

Data Warehousing

資料倉儲

Introduction to Data Warehousing

1001DW01

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

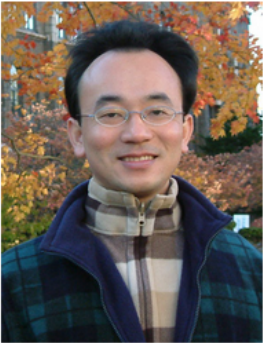
淡江大學 資訊管理學系

<http://mail.im.tku.edu.tw/~myday/>

2011-09-06

Min-Yuh Day

mail.im.tku.edu.tw/~myday/



Min-Yuh Day

Assistant Professor
[Department of Information Management, Tamkang University](#)

Tel: 886-2-26215656 ext. 2347
Fax: 886-2-26209737
Office: I716, Chueh-sheng Memorial Hall [\[Office Hours\]](#)
Address: No.151, Yingzhuan Rd., Danshui Dist., New Taipei City 25137, Taiwan (R.O.C.)
Email: myday@mail.tku.edu.tw
Web: <http://mail.im.tku.edu.tw/~myday/>

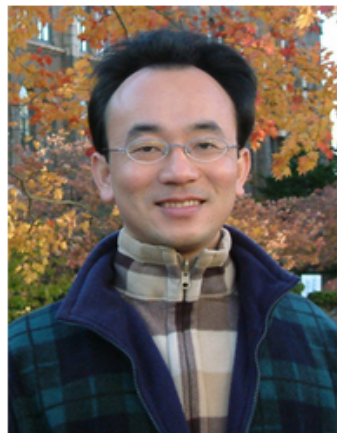
[\[Vita | Education | Research | Publications | Teaching | Professional Activities \]](#) [\[中文版\]](#)

Vita:

Dr. Min-Yuh Day is an Assistant Professor in the Department of Information Management at Tamkang University, Taiwan. Prior to joining the faculty at TKU in 2011, he was a Postdoctoral Fellow in the [Intelligent Agent Systems Lab](#), [Institute of Information Science](#), [Academia Sinica](#), Taiwan. He received the Ph.D. degree from the [Department of Information Management](#) at [National Taiwan University](#), Taiwan. He received his MBA in Management Information System from [Tamkang University](#), Taiwan. His current research interests include Knowledge Management, Electronic Commerce, Information Systems Evaluation, Social Media Service, Question Answering Systems, Data Mining and Text Mining, and Biomedical Informatics. He has published papers in *Information & Management*, *Decision Support Systems*, *Integrated Computer-Aided Engineering*, *ACM Transactions on Asian Language Information Processing*, and a number of international conference proceedings.

Education:

- Ph.D. Department of Information Management, National Taiwan University, 2001-2010
Dissertation: A Study of Evaluation Model of User Satisfaction with Social Network Services
Advisor: Dr. Chorng-Shyong Ong
- M.B.A. Department of Information Management, Tamkang University, 1993-1995
Thesis: Research of Applying Genetic Algorithms to Fuzzy Forecasting - Focus on Sales Forecasting
Advisor: Dr. Hung-Chang Lee
- B.B.A. Department of Information Management, Tamkang University, 1989-1993



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) [Office Hours]

地址：25137 新北市淡水區英專路151號

Email：myday@mail.tku.edu.tw

網址：<http://mail.im.tku.edu.tw/~myday/>

[簡介 | 教育 | 研究 | 論文發表 | 教學 | 學術活動] [English Version]

簡介 (Vita):

戴敏育博士目前是淡江大學資管系專任助理教授。他於2011年加入淡江大學專任教師之前，曾任職於中央研究院資訊科學研究所智慧型代理人系統實驗室博士後研究員。他於2010年取得國立台灣大學資訊管理博士學位，他在淡江大學資訊管理學系取得碩士學位。他目前的研究興趣包括知識管理 (Knowledge Management)、電子商務 (Electronic Commerce)、資訊系統評量 (Information Systems Evaluation)、社會媒體服務 (Social Media Service)、問答系統 (Question Answering Systems)、資料與文字探勘 (Data Mining and Text Mining)、生物醫學資訊 (Biomedical Informatics)。他的學術研究論文已發表在 *Information & Management*, *Decision Support Systems*, *Integrated Computer-Aided Engineering*, *ACM Transactions on Asian Language Information Processing* 等國際期刊和許多國際研討會論文集。

教育 (Education):

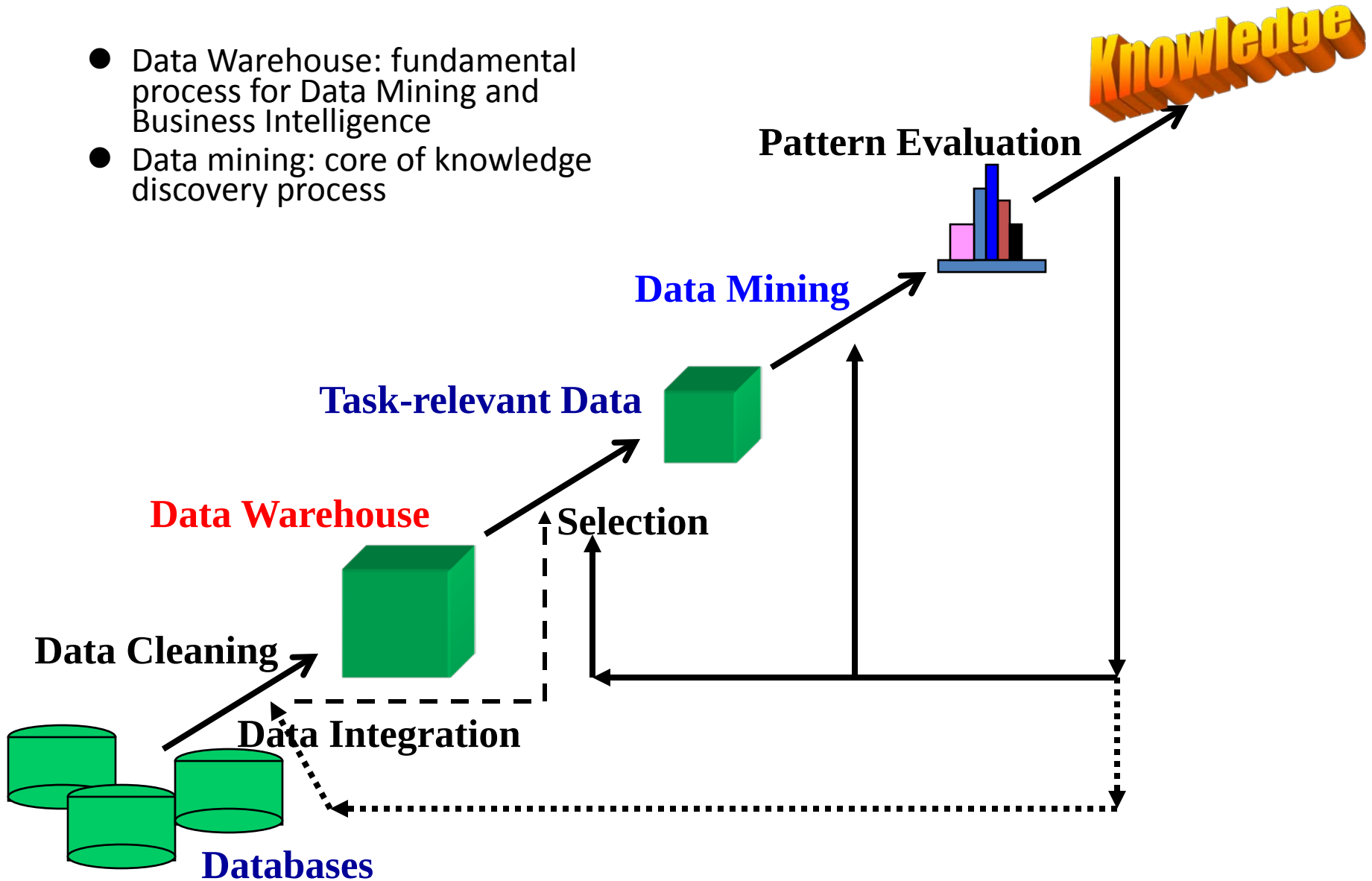
- 博士 國立台灣大學資訊管理研究所 (2001-2010)
博士論文：社交網路服務使用者滿意度評量模式之研究
- 碩士 淡江大學資訊管理研究所 (1993-1995)
碩士論文：應用遺傳演算法發展模糊預測之研究－以銷售預測為例
- 學士 淡江大學資訊管理學系 (1989-1993)

課程資訊

- 課程名稱：資料倉儲 (Data Warehousing)
- 授課教師：戴敏育 (Min-Yuh Day)
- 開課系級：資管四 (MI4)
- 開課資料：選修 單學期 2學分
- 上課時間：週二 6, 7 (Tue 13:10-15:00)
- 上課教室：B427

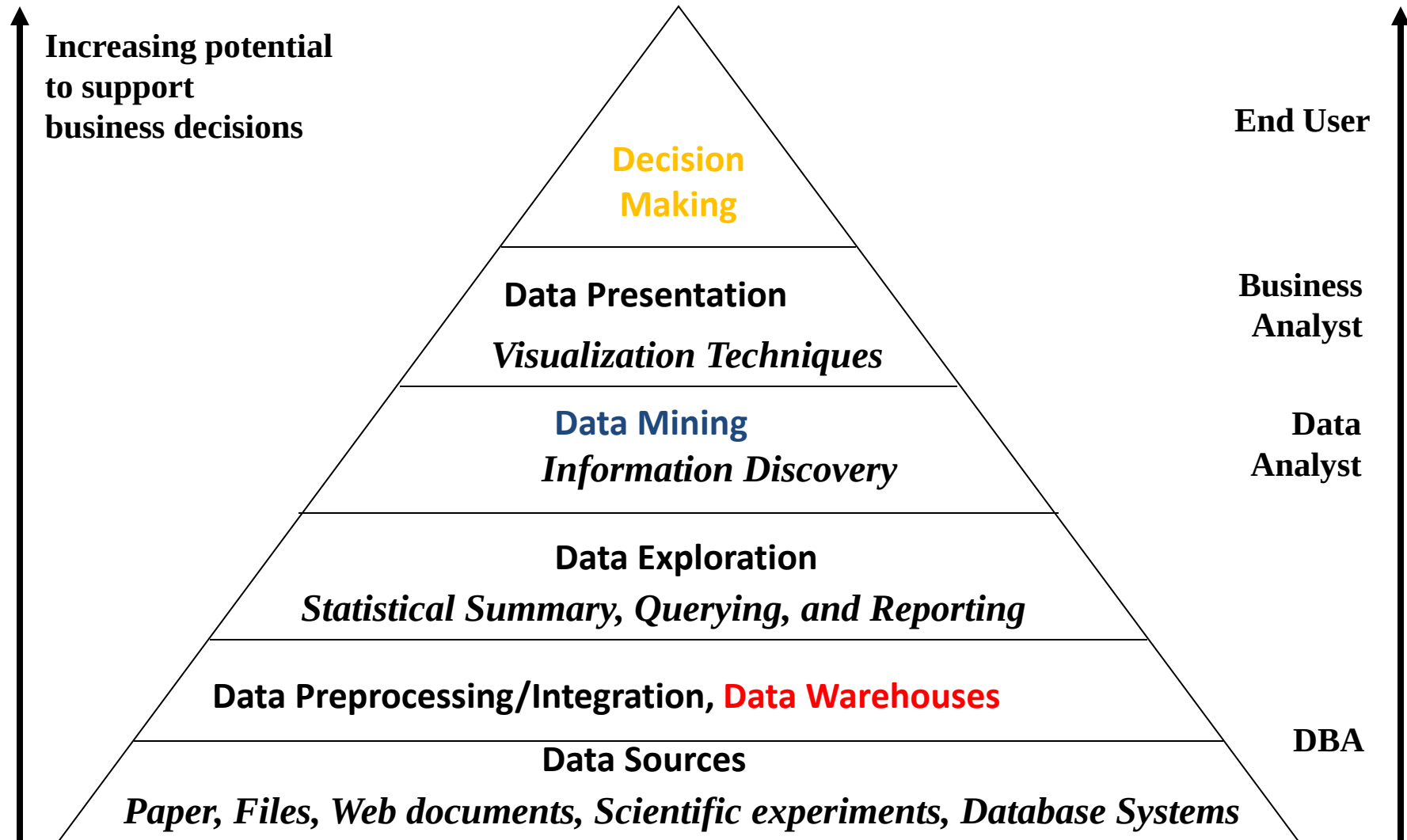
Knowledge Discovery (KDD) Process

- Data Warehouse: fundamental process for Data Mining and Business Intelligence
- Data mining: core of knowledge discovery process



Data Warehouse

Data Mining and Business Intelligence



課程簡介

- 本課程介紹資料倉儲的基本概念及技術。
- 課程內容包括資料倉儲、OLAP、資料探勘、商業智慧、即時分析處理，資料方塊，關聯分析、分類、分群、社會網路分析、文字探勘、與網頁探勘。

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	Classification and Prediction
12	100/11/22	Cluster Analysis
13	100/11/29	Sequence Data Mining
14	100/12/06	Social Network Analysis
15	100/12/13	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

Course Introduction

- This course introduces the fundamental concepts and technology of data warehousing.
- Topics include data warehousing, data mining, business intelligence, OLAP, data cube, association analysis, classification, cluster analysis, social network analysis, text mining, and web mining.

Objective

- Students will be able to understand and apply the fundamental concepts and technology of data warehousing.

教學目標之教學策略與評量方法

- 教學目標
 - － 學生將能夠瞭解及應用資料倉儲的基本概念及技術。
- 教學策略
 - － 課堂講授、分組討論
- 評量方法
 - － 出席率、報告、討論、期中考、期末考

教材課本

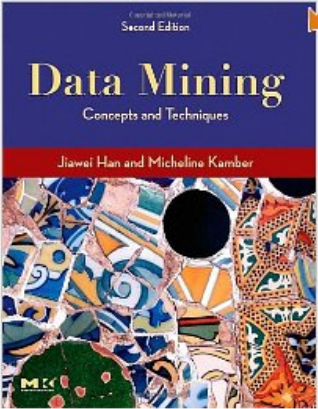
- Data Mining: Concepts and Techniques, Second Edition, Jiawei Han and Micheline Kamber, 2006, Elsevier
- 參考書籍
 - 資料探勘：概念與方法，王派洲 譯，2008，滄海
 - SQL Server 2008 R2 資料採礦與商業智慧，謝邦昌、鄭宇庭、蘇志雄，2011，基峯
 - 資料庫理論與實務SQL Server 2008，施威銘研究室，2010，旗標
 - Web 資料採掘技術經典，孫惠民，2008，松崗

Data Mining: Concepts and Techniques (Second Edition)

Amazon.com: Data Min... x

www.amazon.com/Data-Mining-Concepts-Techniques-Management/dp/1558609016

Click to **LOOK INSIDE!**



Share your own customer images
[Search inside this book](#)

Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems) [Hardcover]
[Jiawei Han](#) (Author), [Micheline Kamber](#) (Author), [Jian Pei](#) (Author)
★ ★ ★ ★ ☆ (31 customer reviews) Like (0)

List Price: ~~\$68.95~~
Price: **\$53.09** & this item ships for **FREE with Super Saver Shipping**. [Details](#)
You Save: **\$15.86 (23%)**
[Special Offers Available](#)

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Wednesday, February 16? Order it in the next **18 hours and 54 minutes**, and choose **One-Day Shipping** at checkout. [Details](#)

22 new from \$51.71 **30 used** from \$30.00

 **FREE Two-Day Shipping for Students.** [Learn more](#)

Formats	Amazon Price	New from	Used from
 Hardcover	\$53.09	\$51.71	\$30.00
 Paperback	--	--	\$50.00

 [Show 4 more formats](#)

Buy New
Quantity: 1
 **Add to Cart**
or
[Sign in](#) to turn on 1-Click ordering.
or
 **Add to Cart with FREE Two-Day Shipping**
Amazon Prime Free Trial required. Sign up when you check out.
[Learn More](#)
Add to Wish List

Buy Used
Used - Good [See details](#)
\$34.99 & this item ships for **FREE with Super Saver Shipping**.
[Details](#)
Fulfilled by Amazon
 **Add to Cart**
or
[Sign in](#) to turn on 1-Click ordering.

More Buying Choices
52 used & new from **\$30.00**

 **Sell This Book Back for \$12.19**
Whether you buy it used on Amazon for [\\$30.00](#) or somewhere else, you can sell it back to our Textbook Buyback Store at the current price of [\\$12.19](#). [Restrictions Apply](#)

Used Price \$30.00
Buyback Price **\$12.19**
Price after Buyback **\$17.81**

<http://www.amazon.com/Data-Mining-Concepts-Techniques-Management/dp/1558609016>

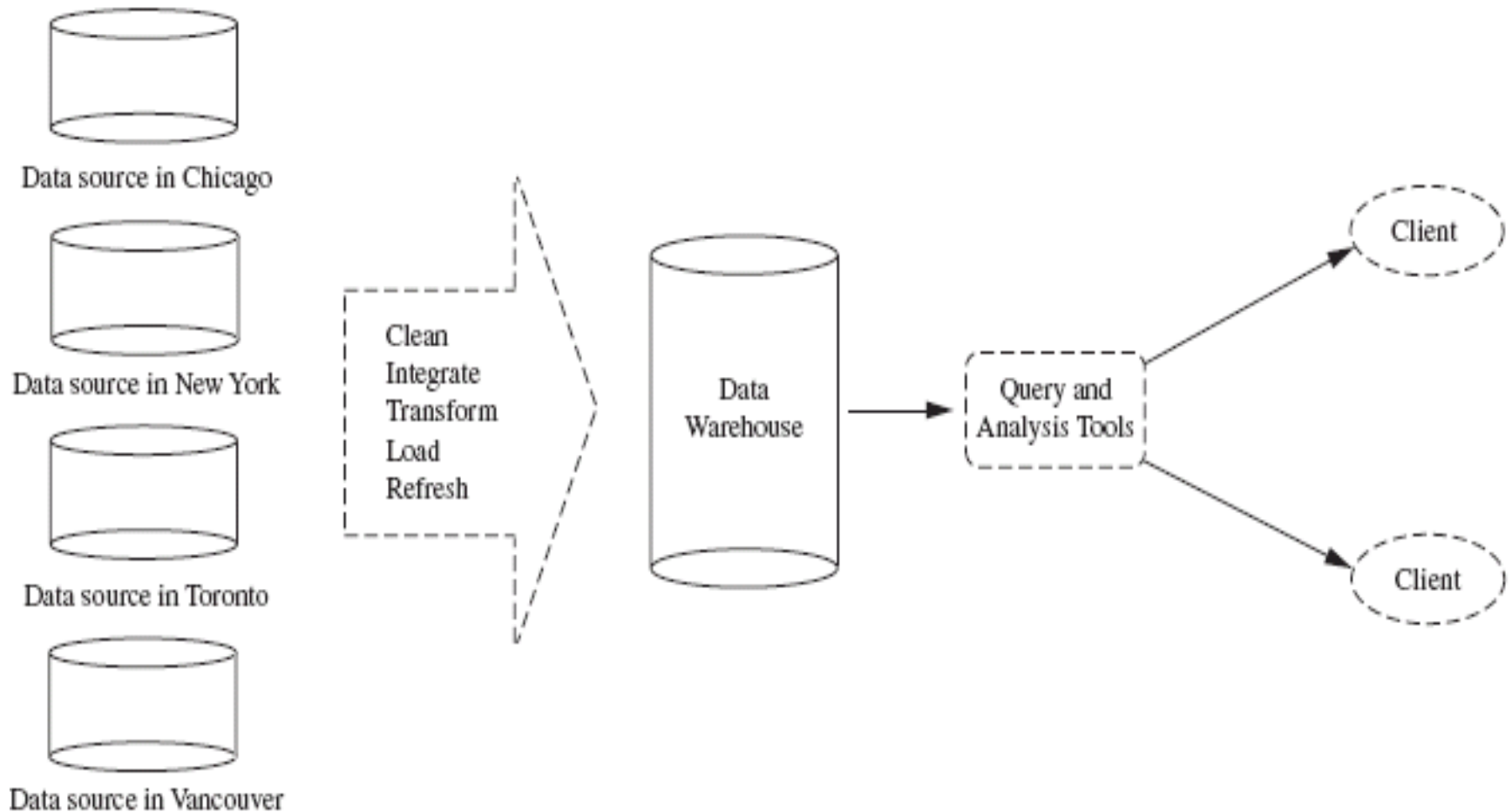
作業與學期成績計算方式

- 批改作業篇數
 - 1篇（Team Term Project）
- 學期成績計算方式
 - ☒ 期中考成績：30 %
 - ☒ 期末考成績：30 %
 - ☒ 作業成績：20 %（Team Term Project）
 - ☒ 其他（課堂參與及報告討論表現）：20 %

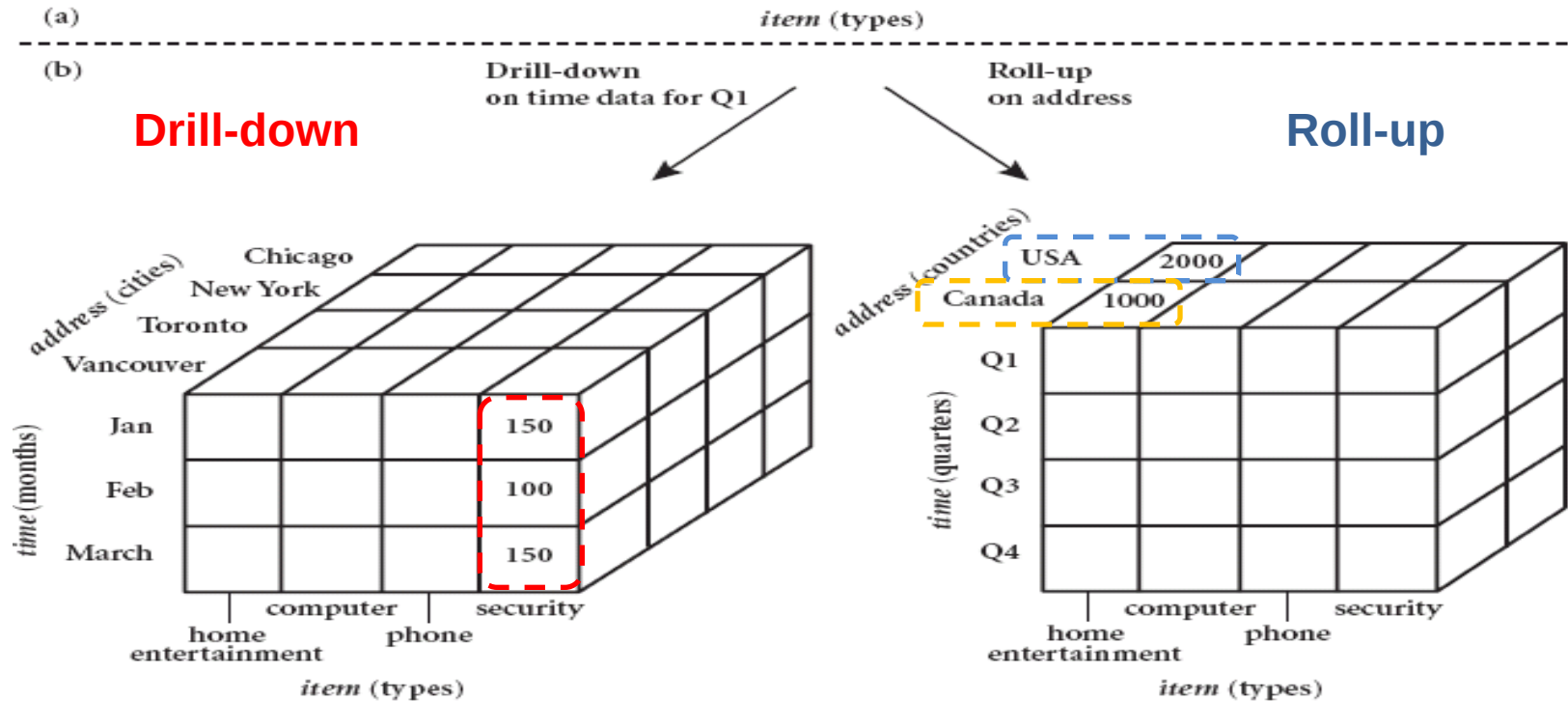
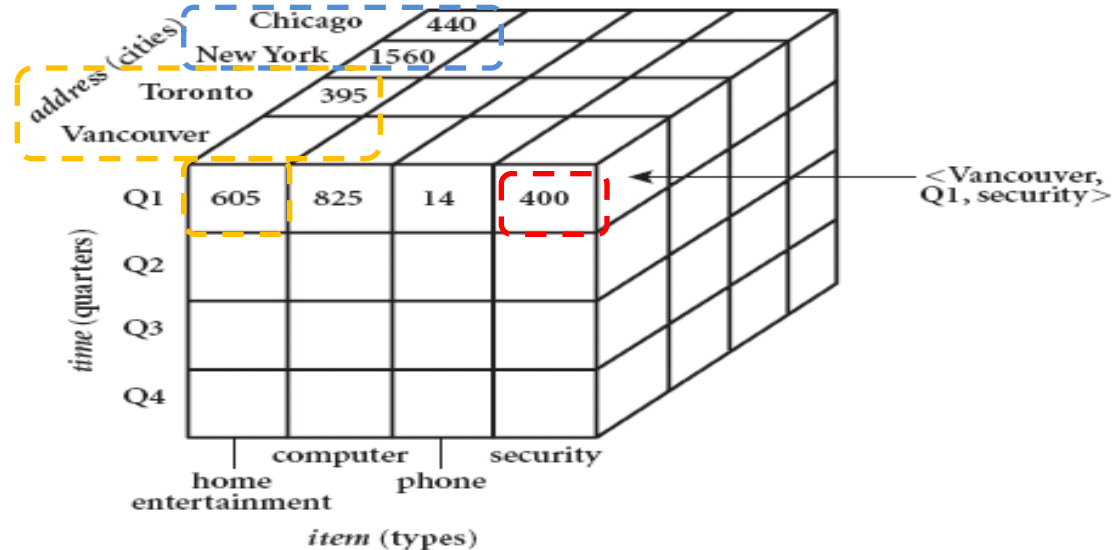
Team Term Project

- Term Project Topics
 - Data Warehousing
 - Business Intelligence
 - Data mining
 - Text mining
 - Web mining
 - Social Network Analysis
 - Link Mining
- 3-5 人為一組
 - 分組名單於 2011.09.20 (二) 課程下課時繳交
 - 由班代統一收集協調分組名單

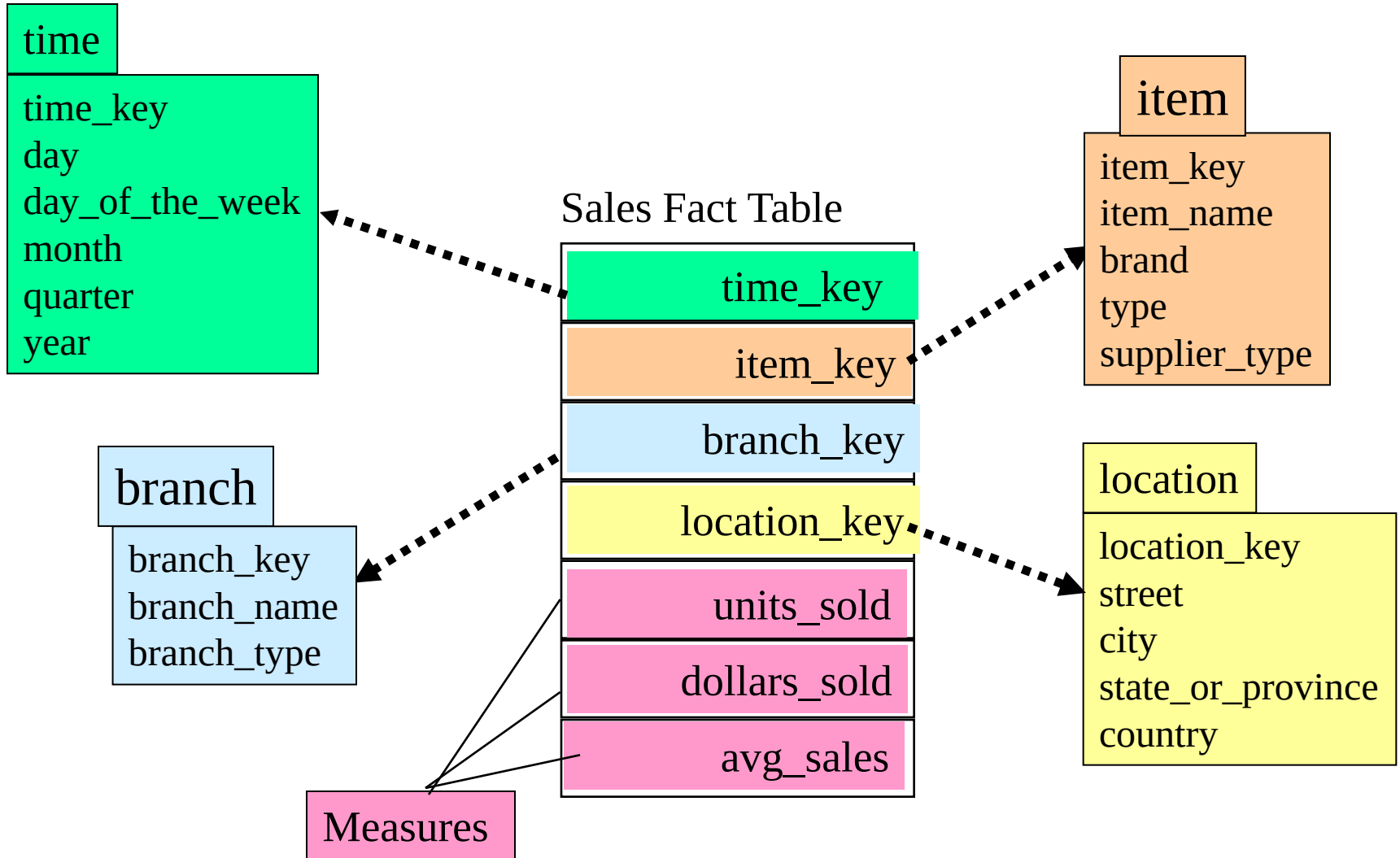
Typical framework of a data warehouse



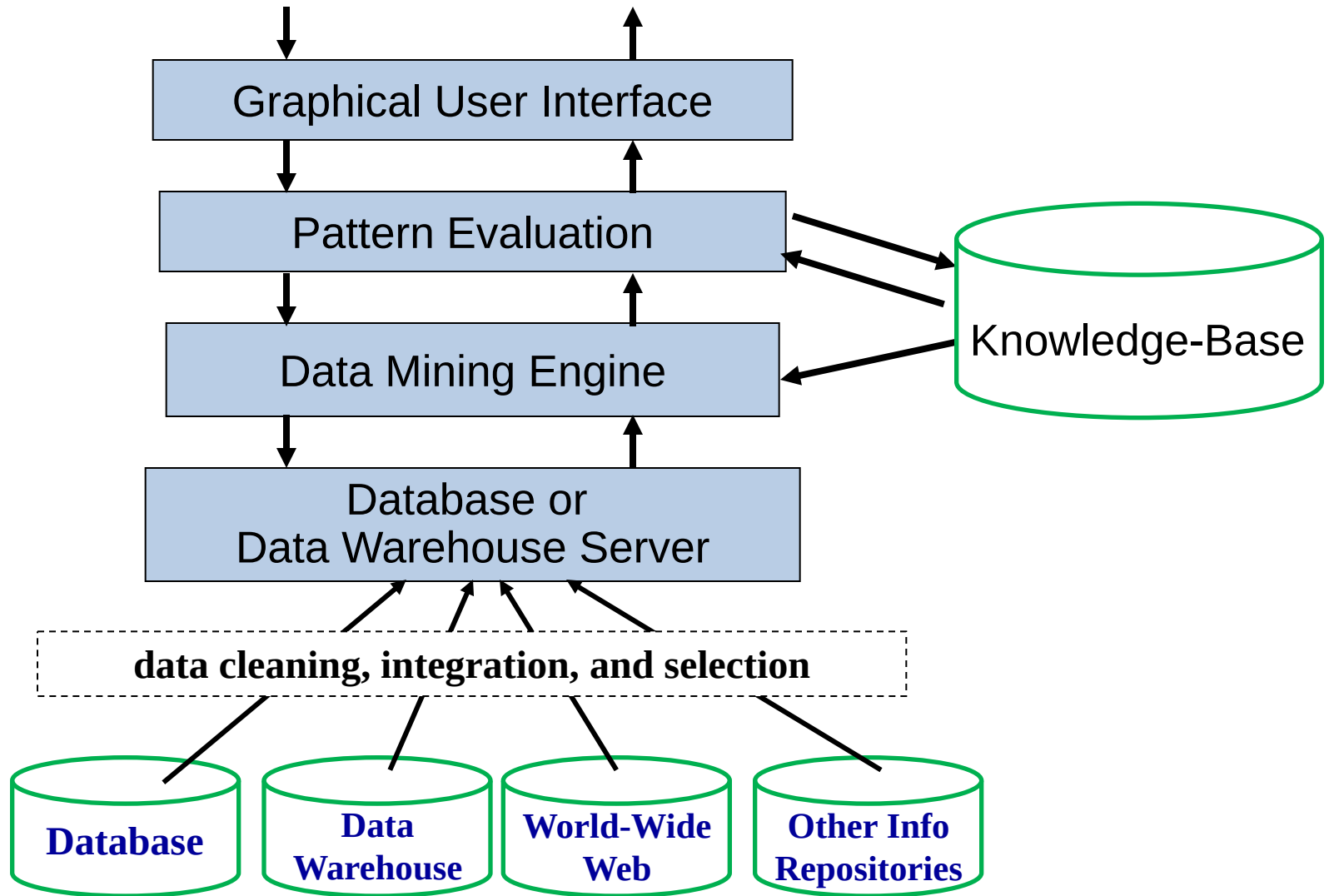
Multidimensional data cube for data warehousing



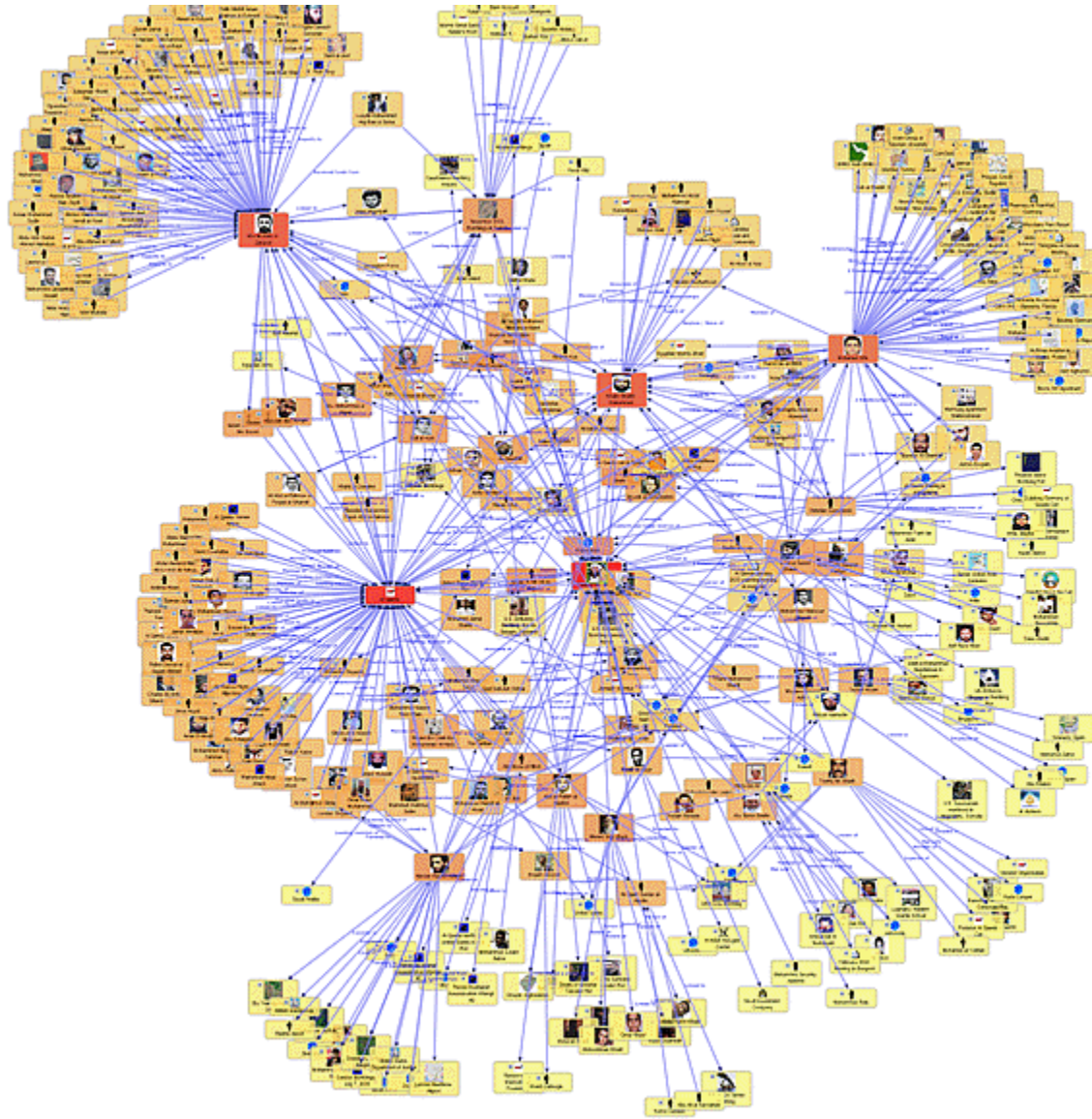
Example of Star Schema



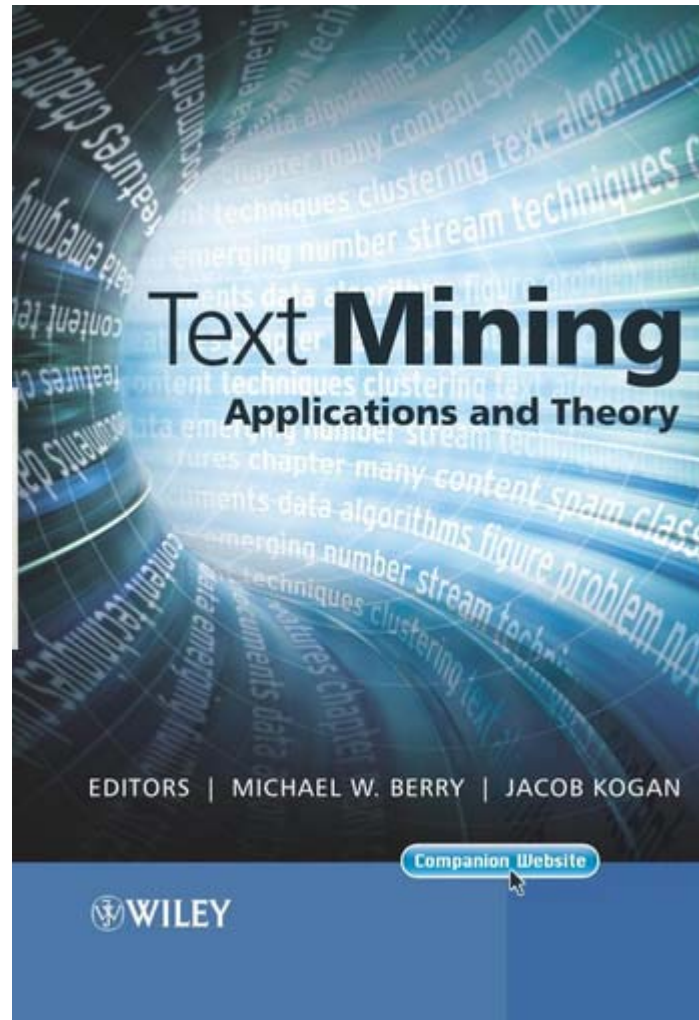
Architecture of a typical data mining system



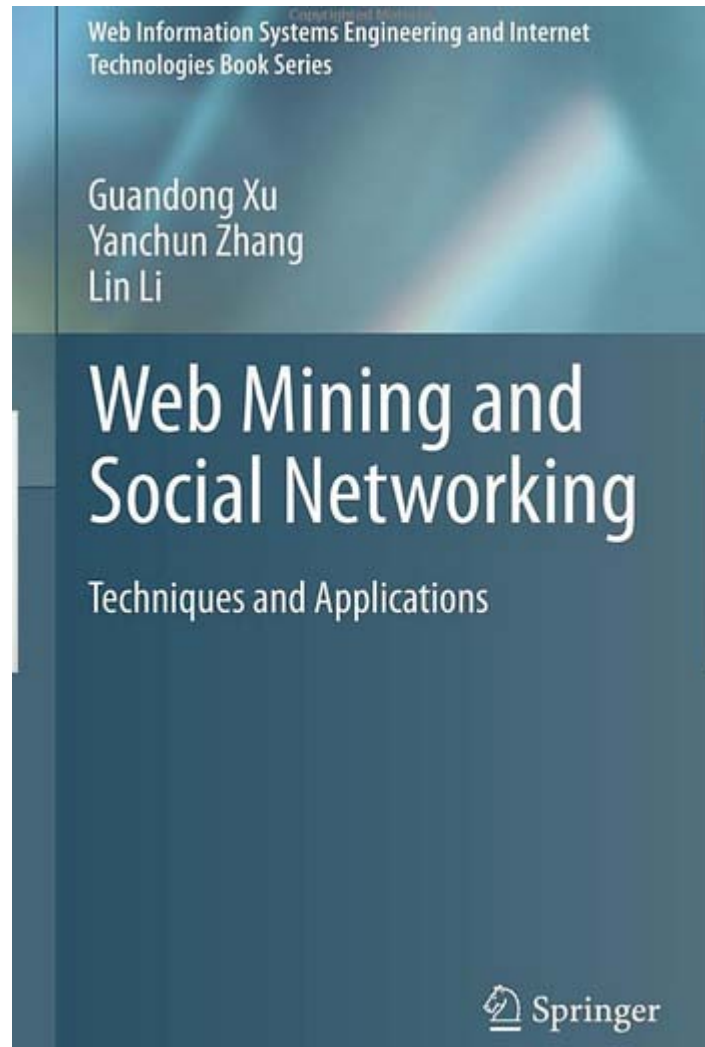
Social Network Analysis



Text Mining



Web Mining and Social Networking



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

*Analyzing Data from Facebook, Twitter, LinkedIn,
and Other Social Media Sites*

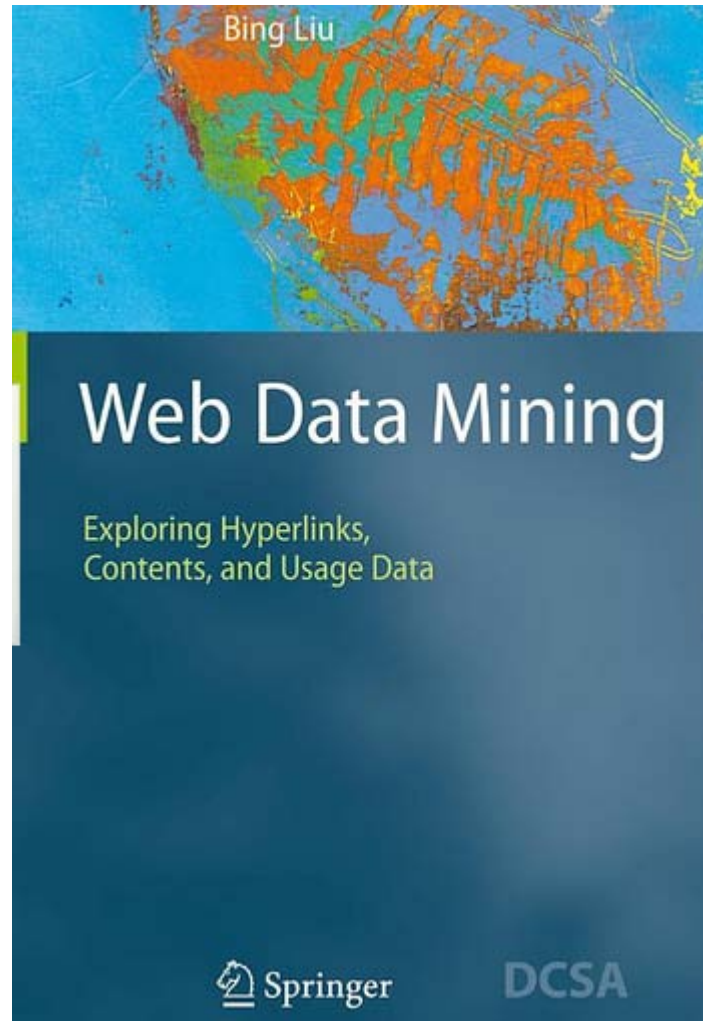


O'REILLY®

Matthew A. Russell

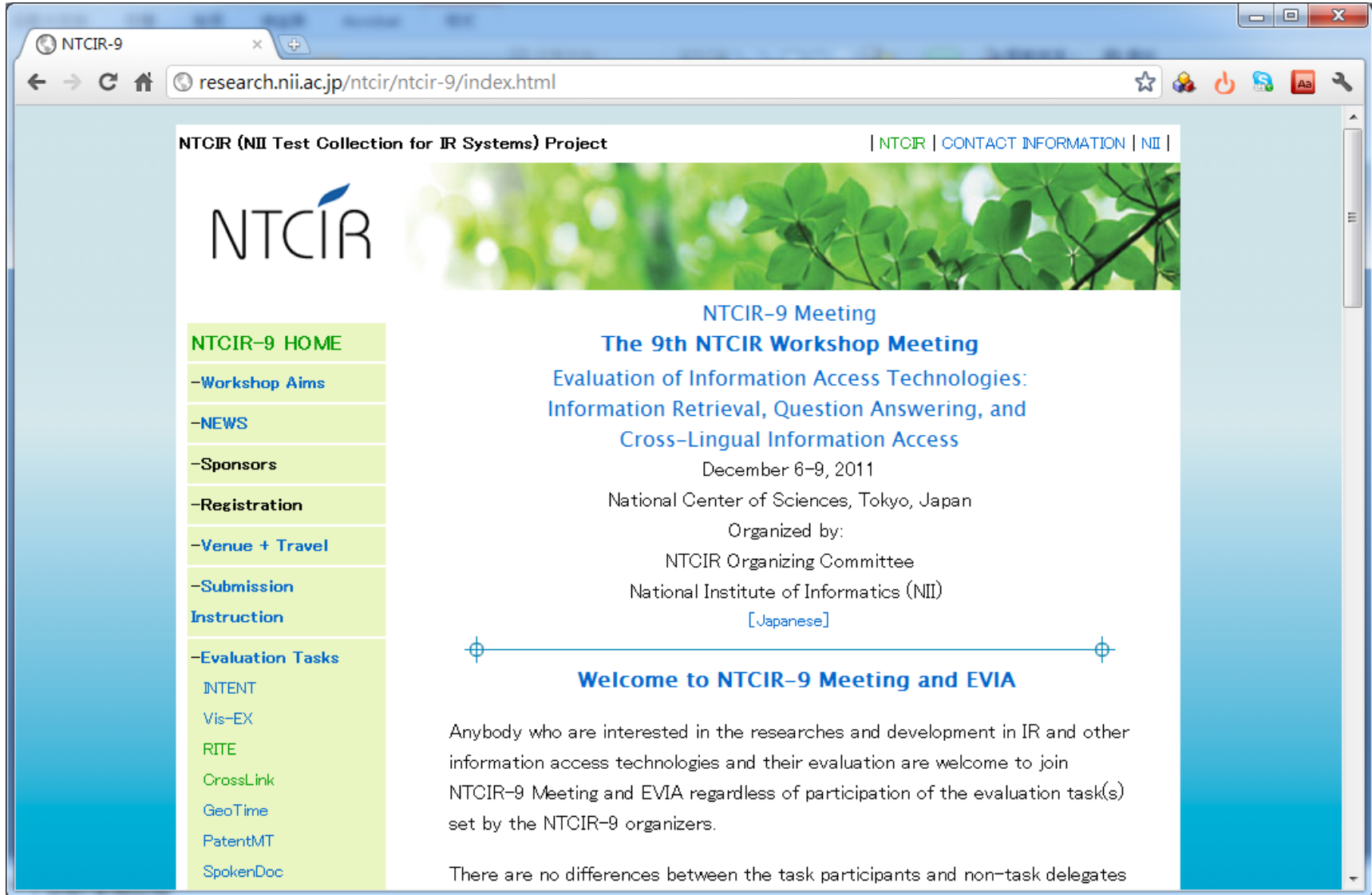
Web Data Mining:

Exploring Hyperlinks, Contents, and Usage Data



NTCIR Project

(NII Test Collection for IR Systems)

A screenshot of a web browser displaying the NTCIR-9 website. The browser's address bar shows the URL "research.nii.ac.jp/ntcir/ntcir-9/index.html". The website has a light blue header with the title "NTCIR (NII Test Collection for IR Systems) Project" and navigation links for "NTCIR", "CONTACT INFORMATION", and "NII". Below the header is a banner image of green leaves. On the left side, there is a vertical menu with links: "NTCIR-9 HOME", "-Workshop Aims", "-NEWS", "-Sponsors", "-Registration", "-Venue + Travel", "-Submission", "Instruction", "-Evaluation Tasks", "INTENT", "Vis-EX", "RITE", "CrossLink", "GeoTime", "PatentMT", and "SpokenDoc". The main content area features the "NTCIR-9 Meeting" title, followed by "The 9th NTCIR Workshop Meeting" and the subtitle "Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access". It specifies the dates "December 6-9, 2011" and the location "National Center of Sciences, Tokyo, Japan". It also mentions the organizers: "Organized by: NTCIR Organizing Committee" and "National Institute of Informatics (NII)", with a link for "[Japanese]". A horizontal line with a crosshair at each end separates this section from the next. The next section is titled "Welcome to NTCIR-9 Meeting and EVIA" and contains a paragraph: "Anybody who are interested in the researches and development in IR and other information access technologies and their evaluation are welcome to join NTCIR-9 Meeting and EVIA regardless of participation of the evaluation task(s) set by the NTCIR-9 organizers." Below this, it states: "There are no differences between the task participants and non-task delegates".

NTCIR-9

research.nii.ac.jp/ntcir/ntcir-9/index.html

NTCIR (NII Test Collection for IR Systems) Project | NTCIR | CONTACT INFORMATION | NII |

NTCIR

NTCIR-9 HOME

- Workshop Aims
- NEWS
- Sponsors
- Registration
- Venue + Travel
- Submission
- Instruction
- Evaluation Tasks
 - INTENT
 - Vis-EX
 - RITE
 - CrossLink
 - GeoTime
 - PatentMT
 - SpokenDoc

NTCIR-9 Meeting

The 9th NTCIR Workshop Meeting

Evaluation of Information Access Technologies:
Information Retrieval, Question Answering, and
Cross-Lingual Information Access

December 6-9, 2011

National Center of Sciences, Tokyo, Japan

Organized by:
NTCIR Organizing Committee
National Institute of Informatics (NII)

[Japanese]

Welcome to NTCIR-9 Meeting and EVIA

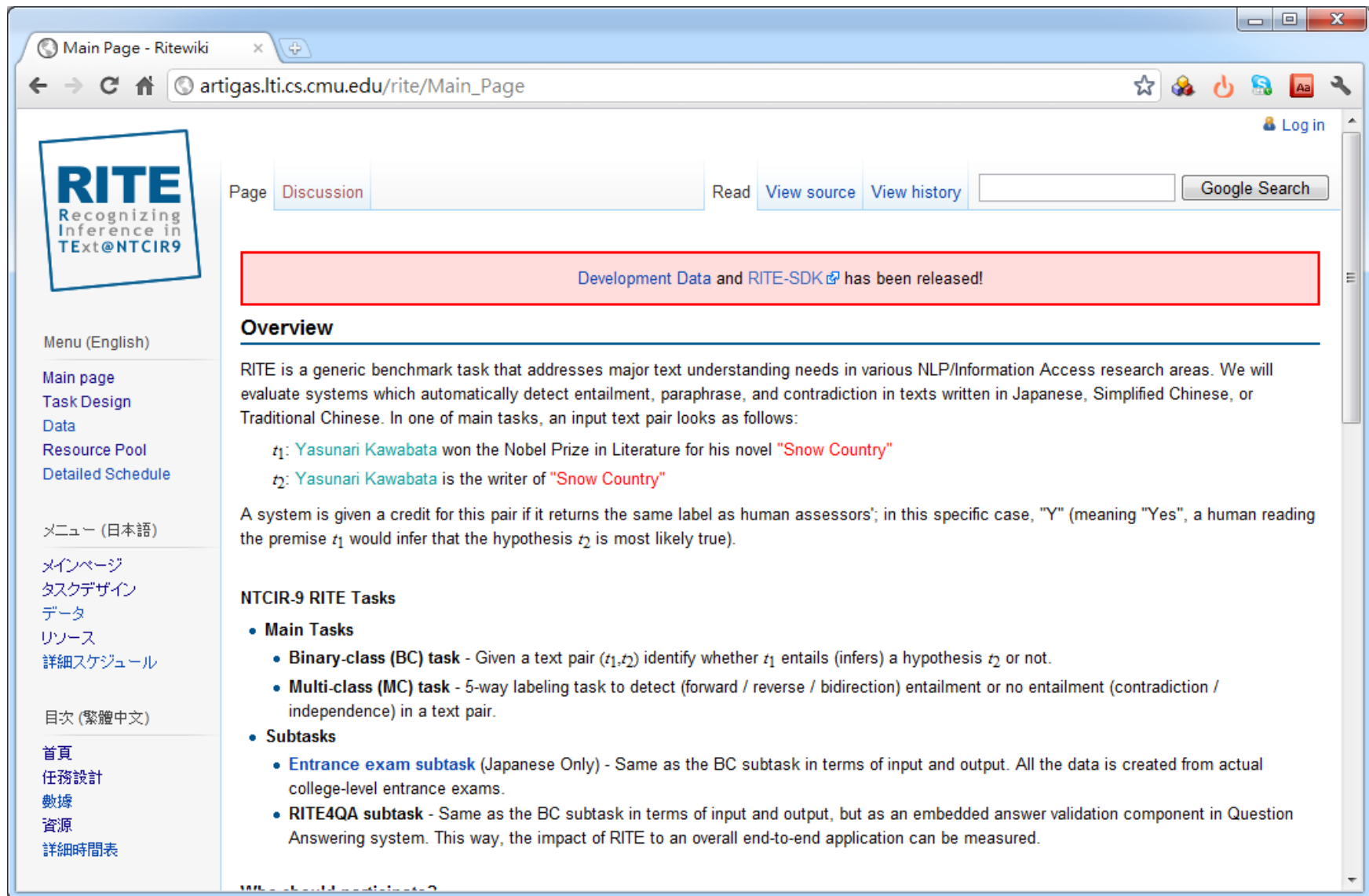
Anybody who are interested in the researches and development in IR and other information access technologies and their evaluation are welcome to join NTCIR-9 Meeting and EVIA regardless of participation of the evaluation task(s) set by the NTCIR-9 organizers.

There are no differences between the task participants and non-task delegates

Source: <http://research.nii.ac.jp/ntcir/ntcir-9/index.html>

NTCIR-9 RITE

Recognizing Inference in TExt @NTCIR9



The screenshot shows a web browser window with the URL artigas.lti.cs.cmu.edu/rite/Main_Page. The page features a sidebar with navigation links in English and Japanese, a main content area with a 'Discussion' tab, and a red banner announcing the release of development data and the RITE-SDK. The 'Overview' section describes the RITE benchmark task, and the 'NTCIR-9 RITE Tasks' section lists the main tasks and subtasks.

RITE
Recognizing Inference in TExt@NTCIR9

Menu (English)
Main page
Task Design
Data
Resource Pool
Detailed Schedule

メニュー (日本語)
メインページ
タスクデザイン
データ
リソース
詳細スケジュール

目次 (繁體中文)
首頁
任務設計
數據
資源
詳細時間表

Page **Discussion** Read View source View history Google Search

Development Data and RITE-SDK has been released!

Overview

RITE is a generic benchmark task that addresses major text understanding needs in various NLP/Information Access research areas. We will evaluate systems which automatically detect entailment, paraphrase, and contradiction in texts written in Japanese, Simplified Chinese, or Traditional Chinese. In one of main tasks, an input text pair looks as follows:

t_1 : Yasunari Kawabata won the Nobel Prize in Literature for his novel "Snow Country"
 t_2 : Yasunari Kawabata is the writer of "Snow Country"

A system is given a credit for this pair if it returns the same label as human assessors'; in this specific case, "Y" (meaning "Yes", a human reading the premise t_1 would infer that the hypothesis t_2 is most likely true).

NTCIR-9 RITE Tasks

- **Main Tasks**
 - **Binary-class (BC) task** - Given a text pair (t_1, t_2) identify whether t_1 entails (infers) a hypothesis t_2 or not.
 - **Multi-class (MC) task** - 5-way labeling task to detect (forward / reverse / bidirection) entailment or no entailment (contradiction / independence) in a text pair.
- **Subtasks**
 - **Entrance exam subtask** (Japanese Only) - Same as the BC subtask in terms of input and output. All the data is created from actual college-level entrance exams.
 - **RITE4QA subtask** - Same as the BC subtask in terms of input and output, but as an embedded answer validation component in Question Answering system. This way, the impact of RITE to an overall end-to-end application can be measured.

NTCIR-9 RITE

Recognizing Inference in Text @NTCIR9

RITE

Recognizing Inference in Text@NTCIR9

Menu (English)

Main page

Task Design

Data

Resource Pool

Detailed Schedule

メニュー (日本語)

メインページ

タスクデザイン

データ

リソース

詳細スケジュール

目次 (繁體中文)

首頁

任務設計

數據

資源

詳細時間表

Page

Discussion

Read

View source

View history

Google Search

Development Data and RITE-SDK has been released!

任務概觀

RITE是一個效能評鑑任務(task)，目的在評鑑系統自動偵測特定語句「關係」的能力。語句間的關係有很多種，比如說蘊涵(entailment)、意譯(paraphrase)以及矛盾(contradiction)等。本次參與RITE評鑑的語言有日文、簡體中文以及繁體中文，每個語言都包含三個「子任務」(subtask)。參與評鑑的系統需辨識給定的兩個文本(text)，輸出二選一或五選一的關係標記(label)：

- **Binary-class (BC)子任務** - 給定一個「文本對」(t_1, t_2)，辨識 t_1 是否能蘊涵或推論出 t_2 這個假設(hypothesis)。
- **Multi-class (MC)子任務** - MC子任務是一個五選一(5-way)的標記任務，需辨識文本對之間是否為正向(forward)蘊涵、逆向(reverse)蘊涵、雙向(bidirection)蘊涵、矛盾(contradiction)或獨立(independence)關係。
- **RITE4QA子任務** - 此子任務的輸入輸出與BC子任務相同，差別在於此任務被特別設計成內嵌於問答系統(Question Answering)，負責答案確認(answer validation)的一個子元件。希望能透過這樣的設定，來度量RITE任務對一個端對端應用的影響。

以下為一個BC子任務的「文本對」範例：

t_1 : 川端康成因其著名小說「雪國」獲得諾貝爾文學獎。

t_2 : 川端康成是「雪國」的作者。

誰應該參加？

不少研究領域，如資訊檢索、問答系統、自動摘要、意見分析等，都有文本理解的需求。我們希望透過RITE這個通用的任務，在文本理解這個關鍵的議題上，匯集相關領域學者的研究能量。我們也非常歡迎其他自然語言處理領域以及機器學習學者一起來參與這場盛會。RITE是一個非對抗性的評鑑活動，非常歡迎任何具野心或探索性質但卻不一定成功的嘗試。您獨一無二的研究對這個研究社群來說絕對是不可或缺的重要貢獻。我們將在官網上(草稿)彙整一系列相關語言資源或工具的資訊，供您建構參賽系統之參考。

任務設計 (草稿)

您可以在任務設計(英文版草稿)網頁找到更詳細的任務細節。

Source: [http://artigas.lti.cs.cmu.edu/rite/Main_Page_\(TC\)](http://artigas.lti.cs.cmu.edu/rite/Main_Page_(TC))

28

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.im.tku.edu.tw/~myday/>