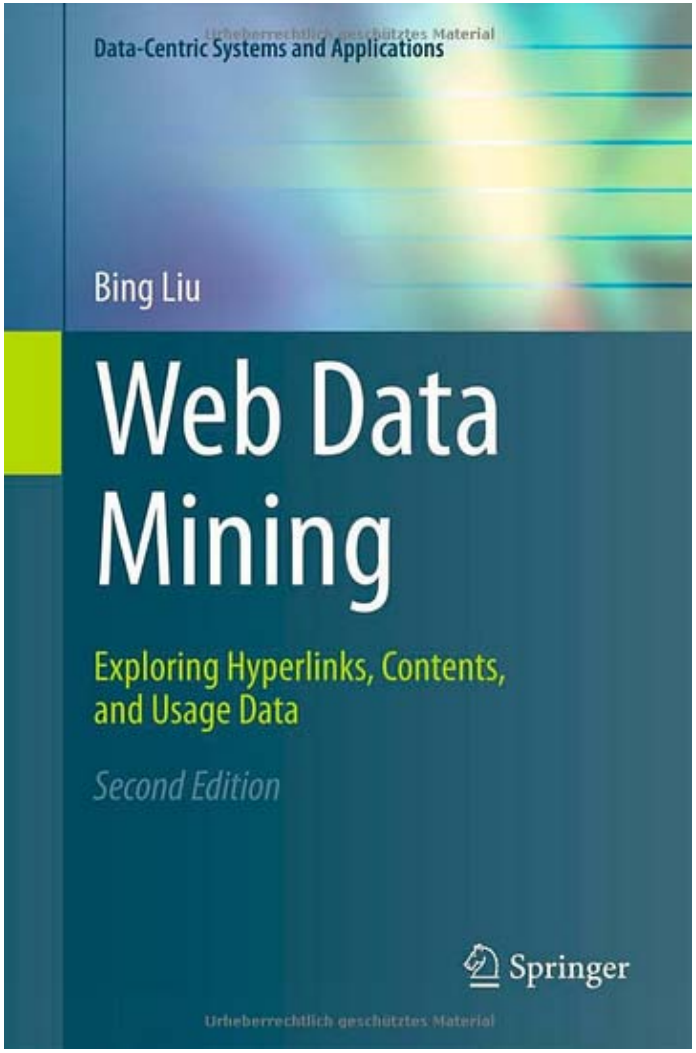
- Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data
- Web Mining and Social Networking
- Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites
- Text Mining: Applications and Theory
- Search Engines – Information Retrieval in Practice

ACM Categories and Subject Descriptors

- I.2.7 [Artificial Intelligence]
  - Natural Language Processing
    - Text analysis
- H.2.8 [Database Management]
  - Database Applications
    - Data mining

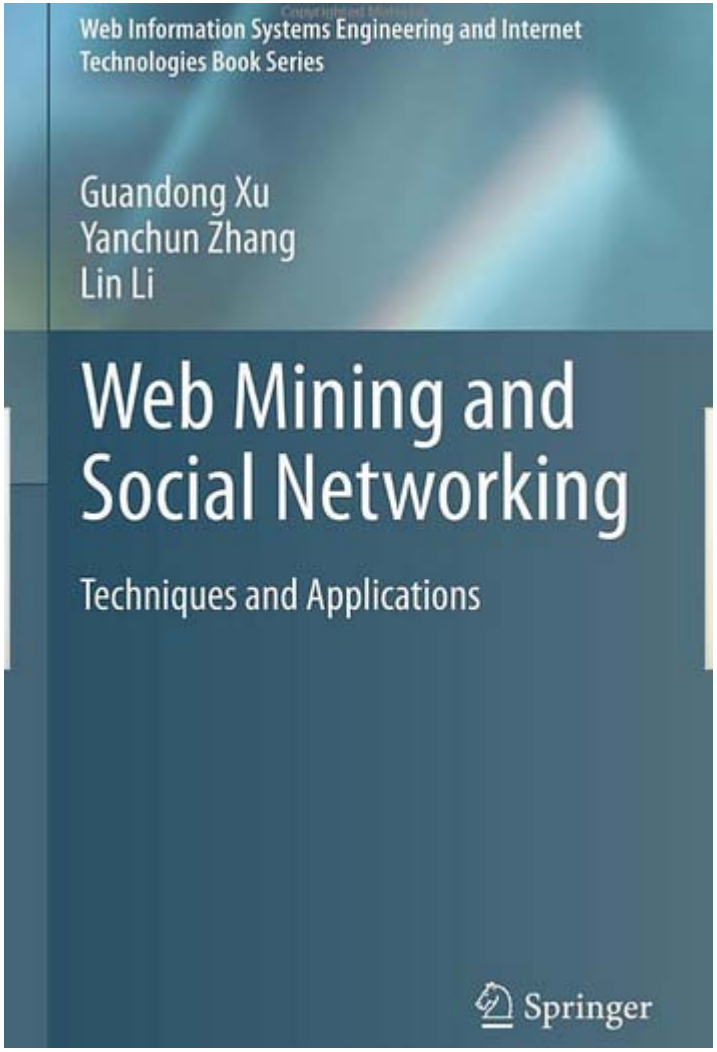
Web Data Mining:

Exploring Hyperlinks, Contents, and Usage Data



Source: http://www.amazon.com/Web-Data-Mining-Data-Centric-Applications/dp/3642194591

Web Mining and Social Networking

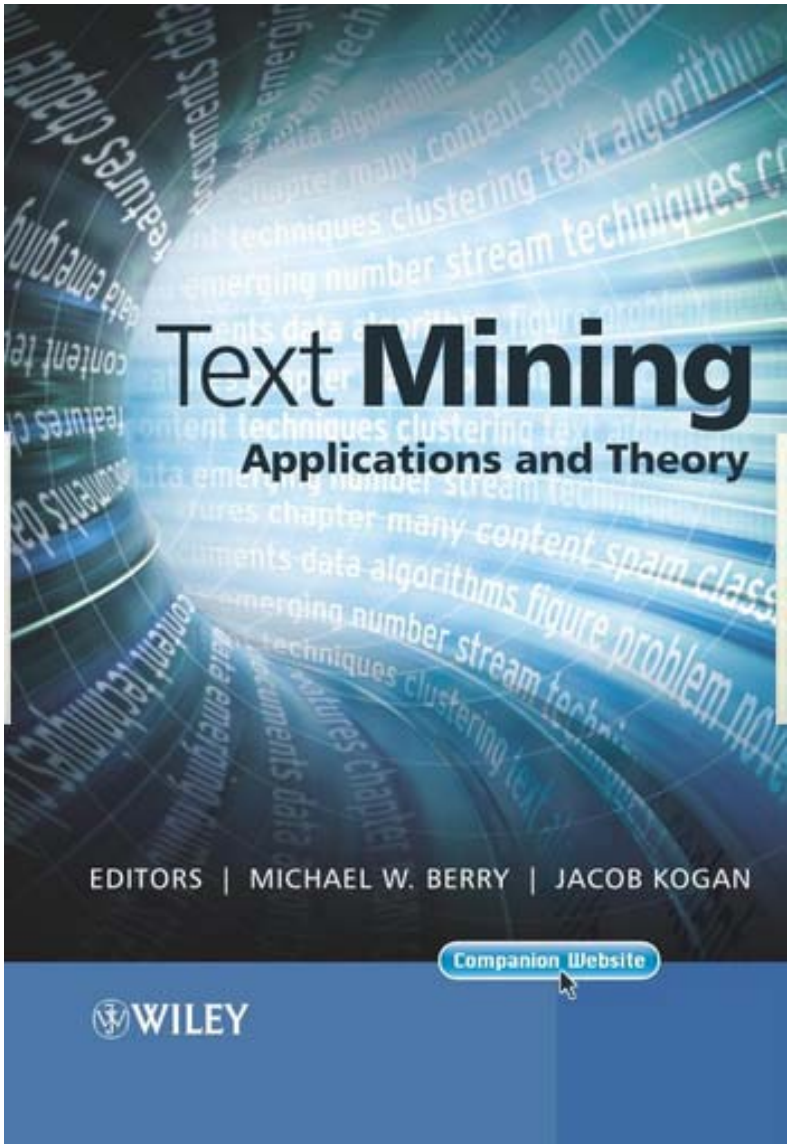


http://www.amazon.com/Web-Mining-Social-Networking-Applications/dp/1441977341

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

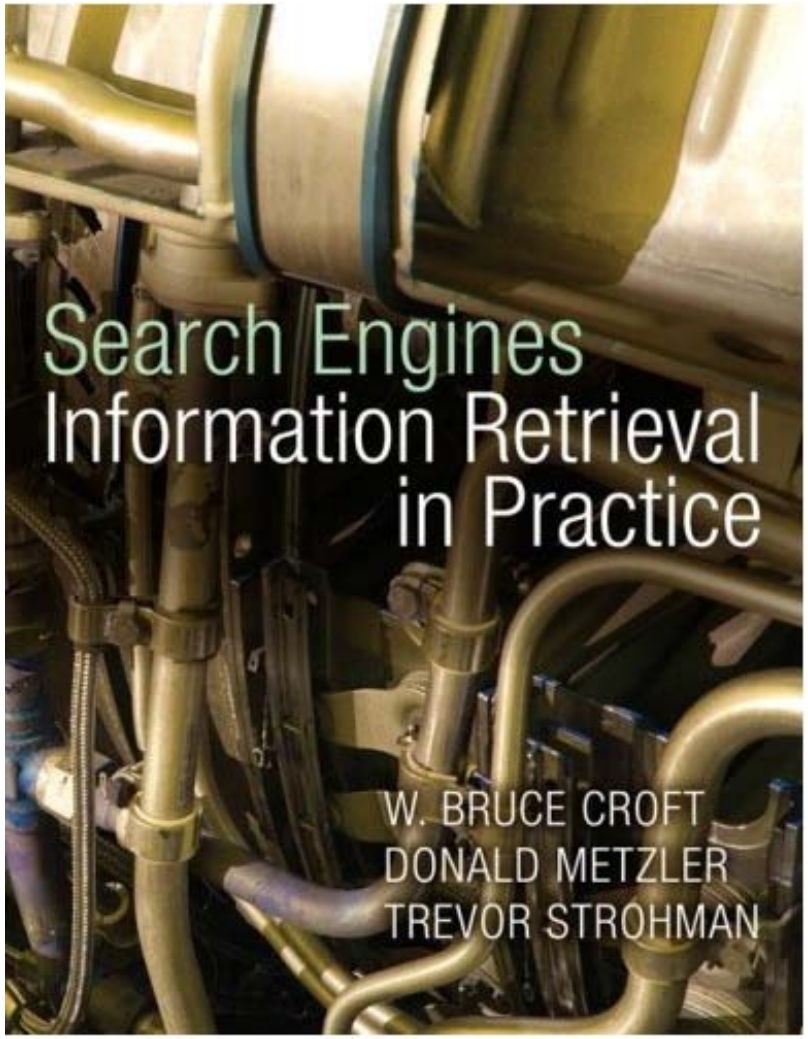
http://www.amazon.com/Mining-Social-Web-Analyzing-Facebook/dp/1449388345

Text Mining



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

Search Engines: Information Retrieval in Practice



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

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

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.

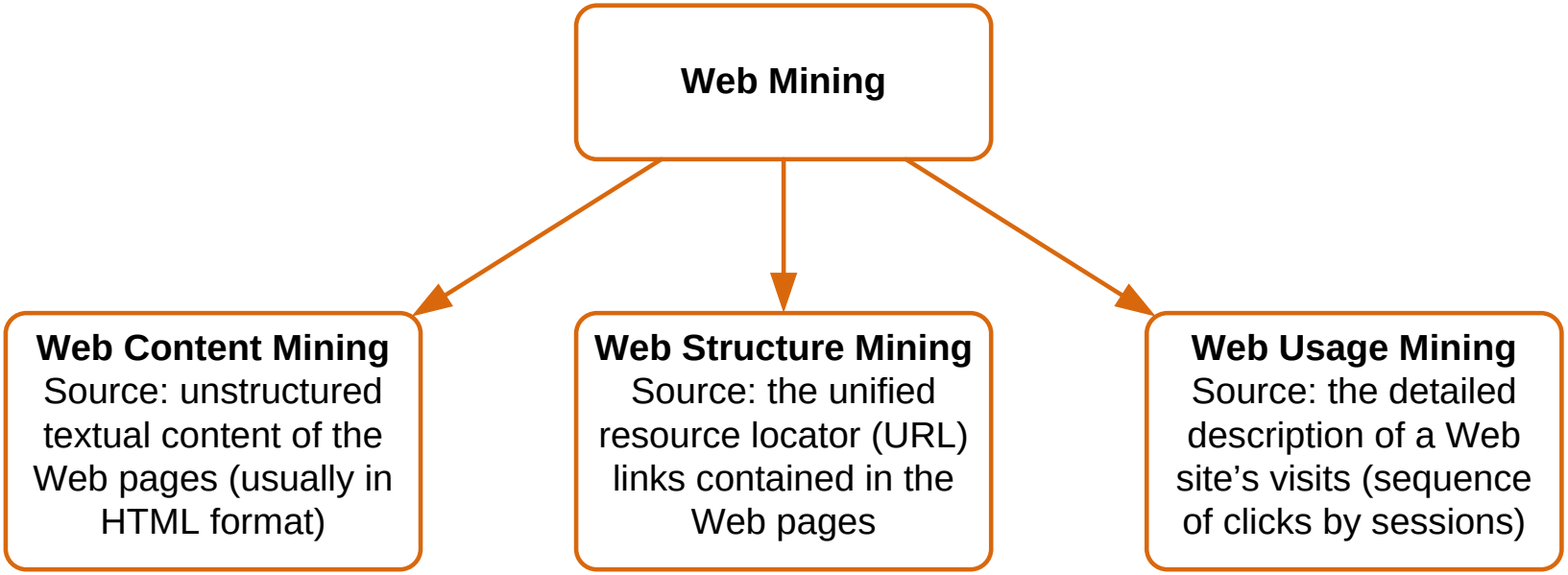
http://en.wikipedia.org/wiki/Text_mining

Web Mining 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

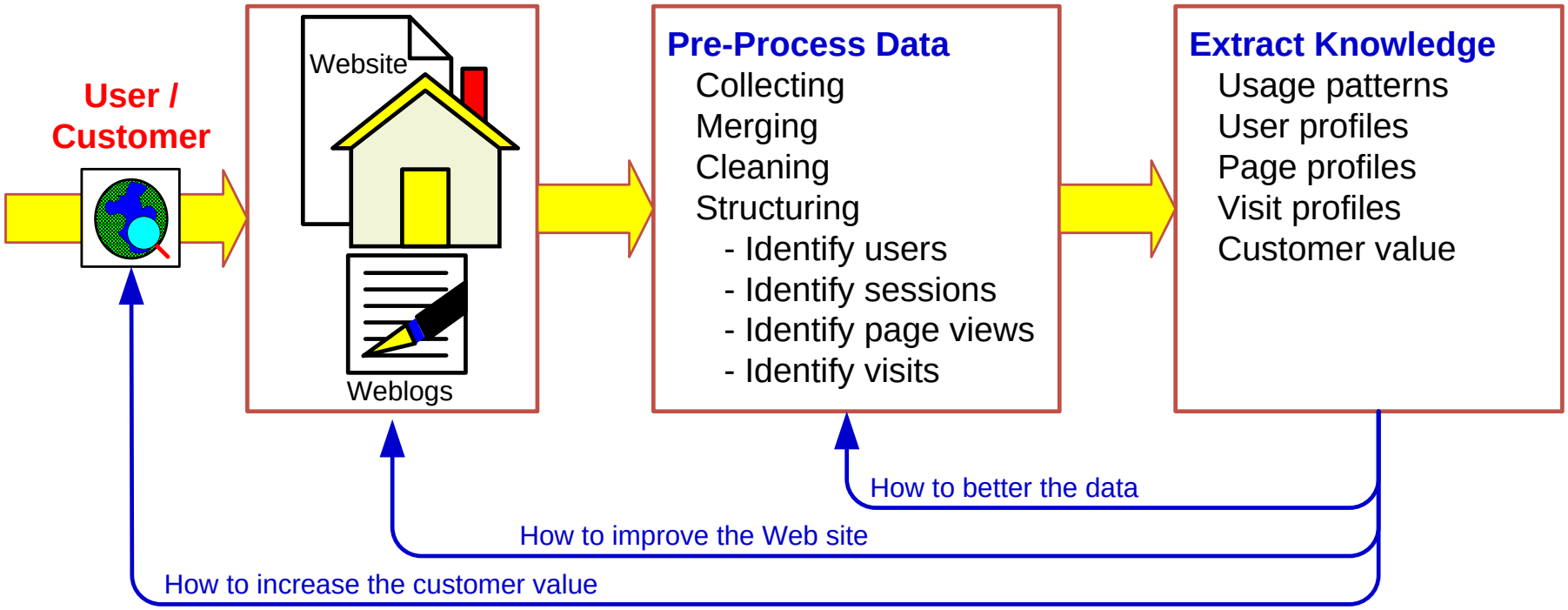
Web Usage Mining

- 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…

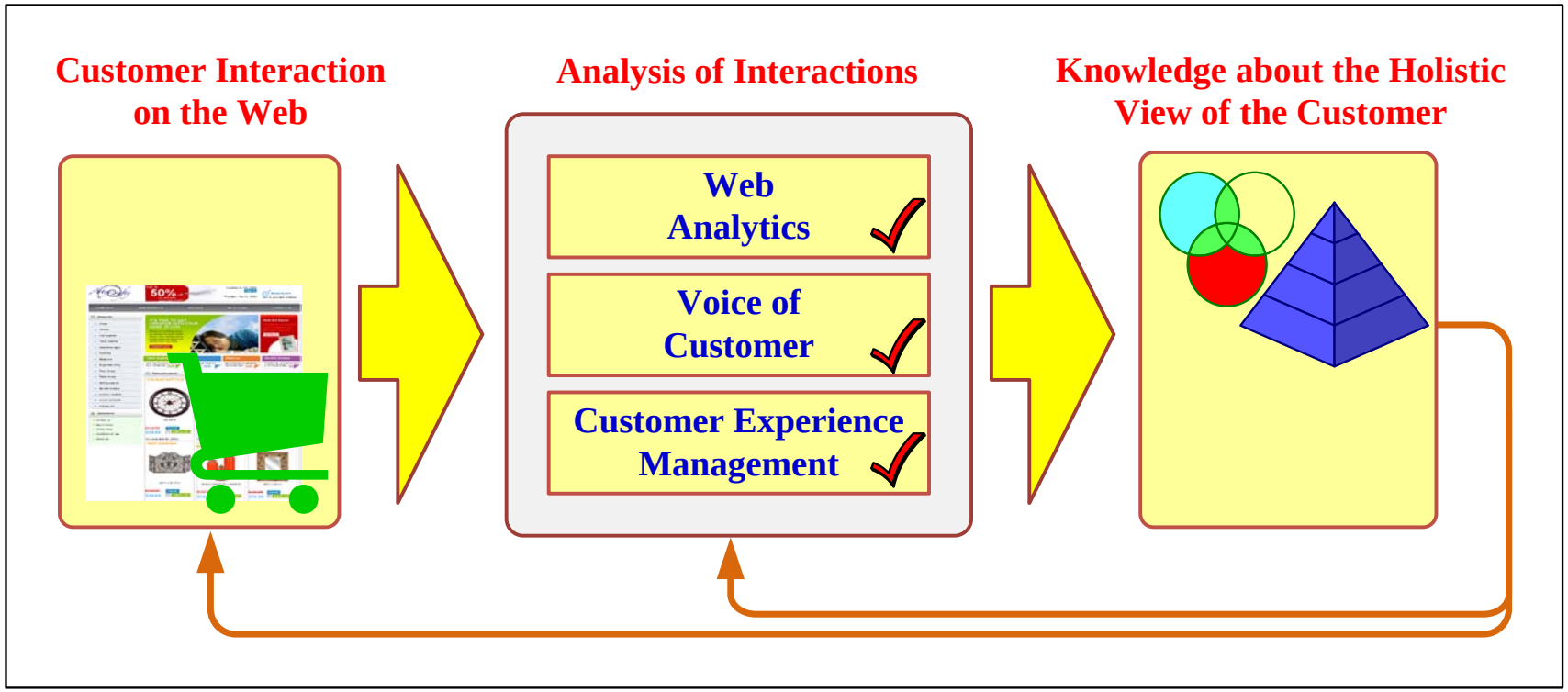
Web Usage Mining
(clickstream analysis)



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

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

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

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)

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

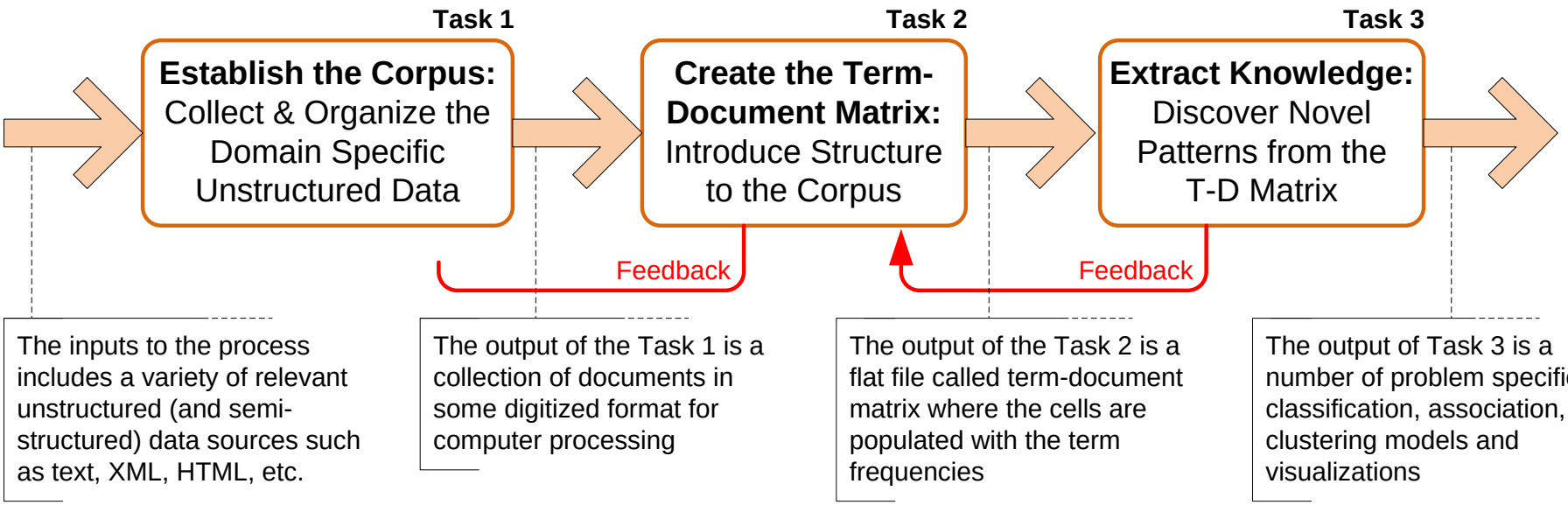
- Information retrieval (IR)
- Information extraction (IE)
- Named-entity recognition (NER)
- Question answering (QA)
- Automatic summarization
- Natural language generation and understanding (NLU)
- Machine translation (ML)
- Foreign language reading and writing
- Speech recognition
- Text proofing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

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

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 Tools

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

Summary

- This course introduces the fundamental concepts and technology of web mining.
- Topics include
  - Introduction to Web Mining,
  - Association Rules and Sequential Patterns,
  - Supervised Learning,
  - Unsupervised Learning,
  - Partially Supervised Learning,
  - Information Retrieval and Web Search,
  - Social Network Analysis,
  - Web Crawling,
  - Structured Data Extraction,
  - Information Integration,
  - Opinion Mining and Sentiment Analysis, and
  - Web Usage Mining.

References

- Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition.
- Efraim Turban, Ramesh Sharda, Dursun Delen (2011), “Decision Support and Business Intelligence Systems,” Pearson, Ninth Edition.

Contact Information

戴敏育 博士 (Min-Yuh Day, Ph.D.)

專任助理教授
淡江大學 資訊管理學系

電話：02-26215656 #2347
傳真：02-26209737
研究室：i716 (覺生綜合大樓)
地址： 25137 新北市淡水區英專路151號
Email： myday@mail.tku.edu.tw
網址：http://mail.tku.edu.tw/myday/