

資料倉儲

Data Warehousing

992DW01

MI4

二 8,9 15:10-17:00 L413

淡江大學資訊管理系

戴敏育

Min-Yuh Day

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

2011-02-15

Min-Yuh Day x

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



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
Address: No.151, Yingzhuan Rd., Danshui Dist., New Taipei City 25137, Taiwan (R.O.C.)
Email: myday@mail.im.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. 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

戴敏育 (Min-Yuh Day)

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



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 Hour](#)]
地址：25137 新北市淡水區英專路151號
Email：myday@mail.im.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):

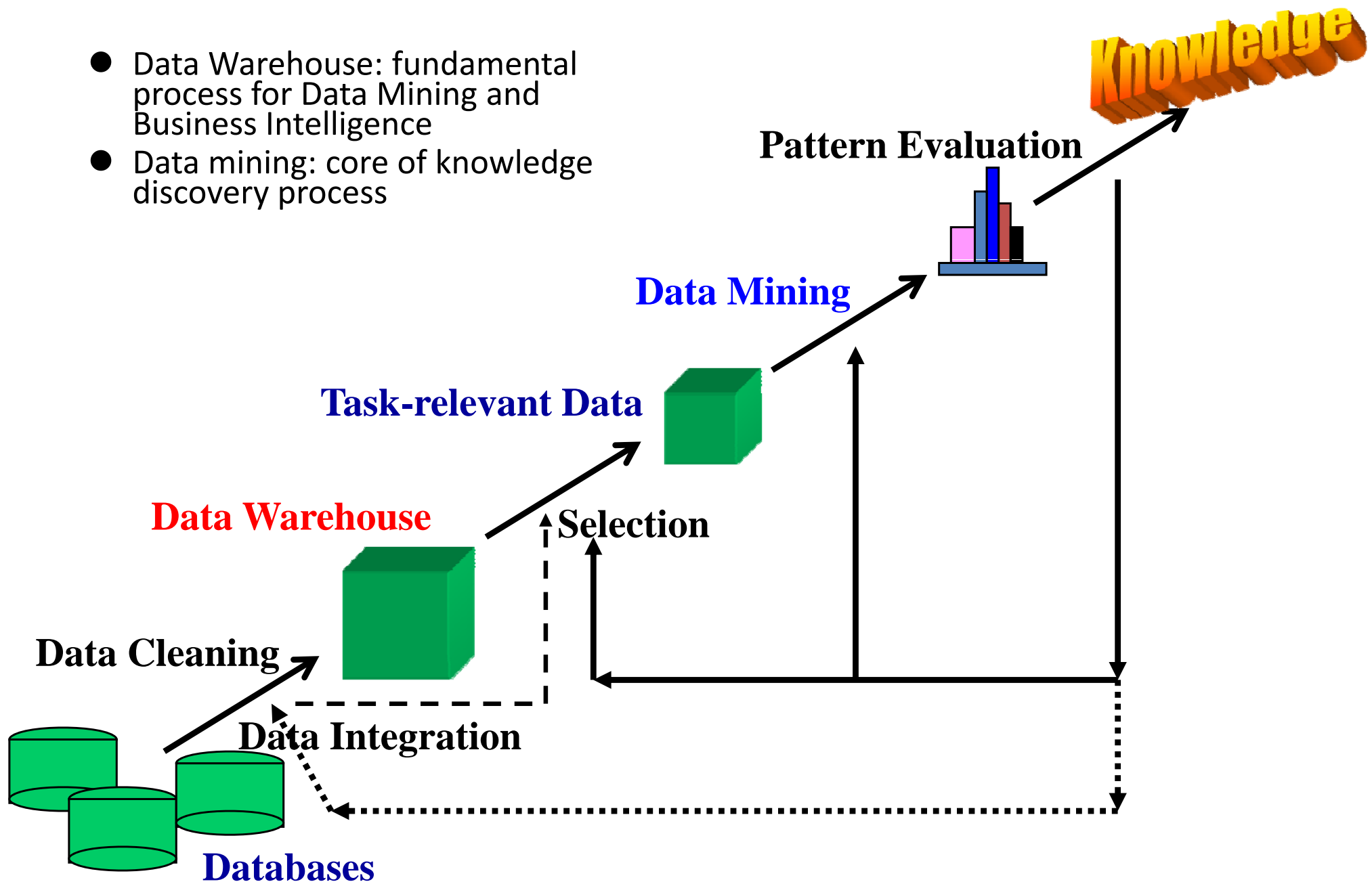
- 博士 國立台灣大學資訊管理研究所 (2001-2010)
博士論文：社交網路服務使用者滿意度評量模式之研究
- 碩士 淡江大學資訊管理研究所 (1993-1995)
碩士論文：應用遺傳演算法發展模糊預測之研究－以銷售預測為例
- 學士 淡江大學資訊管理學系 (1989-1993)

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 8, 9 (Tue 15:10-17:00)
- 上課教室：L413

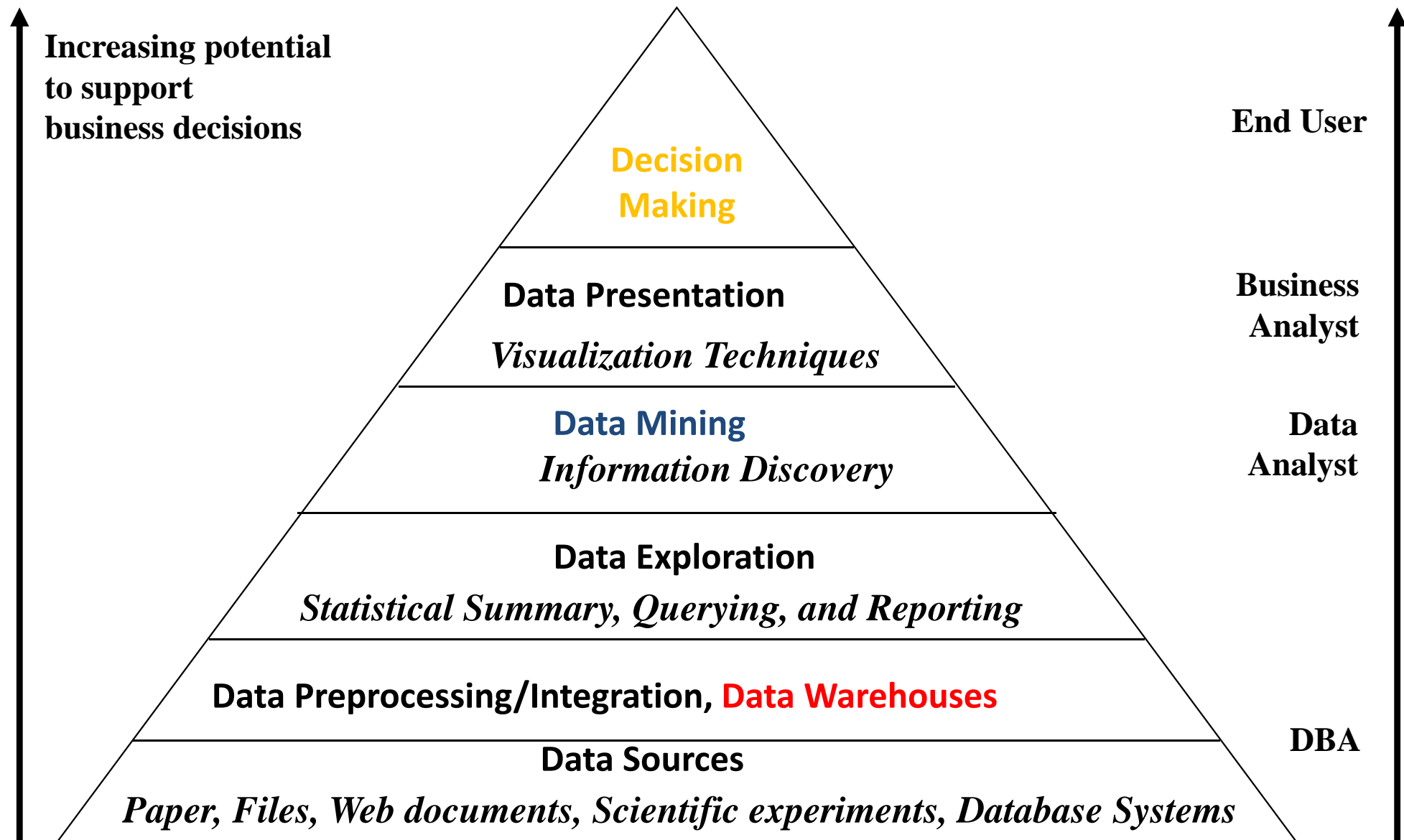
Knowledge Discovery (KDD) Process

- 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。

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

授課進度表

週次	月／日	內容 (Subject/Topics)	備註
1	100/02/15	Introduction to Data Warehousing	
2	100/02/22	Data Warehousing, Data Mining, and Business Intelligence	
3	100/03/01	Data Preprocessing: Integration and the ETL process	
4	100/03/08	Data Warehouse and OLAP Technology	
5	100/03/15	Data Cube Computation and Data Generation	
6	100/03/22	Association Analysis	
7	100/03/29	Classification and Prediction	
8	100/04/05	(放假一天)	100/04/05 (二) 民族掃墓節
9	100/04/12	Cluster Analysis	
10	100/04/19	期中考試週	

授課進度表(續)

週次	月／日	內容 (Subject/Topics)	備註
11	100/04/26	Sequence Data Mining	
12	100/05/03	Social Network Analysis and Link Mining	
13	100/05/10	Text Mining and Web Mining	
14	100/05/17	Project Presentation	
15	100/05/24	畢業班考試	
16	100/05/31	NA	
17	100/06/07	NA	
18	100/06/14	期末考試週	

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
 - 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!**



Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber

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](#)

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

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

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**

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

Han and Kamber: Data ... x

www.cs.uiuc.edu/homes/hanj/bk2/

Text Box:

Jiawei Han and Micheline Kamber

Data Mining: Concepts and Techniques, 2nd ed.

The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor
[Morgan Kaufmann Publishers](#), March 2006. ISBN 1-55860-901-6

“The second edition of Han and Kamber Data Mining: Concepts and Techniques updates and improves the already comprehensive coverage of the first edition and adds coverage of new and important topics, such as mining stream data, mining social networks, and mining spatial, multimedia, and other complex data. This book will be an excellent textbook for courses on Data Mining and Knowledge Discovery.”

-Gregory Piatetsky-Shapiro, President, [KDnuggets](#)

“The second edition is the most complete and up-to-date presentation on this topic. Compared to the already comprehensive and thorough coverage of the first edition, it adds the state-of-the-art research results in new topics such as mining stream, time-series and sequence data as well as mining spatial, multimedia, text and Web data. This book is a must-have for all instructors, researchers, developers and users in the area of data mining and knowledge discovery.”

- Hans-Peter Kriegel, University of Munich, Germany

[Table of Contents in PDF](#)

[Slides \(in PowerPoint form\)](#)

[Art work of the book](#)

Han and Kamber: Data ... x

www.cs.uiuc.edu/homes/hanj/bk2/bk3_slidesindex.htm

Jiawei Han, Micheline Kamber & Jian Pei

Data Mining: Concepts and Techniques, 3rd ed.

Vol. 1.

Morgan Kaufmann Publishers, June 2011.

Slides in PowerPoint form (will be updated without notice!)

- [Chapter 1. Introduction](#)
- [Chapter 2. Know Your Data](#)
- [Chapter 3. Data Preprocessing](#)
- [Chapter 4. Data Warehousing and On-Line Analytical Processing](#)
- [Chapter 5. Data Cube Technology](#)
- [Chapter 6. Mining Frequent Patterns, Associations and Correlations: Basic Concepts and Methods](#)
- [Chapter 7. Advanced Frequent Pattern Mining](#)
- [Chapter 8. Classification: Basic Concepts](#)
- [Chapter 9. Classification: Advanced Methods](#)

作業與學期成績計算方式

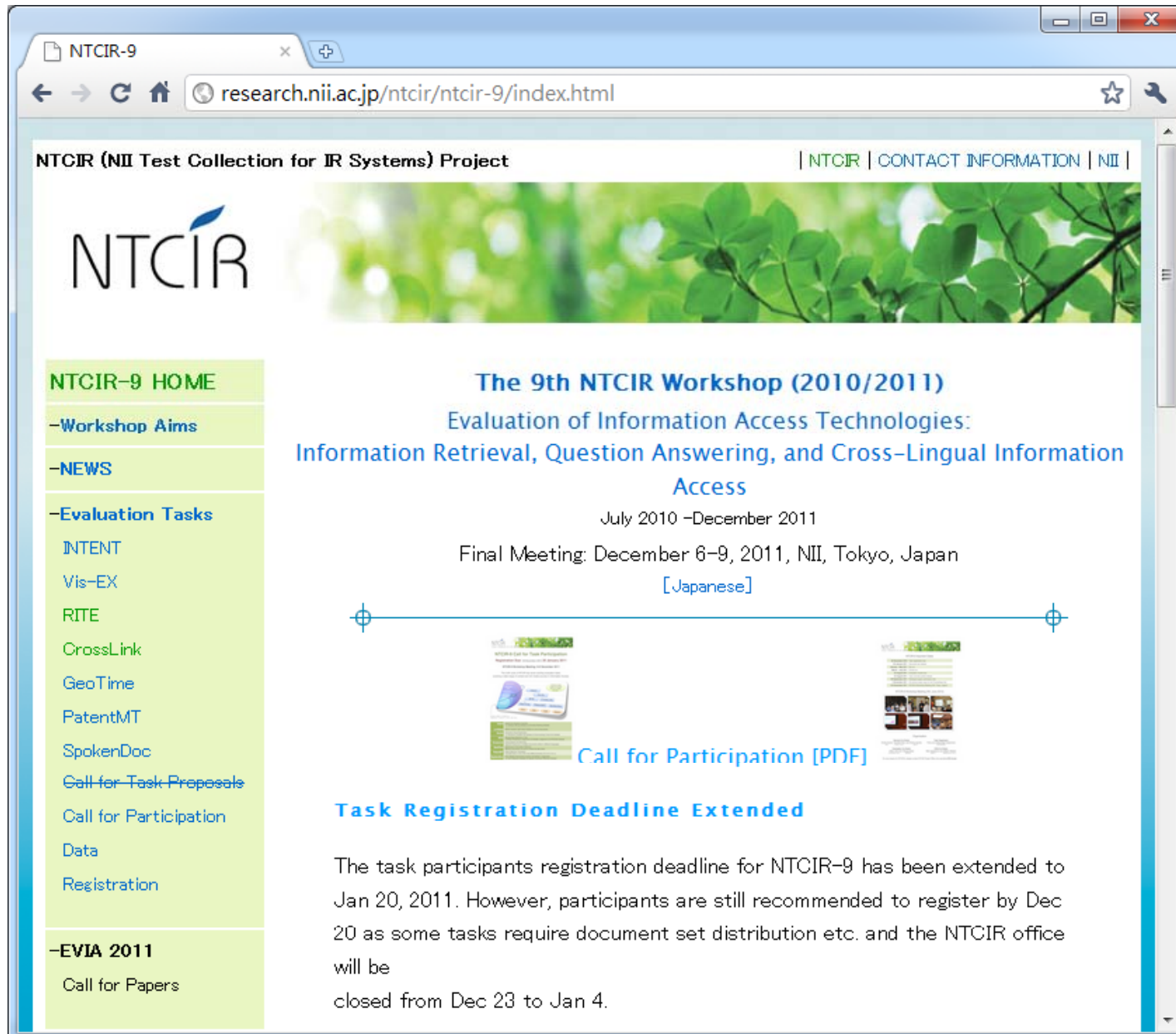
- 批改作業篇數
 - － 1篇（Team Term Project）
- 學期成績計算方式
 - － ☒ 期中考成績：30 %
 - － ☒ 期末考成績：30 %
 - － ☒ 作業成績：20 %（Team Term Project）
 - － ☒ 其他（課堂參與及報告討論表現）：20 %

Term Project

- 參與 NTCIR 國際競賽
 - NTCIR (NII Test Collection for IR Systems) Project
 - NTCIR -9 (July 2010 -December 2011)
 - December 6-9, 2011, NII, Tokyo, Japan
 - NTCIR-9 RITE
 - Recognizing Inference in TExt @NTCIR9
 - http://artigas.lti.cs.cmu.edu/rite/Main_Page
 - NTCIR-9 CrossLink
 - CrossLingual Link Discovery Task
 - <http://ntcir.nii.ac.jp/CrossLink/>
- Open Topic Project
 - Topics related to Data Warehousing, Business Intelligence, Data mining, Text mining, Web mining, Social Network Analysis, Link Mining.

NTCIR Project

(NII Test Collection for IR Systems)

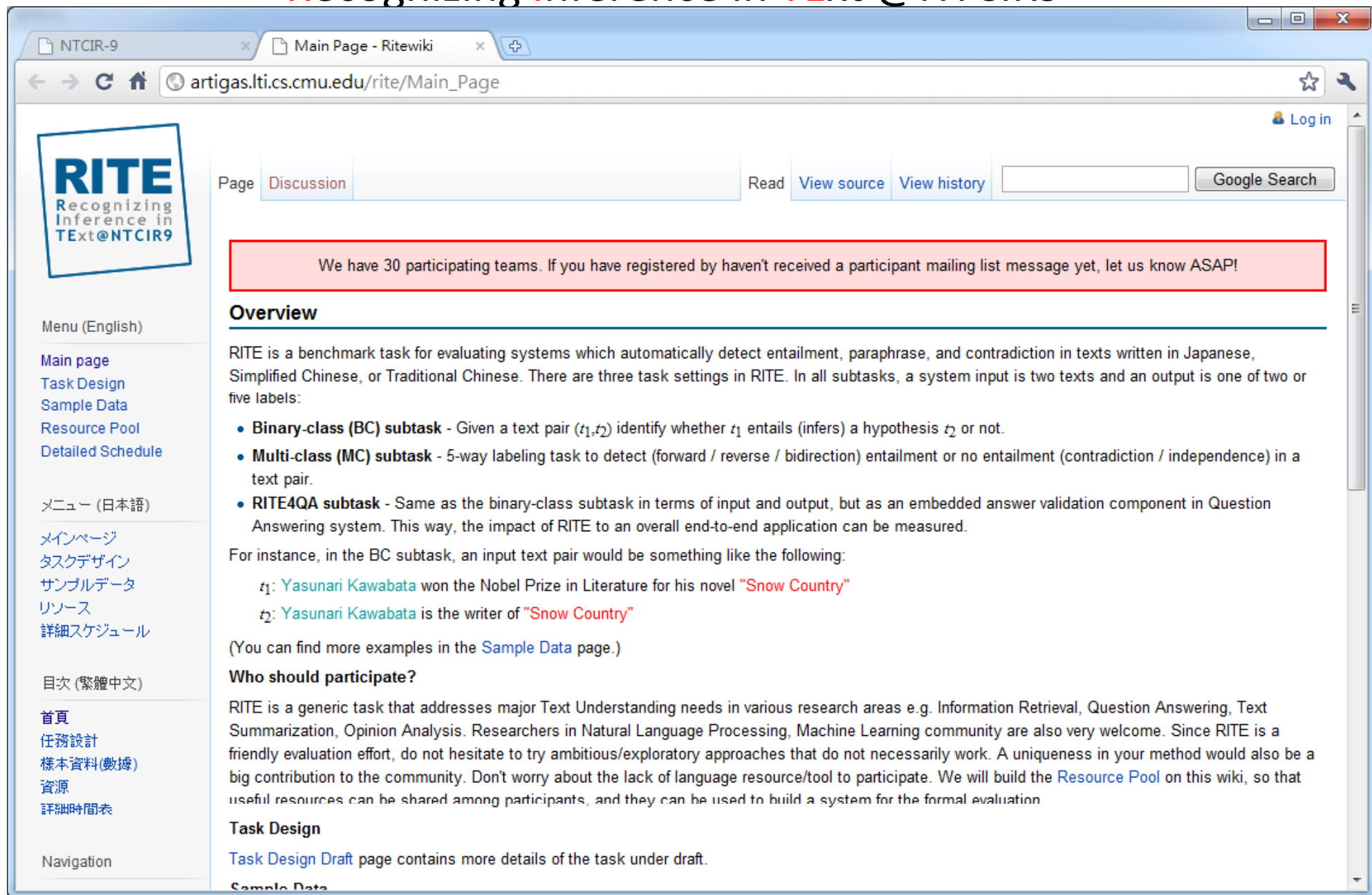


The screenshot shows a web browser window with the address bar displaying research.nii.ac.jp/ntcir/ntcir-9/index.html. The page title is "NTCIR (NII Test Collection for IR Systems) Project". The main content area features a banner with the NTCIR logo and a background image of green leaves. Below the banner, the text reads "The 9th NTCIR Workshop (2010/2011)" followed by "Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access". The dates "July 2010 -December 2011" and the location "Final Meeting: December 6-9, 2011, NII, Tokyo, Japan" are listed, along with a link to "[Japanese]". A horizontal line with a crosshair at each end separates this section from the next. Below the line, there are two small images of workshop materials and a link to "Call for Participation [PDF]". The section "Task Registration Deadline Extended" follows, stating that the registration deadline has been extended to Jan 20, 2011, and that the NTCIR office will be closed from Dec 23 to Jan 4. On the left side of the page, there is a green sidebar with a table of contents.

NTCIR-9 HOME
-Workshop Aims
-NEWS
-Evaluation Tasks
INTENT
Vis-EX
RITE
CrossLink
GeoTime
PatentMT
SpokenDoc
Call for Task Proposals
Call for Participation
Data
Registration
-EVIA 2011
Call for Papers

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 logo for RITE (Recognizing Inference in Text@NTCIR9) on the left. A navigation menu on the left side includes links for Main page, Task Design, Sample Data, Resource Pool, and Detailed Schedule, as well as a Japanese menu and a table of contents. The main content area has tabs for Discussion, Read, View source, and View history. A red-bordered box contains a message about 30 participating teams. Below this is an 'Overview' section describing the RITE benchmark task and its subtasks: Binary-class (BC), Multi-class (MC), and RITE4QA. Examples of text pairs are provided for the BC subtask. The page also includes sections for 'Who should participate?' and 'Task Design'.

Page [Discussion](#) [Read](#) [View source](#) [View history](#) [Google Search](#)

We have 30 participating teams. If you have registered by haven't received a participant mailing list message yet, let us know ASAP!

Overview

RITE is a benchmark task for evaluating systems which automatically detect entailment, paraphrase, and contradiction in texts written in Japanese, Simplified Chinese, or Traditional Chinese. There are three task settings in RITE. In all subtasks, a system input is two texts and an output is one of two or five labels:

- **Binary-class (BC) subtask** - Given a text pair (t_1, t_2) identify whether t_1 entails (infers) a hypothesis t_2 or not.
- **Multi-class (MC) subtask** - 5-way labeling task to detect (forward / reverse / bidirection) entailment or no entailment (contradiction / independence) in a text pair.
- **RITE4QA subtask** - Same as the binary-class 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.

For instance, in the BC subtask, an input text pair would be something like the following:

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"

(You can find more examples in the [Sample Data](#) page.)

Who should participate?

RITE is a generic task that addresses major Text Understanding needs in various research areas e.g. Information Retrieval, Question Answering, Text Summarization, Opinion Analysis. Researchers in Natural Language Processing, Machine Learning community are also very welcome. Since RITE is a friendly evaluation effort, do not hesitate to try ambitious/exploratory approaches that do not necessarily work. A uniqueness in your method would also be a big contribution to the community. Don't worry about the lack of language resource/tool to participate. We will build the [Resource Pool](#) on this wiki, so that useful resources can be shared among participants, and they can be used to build a system for the formal evaluation

Task Design

[Task Design Draft](#) page contains more details of the task under draft.

[Sample Data](#)

NTCIR-9 CrossLink

CrossLingual Link Discovery Task

The screenshot shows a web browser window with the address bar displaying ntcir.nii.ac.jp/CrossLink/. The page features a green-themed header with the NTCIR logo and navigation links: NTCIR HOME, NTCIR-9, NTCIR CMS HOME, NTCIR Organizers, Data, and Important Dates. A search bar is located on the right. The main content area is titled "CrossLingual Link Discovery Task" and includes a "Last Update: 02 Feb 2011" notice. A "ChangLog" section lists updates: 02/02/2011 (Crosslink training topics released), 31/01/2011 (Crosslink validation tool released), and 11/01/2011 (Wikipedia CJK XML collections released). The "1. Introduction" section explains that Cross-lingual link discovery (CLLD) is a way of automatically finding potential links between documents in different languages. It notes that CLLD actively recommends a set of meaningful anchors in the source document and uses them as queries with contextual information to establish links with documents in other languages. A paragraph about Wikipedia's multilingual nature and its extensive hypertext links is also present. A sidebar on the left contains links for NTCIR-9Tasks:CrossLink (Submission, Evaluation, Forum), NTCIR-9Tasks:PatentMT, Data, Proceedings, Papers on NTCIR/using NTCIR, NTCIR Blog, Album, Your Voice, Past Workshop, and Calendar. At the bottom, a calendar for February 2011 is shown.

NTCIR-9Tasks:CrossLink

Submission
Evaluation
Forum

NTCIR-9Tasks:PatentMT

Data
Proceedings
Papers on NTCIR/using NTCIR
NTCIR Blog
Album
Your Voice
Past Workshop
Calendar

2011
Feb
S M T W T F S

CrossLingual Link Discovery Task

Last Update: 02 Feb 2011

ChangLog:
02/02/2011 Crosslink training topics released
31/01/2011 Crosslink validation tool released
11/01/2011 Wikipedia CJK XML collections released

1. Introduction

Cross-lingual link discovery (CLLD) is a way of automatically finding potential links between documents in different languages. It is not directly related to traditional cross-lingual information retrieval (CLIR) because CLIR can be viewed as a process of creating a virtual link between the provided cross-lingual query and the retrieved documents; but CLLD actively recommends a set of meaningful anchors in the source document and uses them as queries with the contextual information from the text to establish links with documents in other languages.

Wikipedia is an online multilingual encyclopaedia that contains a very large numbers of articles covering most written languages and so it includes extensive hypertext links between documents of same language for easy reading and referencing. However, the pages in different languages are rarely linked except for the cross-lingual link between pages about the same subject. This could pose serious difficulties to users who try to seek information or knowledge from different lingual sources. Therefore, cross-lingual link discovery tries to break the language barrier in knowledge sharing. With CLLD users are able to discover documents in languages which they either are familiar with, or which have a richer set of documents than in their language of choice.

Term Project Teams

- 5-7 人為一組
 - 分組名單於 2011.02.22 (二) 課程下課時繳交
 - 由班代統一收集協調分組名單
- NTCIR Project
 - NTCIR-9 RITE (Project 1 Teams)
 - NTCIR-9 CrossLink (Project 2 Teams)
- Open Topic Project (Project 3 Teams)
 - Topics related to Data Warehousing, Business Intelligence, Data mining, Text mining, Web mining, Social Network Analysis, Link Mining.

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) [[Office Hour](#)]

地址：25137 新北市淡水區英專路151號

Email：myday@mail.im.tku.edu.tw

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