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

Assistant Professor
Department of Information Management, Tamkang University

Tel: 886-2-26215656 ext. 2347
Fax: 886-2-26209737
Office: 1716, Chueh-sheng Memorial Hall
Address: No. 151, Yingzhu 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:

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. Chong-Shyang 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

http://mail.im.tku.edu.tw/~myday/
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：1/16 (學生綜合大樓) [Office Hour]
地址：25137 新北市淡水區英華路151號
Email：myday@mail.im.tku.edu.tw
網址：http://mail.im.tku.edu.tw/~myday/

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，曾任職於中興大學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，他在淡江大學資管學系取得碩士學位。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)、問題解答系統(Question Answering Systems)、資料挖掘 (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：淡江大學資訊管理學系 (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)

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

Pattern Evaluation

Data Mining

Task-relevant Data

Selection

Data Warehouse

Data Cleaning

Data Integration

Databases

Source: Han & Kamber (2006)
Data Warehouse

Data Mining and Business Intelligence

- Increasing potential to support business decisions
- Decision Making
- Data Presentation
  - Visualization Techniques
- Data Mining
  - Information Discovery
- Data Exploration
  - Statistical Summary, Querying, and Reporting
- Data Preprocessing/Integration, Data Warehouses
- Data Sources:
  - Paper, Files, Web documents, Scientific experiments, Database Systems

Source: Han & Kamber (2006)
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
<table>
<thead>
<tr>
<th>週次</th>
<th>月／日</th>
<th>內容（Subject/Topics）</th>
<th>備註</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100/02/15</td>
<td>Introduction to Data Warehousing</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>100/02/22</td>
<td>Data Warehousing, Data Mining, and Business Intelligence</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>100/03/01</td>
<td>Data Preprocessing: Integration and the ETL process</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>100/03/08</td>
<td>Data Warehouse and OLAP Technology</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>100/03/15</td>
<td>Data Cube Computation and Data Generation</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>100/03/22</td>
<td>Association Analysis</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>100/03/29</td>
<td>Classification and Prediction</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>100/04/05</td>
<td>(放假一天)</td>
<td>100/04/05 (二) 民族掃墓節</td>
</tr>
<tr>
<td>9</td>
<td>100/04/12</td>
<td>Cluster Analysis</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>100/04/19</td>
<td>期中考試週</td>
<td></td>
</tr>
</tbody>
</table>
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
Jiawei Han and Micheline Kamber

Data Mining: Concepts and Techniques, 2nd ed.

The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor

“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

http://www.cs.uiuc.edu/homes/hanj/bk2/
Jiawei Han, Micheline Kamber & Jian Pei


Vol. 1.


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
  – 期中成績：30 %
  – 期末成績：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 9th NTCIR Workshop (2010/2011)
Evaluation of Information Access Technologies:
Information Retrieval, Question Answering, and Cross-Lingual Information Access
July 2010 - December 2011
Final Meeting: December 6-9, 2011, NII, Tokyo, Japan
(Japanese)

Task Registration Deadline Extended
The task participants registration deadline for NTCIR-9 has been extended to Jan 20, 2011. However, participants are still recommended to register by Dec 20 as some tasks require document set distribution etc. and the NTCIR office will be closed from Dec 23 to Jan 4.

http://research.nii.ac.jp/ntcir/ntcir-9/index.html
NTCIR-9 RITE

Recognizing Inference in TExt @NTCIR9

http://artigas.lti.cs.cmu.edu/rite/Main_Page
CrossLingual Link Discovery Task

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.

http://ntcir.nii.ac.jp/CrossLink/
Term Project Teams

- 5-7 人为一组
  - 分组名单于 2011.02.22 (二) 考课下课时递交
  - 由班代统一收集协调分组名单

- 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/