

CSC8417

ADVANCED WEB DATA MANAGEMENT

Faculty of Sciences

Introductory/Study book

Semester 2 2006

Published by

**University of Southern Queensland
Toowoomba Queensland 4350
Australia**

<http://www.usq.edu.au>

© University of Southern Queensland, 2006.2.

Copyrighted materials reproduced herein are used under the provisions of the Copyright Act 1968 as amended, or as a result of application to the copyright owner.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without prior permission.

Produced by the Distance and e-Learning Centre using the *GOOD* Publishing System.

Table of contents

Introductory material	Page
Course specification	
Course introduction	1
Welcome	1
Course personnel	1
Course overview	2
Course outline	2
How to study this course	3
Support	4
Types of enquiries	4
Types of support	4
USQConnect	6
Study Desk	6
USQAdmin	6
Other links	6
Study schedule	7
Course evaluation	8
Course assessment	9
Assessment scheme	9
Examination	9
Assignment 1	11
Research proposal	11
Submission	12
Assignment 2	13
Dynamic website development	13
Submission	13
Marking criteria	13
Plagiarism	14
Assignment 3	15
Research report	15
Marking criteria	16
Submission	16
Plagiarism	16

Additional resources	17
Study materials	17
Online support	17
Course mailing list	17

Study modules	Page
----------------------	-------------

Preface

1 Web technology and databases

Objectives	1.1
Learning resources	1.1
Key concepts	1.1
Introduction	1.1
1.1 Time allocation	1.2
1.2 Relevance	1.2
1.3 Possible difficulties	1.2
1.4 Learning hints	1.2
1.5 Activities	1.2

2 Semi-structured data and XML

Objectives	2.1
Learning resources	2.1
Key concepts	2.1
Introduction	2.2
2.1 Time allocation	2.2
2.2 Relevance	2.2
2.3 Possible difficulties	2.2
2.4 Learning hints	2.2
2.5 Activities	2.2

3 Query processing

Objectives	3.1
Learning resources	3.1
Key concepts	3.1
Introduction	3.1
3.1 Time allocation	3.2
3.2 Relevance	3.2
3.3 Possible difficulties	3.2
3.4 Learning hints	3.2
3.5 Activities	3.2

4 Transaction management

Objectives	4.1
------------	-----

Learning resources	4.1
Key concepts	4.1
Introduction	4.2
4.1 Time allocation	4.2
4.2 Relevance	4.2
4.3 Possible difficulties	4.2
4.4 Learning hints	4.2
4.5 Activities	4.3
5 Distributed DBMSs	
Objectives	5.1
Learning resources	5.1
Key concepts	5.1
Introduction	5.1
5.1 Time allocation	5.1
5.2 Relevance	5.2
5.3 Possible difficulties	5.2
5.4 Learning hints	5.2
5.5 Activities	5.2
6 Object-oriented and object-relational DBMSs	
Objectives	6.1
Learning resources	6.1
Key concepts	6.1
Introduction	6.1
6.1 Time allocation	6.2
6.2 Relevance	6.2
6.3 Possible difficulties	6.2
6.4 Learning hints	6.2
6.5 Activities	6.2
7 Data warehousing	
Objectives	7.1
Learning resources	7.1
Key concepts	7.1
Introduction	7.1
7.1 Time allocation	7.2
7.2 Relevance	7.2
7.3 Possible difficulties	7.2
7.4 Learning hints	7.2
7.5 Activities	7.2
8 OLAP and data mining	
Objectives	8.1
Learning resources	8.1
Key concepts	8.1
Introduction	8.1

8.1	Time allocation	8.2
8.2	Relevance	8.2
8.3	Possible difficulties	8.2
8.4	Learning hints	8.2
8.5	Activities	8.2

9 Information retrieval and web search

	Objectives	9.1
	Learning resources	9.1
	Key concepts	9.1
	Introduction	9.1
9.1	Time allocation	9.1
9.2	Relevance	9.2
9.3	Possible difficulties	9.2
9.4	Learning hints	9.2
9.5	Activities	9.2

Introductory material



The University of Southern Queensland

Course specification

Description: Advanced Web Data Management

Subject	Cat-nbr	Class	Term	Mode	Units	Campus
CSC	8417	54323	2, 2006	EXT	1.00	TWMBA

Academic group:	FOSCI
Academic org:	FOS003
Student contribution band:	2
ASCED code:	020199

STAFFING

Examiner: Stijn Dekeyser
Moderator: Hua Wang

REQUISITES

Pre-requisite: CSC3400 and Students must be enrolled in one of the following Programs: BINH or GCAC or GCPC or GDAC or GDPC or MCOP or MPIT or MPCP

RATIONALE

While database systems were already very important to large organizations before the advent of the World Wide Web, the IT revolution of the last decade has made them indispensable. Advanced database features such as transaction support, query optimization, data distribution, semi-structured data management, and information retrieval have all received extended attention to better support data-intensive web-based applications.

SYNOPSIS

This research-oriented course introduces students to advanced web data management concepts such as transaction support, query optimization, data distribution, semi-structured data management, on-line analytical processing, data mining, information retrieval, and web search engine architectures.

OBJECTIVES

On completion of this course, students will be able to:

1. demonstrate a general understanding of advanced database management issues;
2. demonstrate a good understanding of recent technologies linking databases to web sites;
3. develop database-driven web sites.

TOPICS

	Description	Weighting (%)
1.	Web-database integration techniques	20.00
2.	Transaction Management : transactions, concurrency control and recovery	10.00
3.	Query processing: optimization of Relational Algebra and SQL expressions	10.00
4.	Distributed Databases: concepts, functions, design, transactions, and query optimization in a DDBMS	10.00
5.	Semi-structured data and XML	10.00
6.	On-line analytical processing (OLAP)	10.00
7.	Data mining	10.00
8.	Information Retrieval: models and techniques	10.00
9.	Web Search: search engine techniques	10.00

TEXT and MATERIALS required to be PURCHASED or ACCESSED:

ALL textbooks and materials are available for purchase from USQ BOOKSHOP (unless otherwise stated). Orders may be placed via secure internet, free fax 1800642453, phone 07 46312742 (within Australia), or mail. Overseas students should fax +61 7 46311743, or phone +61 7 46312742. For costs, further details, and internet ordering, use the 'Textbook Search' facility at <http://bookshop.usq.edu.au> click 'Semester', then enter your 'Course Code' (no spaces).

Connolly, T & Begg, C 2005, *Database systems, a practical approach to design, implementation and management*, 4th edn, Addison-Wesley, Boston.

REFERENCE MATERIALS:

Reference materials are materials that, if accessed by students, may improve their knowledge and understanding of the material in the course and enrich their learning experience.

On-line lecture notes, tutorial material and resources at
<http://www.sci.usq.edu.au/courses/csc8417>

Baeza-Yates, R & Ribiro-Neto, B 1999, *Modern information retrieval*, Addison-Wesley,
Silberschatz, A, Korth, HF & Sudarshan, S 2005, *Database system concepts*, 5th edn, McGraw-Hill,

STUDENT WORKLOAD REQUIREMENTS:

ACTIVITY	HOURS
Assessment	50.00
Private Study	120.00

ASSESSMENT DETAILS

Description	Marks out of	Wtg(%)	Due date
ASSIGNMENT 1	100.00	10.00	25 Aug 2006
ASSIGNMENT 2	100.00	45.00	09 Oct 2006
ASSIGNMENT 3	100.00	45.00	30 Oct 2006

IMPORTANT ASSESSMENT INFORMATION

- 1 Attendance requirements:
There are no attendance requirements for this course. However, it is the students' responsibility to study all material provided to them or required to be accessed by them to maximise their chance of meeting the objectives of the course and to be informed of course-related activities and administration.
- 2 Requirements for students to complete each assessment item satisfactorily:
To complete each of the assessment items satisfactorily, students must obtain at least 50% of the marks available for each assessment item.
- 3 Penalties for late submission of required work:
If students submit assignments after the due date without prior approval then a penalty of 10% of the total marks gained by the student for the assignment will apply for each working day late.
- 4 Requirements for student to be awarded a passing grade in the course:
To be assured of receiving a passing grade a student must submit all of the summative assessment items and achieve at least 50% of the available marks for those items.
- 5 Method used to combine assessment results to attain final grade:
The final grades for students will be assigned on the basis of the aggregate of the weighted marks obtained for each of the summative assessment items in the course.
- 6 Examination information:
There is no examination in this course.
- 7 Examination period when Deferred/Supplementary examinations will be held:
There will be no Deferred or Supplementary examinations in this course.
- 8 University Regulations:
Students should read USQ Regulations 5.1 Definitions, 5.6. Assessment, and 5.10 Academic Misconduct for further information and to avoid actions which might contravene University Regulations. These regulations can be found at the URL

<http://www.usq.edu.au/corporateservices/calendar/part5.htm> or in the current USQ Handbook.

ASSESSMENT NOTES

- 9 Students must retain a copy of each item submitted for assessment. If requested, students will be required to provide a copy of assignments submitted for assessment purposes. Such copies should be despatched to USQ within 24 hours of receipt of a request being made.
- 10 The due date for an assignment is the date by which a student must despatch the assignment to the USQ. The onus is on the student to provide proof of the despatch date, if requested by the Examiner.
- 11 Students must retain a copy of each item submitted for assessment. This must be produced within 24 hours if required by the Examiner.
- 12 The examiner of a course may grant an extension of the due date of an assignment in extenuating circumstances.
- 13 The Faculty will NOT accept submission of assignments by facsimile.
- 14 Students who have undertaken all of the required assessments in a course but who have failed to meet some of the specified objectives of a course within the normally prescribed time may be awarded the temporary grade: IM (Incomplete - Make up). An IM grade will only be awarded when, in the opinion of the examiner, a student will be able to achieve the remaining objectives of the course after a period of non directed personal study.
- 15 Students who, for medical, family/personal, or employment-related reasons, are unable to complete an assignment, may apply to defer the assessment. Such a request must be accompanied by appropriate supporting documentation. The temporary grade IDM (Incomplete Deferred Make-up) will be awarded to such students until the deferred assignment has been assessed.
- 16 Students will require access to e-mail and internet access to USQConnect for this course.

Course introduction

Welcome

Welcome to this course on *Advanced web data management*. Due to significant advances in Internet and Web technology over the last decade and a half, the Web has become an important platform for business applications. More and more information is put on the Web and ever more business transactions are conducted on the Web. This course will prepare you for future research and a career in Web based data management and Web Information Systems.

While database systems were already very important to large organizations before the advent of the World Wide Web, the IT revolution referred to above has made them indispensable. Advanced database features such as transaction support, query optimization, data distribution, semi-structured data management, and information retrieval have all received extended attention to better support data-intensive web-based applications.

Hence, this course first extends your knowledge of relational database systems, before delving into other web data management issues and paradigms.

Course personnel

Course team leader – Stijn Dekeyser



Since December 2003 I have been a lecturer at USQ, having come over from the University of Antwerp, Belgium, where I received my PhD (on relational databases and XML) in June of the same year. In Antwerp, I was an assistant at the Advanced Database Research and Modelling research group.

At USQ I primarily teach courses related to Web Information Systems: CSC3400 Database Systems, CSC8409 XML and Semantic Web Services, and CSC8417 Advanced Web Data Management. I have also taught Algorithms & Data Structures and Graphical User Interface Programming.

My first name is pronounced as Stan (although with a slightly softer nasal tone).

Course moderator – Hua Wang



Hua also received his PhD in Computer Science in 2003, having been active in industry before joining USQ. He primarily teaches postgraduate courses, while his research interests lie with Security and Web Information Systems.

Course overview

This course is intended for students who have already taken an introductory database system course such as **CSC3400 Database Systems**, so students who have not had this experience, while still being able to take the course, may need to study some of the material of CSC3400.

The course has two components. The research-oriented component introduces students to advanced web data management concepts such as transaction support, query optimization, data distribution, semi-structured data management, on-line analytical processing, data mining, information retrieval, and web search engine architectures. The practical component involves creating a database-driven dynamic web site.

The broad goals of this course are to:

1. guide students in developing a broad understanding of a number of advanced data management issues in databases
2. guide students in developing an in-depth understanding of one specific topic of your choice related to web data management, and write a quality research report on the topic
3. develop database-driven dynamic web sites.

Course outline

This course has 9 study modules which aim to fulfil the first objective of the course. Each module provides learning objectives, followed by sections for further discussion or presentation on each topic/item. At the end of each module there is a list of reading activities.

The assessment of the course consists only of the three assignments; there is no examination. The assignments can be completed without studying the material of the modules. Completion of the assignments fulfils the second and third objective of the course.

As the material in the study modules is not assessed but contributes to the goals of the course, we rely on the maturity and responsibility of postgraduate students to take the opportunity to learn in the absence of assessment pressure. Note also that your selected research topic for Assignments

1 and 3 may be closely related to one of the study modules, and that module 1 may be very useful in the context of Assignment 2.

Since this is a Honours/Masters level course, covering advanced technologies/material in the selected areas, there is no single textbook with the required coverage. However, the chosen textbook by Connolly and Begg covers 8 of the 9 study modules. Other resources will be the online readings listed at the end of each module in addition to the lecture slides made available via the course web site. External students need to have Web access in order to read the online papers/articles and to complete the assignments/projects.

How to study this course

The purpose of the study modules is to outline the concepts/technologies to be covered. In order to fully understand this material, you will need to read the corresponding readings listed at the end of each module.

The study modules have the following goals:

- to summarise concepts or techniques
- to clarify certain points and concepts
- to point you to the right references for particular technologies/concepts.

Approach the material as follows:

Step 1 - Read the appropriate sections of the study modules, textbook chapters, updated lecture slides (to be provided on line) and online references.

Step 2 - Perform the exercises/assignments. Do not wait until the assignment due dates. Two projects need to be planned/started from the very beginning and they will take a few months to finish.

The course web site at <http://www.sci.usq.edu.au/courses/CSC8417/> will be your most important source of information. You must access it regularly.

Support

Types of enquiries

You have access to a wide variety of support services at USQ. Follow the details below or visit the 'Current Students' website at <http://www.usq.edu.au/currentstudents/default.htm> for more information.

General enquiries

USQ*Assist* is the most efficient method for requesting assistance for:

- administrative queries
- assignment submissions
- study assistance
- contacting your lecturer.

Technical enquiries

Enquiries relating to access to USQ*Connect*, the USQ*StudyDesk*, or other technical issues can also be directed to USQ*Assist*. If you cannot access USQ*Assist*, contact the Student IT HelpDesk on +61 7 4631 1510 or email usqconnect@usq.edu.au for assistance. The Student IT HelpDesk is staffed weekdays between 8.00am and 5.00pm (AEST-Australian Eastern Standard Time), with voicemail after hours.

Types of support

There are a number of ways of accessing support services.

Online support

USQ*Assist* is a web self-serve facility for you to:

- find answers to common questions at any time
- ask any question
- track the progress of your question
- keep a record of questions and responses.

To access USQ*Assist* go to <http://usqassist.usq.edu.au> or click on 'USQ*Assist*' in USQ*Connect*.

Telephone support

If you prefer to telephone, call Outreach Services on 07 46312285 for assistance. Outreach Services is staffed weekdays between 8.30am and 5.30pm (AEST), with voicemail after hours. If you are located in Eastern Australia, contact your Regional Liaison Officer.

International students telephone the USQ International Office on +61 7 46312362, or your Agent. USQ International is staffed weekdays between 9.00am and 5.00pm (AEST), with voicemail after hours.

Fax

International students fax the USQ International Office on +61 7 46362211. All other students fax the Distance and e-Learning Centre on 07 46361049.

Postal address

The Administrator
Distance and e-Learning Centre
University of Southern Queensland
Toowoomba Qld 4350
Australia

USQConnect

USQConnect provides you with online access to information, services and course resources relevant to your studies from a convenient, central point. To access USQConnect, from the USQ home page at <http://www.usq.edu.au> click on USQConnect, or go directly to the URL at <http://usqconnect.usq.edu.au>. You will require your USQConnect username and password to access the system. You will be notified of this username and password on your first Letter of Offer or Enrolment Notice.

Study Desk

Your Study Desk in USQConnect gives access to a home page for every course in which you are currently enrolled. Content available from the course home page will vary according to the teaching requirements of the course, but may include:

- course materials and resources
- electronic course discussion facilities
- access to past examination papers.

As each course has specific learning requirements, availability of these features will vary between courses.

USQAdmin

USQAdmin, also accessed through USQConnect, allows you to access a number of administrative functions such as changing your contact details, checking your enrolment details, accessing learning circles, checking final grades, viewing your exam timetable, changing your exam centre, and more.

Other links

USQConnect also gives access to the Library and the Academic Learning Support site, as well as the Quick Links list of University sections and services.

Study schedule

Week	Module	Activity/Reading	Assessment
1 24–28 July	M1 Web technology and databases	Connolly & Begg, chapter 29 (selected sections)	
2 31 July – 4 August	M1 Web technology and databases	Connolly & Begg, chapter 29 (selected sections)	Start work on Assignments 1 and 2
3 7–11 August	M2 Semi-structured data and XML	Connolly & Begg, chapter 30 (selected sections)	
4 14–18 August	M3 Query processing	Connolly & Begg, chapter 21 (selected sections)	
5 21–25 August	M3 Query processing	Connolly & Begg, chapter 21 (selected sections)	Assignment 1 Due: 15 August 2006 Start work on Assignment 3
6 28 Aug – 1 Sept	M4 Transaction management	Connolly & Begg, chapter 20 (selected sections)	
7 4–8 September	M4 Transaction management	Connolly & Begg, chapter 20 (selected sections)	
8 11–15 September	M5 Distributed DBMSs	Connolly & Begg, chapter 22 (selected sections)	
9 18–22 September	M6 Object-oriented and object-relational DBMSs	Connolly & Begg, chapters 25 & 28 (selected sections)	
10–11 25 Sept – 6 Oct	RECESS		
12 9–13 October	M7 Data warehousing	Connolly & Begg, chapters 31 & 32 (selected sections)	Assignment 2 Due: 9 October 2006
13 16–20 October	M8 OLAP and data mining	Connolly & Begg, chapters 33 & 34 (selected sections)	
14 23–27 October	M9 Information retrieval and web search	Resources listed on course website	
15 30 Oct – 3 Nov	M9 Information retrieval and web search	Resources listed on course website	Assignment 3 Due: 30 October 2006
16–18 6–24 November	EXAMINATION PERIOD There is no exam for this course.		

Course evaluation (external students only)

The University of Southern Queensland is committed to continuous improvement, and **seeks your input** to that process through your participation in our course evaluation process. Please complete and return the questionnaire ‘Student Feedback’ included later in this introductory material.

Your response will be processed so that, unless you wish otherwise, the course examiner will not be aware of your identity. Please help us to help our students by providing feedback on your experiences in this course.

When to return the questionnaire

Please return the questionnaire before the end of this semester’s examination period.

Where to send the questionnaire

1. Insert the completed questionnaire in an envelope, seal and address envelope as follows:

The Course Evaluation Co-ordinator
Information Technology Services
University of Southern Queensland
Toowoomba 4350
Australia

2. The envelope may be posted directly to the above address

OR

attached to the outside of your last assignment for this course and then posted to DeC.

Course assessment

Assessment scheme

The course will be assessed as follows:

Description	Marks out of	Wtg (%)	Due date
ASSIGNMENT 1 – Research proposal	100.00	10.00	25 August 2006
ASSIGNMENT 2 – Developing a dynamic database-driven website	100.00	45.00	9 October 2006
ASSIGNMENT 3 – Research project	100.00	45.00	30 October 2006

All assignments are a compulsory part of the assessment.

Feedback will be provided as soon as possible so that you can learn from your assignment, providing submission has been made on or prior to the due date. Late submissions may carry a severe penalty. If you have special reasons, an extension might be given. You need to provide documentary evidence to support your case.

Examination

There is **no** examination for this course; the three assignments constitute the total assessment for this course.

Assignment 1

Description	Marks out of	Wtg (%)	Due date
Research proposal	100.00	10.00	25 August 2006

Research proposal

Assignments one and three of this course are closely connected. The third assignment involves writing a research report (more specifically, a literature review) on a topic related to web data management. To help you start, this first assignment is meant to develop a proposal before you write the research report itself.

You need to select one topic from the list given in Assignment 3 (an updated list will be available on the course web site), or propose your own topic which needs to be accepted by the lecturer.

The Research proposal is a document of at most three pages in which you indicate:

- your choice of topic
- a description of the topic and how it relates to data management and/or web technologies
- at least 3 sub-topics or questions that you may address in your actual Research report (see Assignment 3)
- your appraisal of how important this topic is for
 - (a) fundamental (theoretical) research
 - (b) general applied computer science
 - (c) cutting edge IT industry
 - (d) general computing industry
 - (e) computer users in general
- a list of initial ‘secondary’ sources: books, encyclopedias, monographs, and (perhaps web-based) review articles, plus a very brief (one or two sentences for each source) description of why you think the source may be useful in your Research report
- a list of at least 3 ‘primary’ sources (published research papers presented at conferences or articles in scientific journals) plus a very brief (one or two sentences per source) description of why you think the source may be useful in your Research report.

It is strongly preferred that you write your document using **LaTeX**. Instructions on how to do this will be made available via the course website. Regardless of your choice of software to produce your research proposal, you must be able to generate a PDF document from it.

Submission

Submission is as a single PDF file by email to the lecturer by the due date. A word processor file (e.g. MS Word .doc) is **not** acceptable.

Assignment 2

Description	Marks out of	Wtg (%)	Due date
Dynamic website development	100.00	45.00	9 October 2006

Dynamic website development

The second assignment is very practical: the aim is to develop a dynamic, database-driven website. Your project will be implemented using XAMPP (see further below), meaning that MySQL will be the database server, and PHP will be the server-side scripting language to be used.

You may develop the website as a team effort, with a partner of your choice (or allocated to you by the lecturer). A team can consist of no more than 2 students. It is important that you adequately and equitably divide the work among the partners, and that you signal any collaboration problems to the lecturer in a timely fashion (i.e. as soon as one team member has a documentable complaint regarding the conduct of the other team member). Remember that it is your responsibility to complete this project well.

In normal circumstances (i.e. when the examiner was not forced to intervene) the marks obtained for the second assignment will be the same for both members of the team.

The requirements for the dynamic, database-driven website will be published on the course website around the start of the semester.

It is strongly recommended that you work on Assignment 2 and 3 concurrently.

You will need to obtain, install, and use the **XAMPP** software package as the platform in which to develop the project. This software is free, can be downloaded from the Internet (<<http://www.apachefriends.org/en/xampp.html>>) and is also available on the Department's DVD-Rom which can be bought from the USQ Bookshop. XAMPP is available for Windows and Linux operating systems.

Submission

Submission instructions will be available on the course web site.

<<http://www.sci.usq.edu.au/courses/CSC8417/>>

Marking criteria

This assignment will assess your practical skills in developing and integrating databases and website. Your product should show that your team can build a sound database, query and update it effectively through a web interface, design and implement a dynamic website that conforms to W3C standards, and overall deliver a usable software product.

Criteria	Marks / 100	Grade
Functional dynamic website without major bugs, database schema is adequate but limited	50–64	C (pass)
Installation and running of product is flawless, database schema is very well designed	65–74	B
The website has no security problems, conforms to modern W3C standards, runs flawlessly and is close to being of industrial quality	75–84	A
The website has all of the requirements of 'A', has additional functionality not specified in the criteria, and is of industrial quality	85–100	HD (High Distinction)

Plagiarism

Plagiarism is a serious issue that carries with it severe penalties. Your *Distance education student guide* provides a detailed section on the definition and implications of plagiarism.

Assignment 3

Description	Marks out of	Wtg (%)	Due date
Research report	100.00	45.00	30 October 2006

Research report

The topic for your research report (or more specifically, literature review) must be the same as the one you selected for Assignment 1, unless the lecturer specifically asked you or agreed to your request to take a different one.

This assignment must be completed by you alone; it is **not** a team project.

You are to conduct extensive readings in the area of choice, then develop a sub-topic or question that you will address in detail in your report. The sub-topic does not necessarily have to be one from your research proposal (Assignment 1), although that would be preferable.

The report must have a title, an abstract, an introduction and at least one section that extensively reviews the literature you have read. Include other sections and subsections to carefully structure your report. It is advisable that you try to include recent developments, current problems, possible solutions, examples of implementations, and future directions, if applicable. The report must be ended with a list of references.

It is **not** the objective of this assignment to create new knowledge, nor to propose novel solutions to a problem. You should also **not** implement any software, as this is a research-only assignment.

In scientific writing, it is important to compare and contrast whenever possible, and remain objective at all times. Present what you have learned in your own words, and if applicable show that you can critically reason about the topic using your own ideas and arguments.

Do **NOT** copy any text directly from any document, digital or on paper, unless clearly quoting it together with reference to the source. Plagiarism will be subject to severe penalties.

You should use **LaTeX** to write your report (instructions and a simple template will be available on the course website), which should be between 12 and 15 pages long (including the list of references). You must include at least 10 references (both ‘secondary’ and ‘primary’), formatted as in a LaTeX document.

Here is a list of research projects. An updated list is to be posted in the course home page.

1. Transaction support for Web Database Transactions
2. Integrating Web and DBMS
3. On Web Database Security
4. On Semi-structured Data Model and Query Processing
5. From XML Schema to RDF

6. Integrating Web Databases via RDF
7. On XML Query Optimization (X-Path)
8. Query Optimization with XML Query Algebra (or XQuery)
9. Web-based Information Retrieval (such as Models, Indexing or Searching)
10. Web Service Management
11. Web Service Modelling
12. Web Service Specification
13. Service Searching Engine
14. Web Service Composition
15. Web Service Scheduling & Execution
16. Service Reliability, Trustiness and Security

Marking criteria

This assignment will assess your research skills. You should show you have gained a deep understanding of a particular topic through extensive reading, comparison and synthesis of an important topic relevant to data management and/or web technologies. You should present your own view on the development of the topic, and organize your presentation in a logical way.

Criteria	Marks/100	Grade
Extensive readings & literature reviews, synthesis, and comparisons.	50-64	C (pass)
Having a logical and clear presentation, in addition to the requirements for 'C'.	65-74	B
A deep understanding shown in the report in addition to requirements for 'B'.	75-84	A
All the requirements of 'A' with additional originality & innovation.	85-100	HD (High Distinction)

Submission

Submission is by email to the examiner. You must mail a single PDF file as a normal attachment to an email, by the due date. **No** word processing files (e.g. MS Word .doc) are allowed.

Plagiarism

Plagiarism is a serious issue that carries with it severe penalties. Your *Distance education student guide* provides a detailed section on the definition and implications of plagiarism.

Additional resources

Study materials

- This Introductory/Study book
- Online reading materials
- Updated lecture slides online
- Textbook – Connolly, T & Begg, C 2005, *Database systems, a practical approach to design, implementation and management*, 4th edn, Addison Wesley, New York

References

Baeza-Yates, R & Ribiro-Neto, B 1999, *Modern information retrieval*, Addison Wesley.

Silberschatz, A, Korth, HF & Sudarshan, S 2002, *Database system concepts*, 4th edn, McGraw Hill, New York.

Online support

You will find the course web site at:

<http://www.sci.usq.edu.au/courses/CSC8417/>,

and an electronic mailing list accessible through the site.

Send electronic mail to the lecturer/examiner at: dekeyser@usq.edu.au

Course mailing list

A course mailing list will be created at the start of the semester. The email address is csc8417.s2@www.sci.usq.edu.au. The list works as follows. An email sent by anyone (student or examiner) to the list will be automatically forwarded by the list server to all registered members of the list. The idea is that if you have a problem or question or comment, you can post it to the list as an email message and everyone else in the course will receive the message. Any student who wishes may respond to the message.

The examiner's response to a message sent to dekeyser@usq.edu.au may also be copied to the mailing list if it is of general relevance.

The course examiner will read all messages on the list and respond to them if necessary. The list will also be used to broadcast messages of general interest. It is also encouraged that students discuss among themselves via this list.

To register your email address for the course email list, please visit the relevant page on the course web site.

Your assignment was received at the University on the date noted here and will be returned to you when marked.

DISTANCE & E-LEARNING CENTRE
UNIVERSITY OF SOUTHERN QUEENSLAND
Toowoomba, Queensland 4350 Australia

OFFICE USE ONLY

assignment acknowledgement

COURSE NAME				COURSE NUMBER	
ASSIGNMENT NUMBER		DATE OF POSTING		STUDENT ID NUMBER	

Please complete details above AND print YOUR NAME and ADDRESS on the reverse side and **AFFIX THE CORRECT POSTAGE** otherwise this card will not be returned to you.

P.T.O.



Your assignment was received at the University on the date noted here and will be returned to you when marked.

DISTANCE & E-LEARNING CENTRE
UNIVERSITY OF SOUTHERN QUEENSLAND
Toowoomba, Queensland 4350 Australia

OFFICE USE ONLY

assignment acknowledgement

COURSE NAME				COURSE NUMBER	
ASSIGNMENT NUMBER		DATE OF POSTING		STUDENT ID NUMBER	

Please complete details above AND print YOUR NAME and ADDRESS on the reverse side and **AFFIX THE CORRECT POSTAGE** otherwise this card will not be returned to you.

P.T.O.



Your assignment was received at the University on the date noted here and will be returned to you when marked.

DISTANCE & E-LEARNING CENTRE
UNIVERSITY OF SOUTHERN QUEENSLAND
Toowoomba, Queensland 4350 Australia

OFFICE USE ONLY

assignment acknowledgement

COURSE NAME				COURSE NUMBER	
ASSIGNMENT NUMBER		DATE OF POSTING		STUDENT ID NUMBER	

Please complete details above AND print YOUR NAME and ADDRESS on the reverse side and **AFFIX THE CORRECT POSTAGE** otherwise this card will not be returned to you.

P.T.O.

PLACE
STAMP
HERE

NAME _____

ADDRESS _____

_____ POSTCODE _____

PLACE
STAMP
HERE

NAME _____

ADDRESS _____

_____ POSTCODE _____

PLACE
STAMP
HERE

NAME _____

ADDRESS _____

_____ POSTCODE _____

Study modules

Preface

Welcome to this course on *Advanced web data management*. Due to significant advances in Internet and Web technology over the last decade and a half, the Web has become an important platform for business applications. More and more information is put on the Web and ever more business transactions are conducted on the Web. This course will prepare you for future research and a career in Web based data management and Web Information Systems.

While database systems were already very important to large organizations before the advent of the World Wide Web, the IT revolution referred to above has made them indispensable. Advanced database features such as transaction support, query optimization, data distribution, semi-structured data management, and information retrieval have all received extended attention to better support data-intensive web-based applications.

Hence, this course first extends your knowledge of relational database systems, before delving into other web data management issues and paradigms.

The course is intended for students who have already taken an introductory database system course such as *CSC3400 Database systems*, so students who have not had this experience, while still being able to take the course, may need to study some of the material of CSC3400 on their own.

The course has two components. The research-oriented component introduces students to advanced web data management concepts such as transaction support, query optimization, data distribution, semi-structured data management, on-line analytical processing, data mining, information retrieval, and web search engine architectures. The practical component involves creating a database-driven dynamic web site.

The broad goals of this course are to:

1. guide students in developing a broad understanding of a number of advanced data management issues in databases
2. guide students in developing an in-depth understanding of one specific topic of your choice related to web data management, and write a quality research report on the topic
3. develop database-driven dynamic web sites.

Note on text book editions

There are 9 modules in this course, of which 8 are largely covered by relevant chapters in the prescribed text book. Note that we use the fourth edition of the book; some students may have the older third edition. While the structure of parts and chapters has changed from the third to the fourth edition, the content of the relevant chapters is largely unchanged, meaning that you may safely use the older edition if you already own it. You will need to be aware that chapter and section numbers mentioned in this study material may differ, however.

Module 1

Web technology and databases

Objectives

In this module, you will learn

- the basics of Web Technology: HTTP, HTML, URLs, etc.
- the advantages and disadvantages of the Web as a database interface
- approaches for web–database integration, especially server-side scripting languages.

Learning resources

Textbook material that must be read:

- 29.1 Introduction to the Internet & Web
- 29.2 The Web
- 29.3 Scripting languages

The other sections in Chapter 29 are optional.

You should also study the PHP simple tutorial (see Reading activity).

Key concepts

- the basics of the Internet, Web, HTTP, HTML, and URLs
- the difference between the two-tier and the three-tier client-server architecture
- the advantages and disadvantages of the Web as a database platform
- approaches for integrating databases into the Web environment:
 - client-side scripting languages (**JavaScript** and VBScript)
 - server-side scripting languages (**PHP** and ASP)
 - common Gateway Interface (CGI)
 - HTTP cookies.

Introduction

The aim of this module is to examine the integration of the DBMS into the Web environment. After providing a brief introduction to basic Internet and Web technology, the textbook chapter examines the appropriateness of the Web as a database application platform and discusses the advantages and disadvantages of this approach. It then considers a number of the different approaches to integrating databases into the Web environment, including scripting languages, CGI, server extensions, Java, Active Server Pages, and Oracle's Internet Platform.

1.1 Time allocation

It is recommended that a maximum of two weeks should be devoted to this module.

Learning to program with the server-side scripting language PHP is mandatory to complete Assignment 2, and may take several weeks depending on your prior programming experience.

1.2 Relevance

This module is of the highest relevance to Web Information System developers, as it introduces the concept of integrating databases with web technology.

1.3 Possible difficulties

There is a significant amount of material in this chapter. Each topic has only been briefly covered. To have a deep understanding of any particular topic you will need to read further related references. However, many of you have already known many of these topics from other courses like *CSC2406 Web publishing* and *CSC8408 E-commerce technology*. Further, the Web is such a dynamic environment at present, and there are many new developments in this area. We will put up-to-date materials or relevant links on the Web.

The material covered in Sections 29.1 and 29.2 acts as a good introduction to the Web environment. However, by itself, the chapter does not provide sufficient material to allow you to create HTML or complete Web-DBMS applications, and so additional supplementary material has to be provided (see the bibliography and further readings for this chapter for some suggestions).

1.4 Learning hints

The Web itself is the source of an enormous amount of material on this topic and there are many Web sites with tutorials and examples of how databases can be accessed from the Web.

1.5 Activities

Exercise 1.1

You must obtain the free XAMPP software package (either from the Internet or from the Departmental DVD-Rom) and install it on your computer. XAMPP is available for both Windows and Linux.

Read the Intro book and course web site for more information.

Reading activity 1.1

Read the relevant sections of Connolly & Begg 2005, Chapter 29 ‘Web technology and DBMSs’.

Find and read related references from the Web. Read the following online references/readings to gain a further understanding of the material covered in this module. In particular, you need to read the following references on HTML , PHP and Database access in order to master the technical details on how to develop web sites.

1. <<http://www.w3schools.com/>>, tutorial materials on HTML etc
2. <<http://au.php.net/tut.php>>, a simple tutorial for learning to work with PHP.

You should also attempt the practical exercises posted on the course web site.

Module 2

Semi-structured data and XML

Objectives

In this module, you will learn

- what semi-structured data is, and how XML relates to it
- the main language features of XML
- how XML documents can be valid over a schema defined in the DTD or XML Schema languages
- what the purpose is of XML tools such as XPath and XSLT
- what XHTML is and why it should be used.

Learning resources

Textbook material that you should read:

- 30.1 Semi-structured data
- 30.2 Introduction XML
- 30.3 XML-related technologies
- 30.6 XML and databases

The other sections in Chapter 30 are optional.

Key concepts

- what semistructured data is
- the concepts of the Object Exchange Model (OEM), a model for semistructured data
- the basics of Lore, a semistructured DBMS, and its query language, Lorel
- the main language elements of XML
- the difference between well-formed and valid XML documents
- how Document Type Definitions (DTDs) can be used to define the valid syntax of an XML document
- how the Document Object Model (DOM) compares with OEM
- about other related XML technologies, such as Namespaces, XSL and XSLT, XPath, XPointer, XLink, and XHTML
- the limitations of DTDs and how the W3C XML Schema overcomes these limitations
- the proposals for a W3C Query Language

Introduction

The aim of this chapter is to examine semi-structured data and its relationship to XML. The chapter also examines XML, its related technologies, and query languages for XML. Semi-structured data (and hence also XML) constitutes a very different paradigm to represent data than the one most commonly known: the relational model. XML is mainly intended as a way to exchange data over a network, especially data that does not conform to the rigid tabular structure imposed by relational databases.

2.1 Time allocation

It is recommended that you devote a maximum of one week to this module.

2.2 Relevance

XML is becoming more and more important in the development of Web Information Systems, especially in building Web Services. So it is good to be aware of it, and be prepared to use it in large projects. However, in this course we build a dynamic web site without making use of XML: only a relational database and a server-side scripting language is used.

2.3 Possible difficulties

There is a significant amount of material in this chapter. Further, XML-related technologies and the XML query languages are still in development, so what you learn here is future-oriented and is not set in stone yet. You should not try to go into details, but obtain a general knowledge of what XML is, what it is intended to do, and what some of its more popular tools are for. The course *CSC8409 XML and semantic web services* offers a very broad look at all things XML.

2.4 Learning hints

As in the previous Module, the Web itself is a rich source of information on XML. There are many tutorials for learning XML, XSLT, XML Schema etc. A good source for up-to-date articles on the use of XML in Web Information Systems is <http://www.xml.com/>.

2.5 Activities

Reading activity 2.1

Read the relevant sections of Connolly & Begg 2005, Chapter 30 'Semi-structured data and XML'.

Find and read related references from the Web to have a deep understanding on related topics.

Check the updated material regarding XML standards at <http://www.w3c.org>

Exercise 2.1

Attempt exercises 30.17 and 30.18 from the textbook on page 1145.

Module 3

Query processing

Objectives

In this module, you will learn

- why query optimization is a vital task of a DBMS
- what the difference is between static and dynamic query optimization
- how relational algebra operations can be implemented and what their cost is
- what role statistics play in query optimization.

Learning resources

Textbook material that you should read:

- 21.1 Overview of query processing
- 21.2 Query decomposition
- 21.3 Heuristical approach to query optimization
- 21.4 Cost estimation for the relational algebra operations

The other sections in Chapter 21 are optional.

Key concepts

- what query optimization is and why it is so important
- how queries are decomposed, simplified, and restructured
- the use of transformation rules to rewrite relational algebra expressions into equivalent but more efficient expressions
- what heuristics are used to determine optimization strategies
- which statistics are used in query processing, and how
- what the costs are of the basic algebra operations (selection, projection, join)
- what the role of indices is in the cost estimation of operations

Introduction

The aim of this chapter is to examine how SQL queries submitted by database users get evaluated by the DBMS. The relational algebra plays a key role here: its transformation rules can guarantee that query statements can be rewritten so that they produce the exact same result in a much more efficient manner. Also detailed are rules-of-thumb (heuristics) that are used in query processing, and the statistics that a DBMS automatically compiles to further enhance query optimization. The chapter further examines the cost of processing individual algebra operations, in the presence or absence of indices defined on attributes in the tables that are being queried.

3.1 Time allocation

It is recommended that you devote a maximum of two weeks to this module.

3.2 Relevance

While query optimization is a task of the DBMS, a sophisticated user who understands how it works can assist the optimizer in its task by writing better SQL queries and hence increase the chance that queries will be evaluated in as little time as possible. And this is of course critical to the success of large-scale web information systems, where response times mean the difference between losing a customer and doing business.

In general, the more expensive the DBMS, the better it is at query optimization. That means that if you use cheap (or free) software, the burden of writing better SQL queries becomes larger! Applied to this course's second Assignment, this means that you should not expect MySQL to be very sophisticated in optimizing your queries. Of course, the project is very small, so optimization does not really become an issue.

Remember that static optimization can be extremely beneficial; therefore, it is useful to store your often-occurring queries as views in the database, so that query execution plans can be reused.

3.3 Possible difficulties

Some of the concepts in this module are far from easy. You must already have a good understanding of the Relational Algebra from a course such as *CSC3400 Database systems*, and also know what an index on a column (or set of columns) in a table is. Also important is that you understand the concept of commutativity, something you learnt in Algebra classes in secondary school.

Slightly less important is the involvement of various file and index storage techniques, such as B+ trees. It is also not necessary to learn the algorithms shown in the various figures of Section 21.4 in any detail. Rather, try to obtain a general understanding of the individual steps involved in query processing.

3.4 Learning hints

Some sophisticated DBMSs, such as Oracle, have a tool that lets you see the query execution plan for a submitted SQL query. If you have access to such a product, it may be worthwhile to spend some time working with this tool.

3.5 Activities

Reading activity 3.1

Read the relevant sections of Connolly & Begg 2005, Chapter 21 'Query processing'.

Exercise 3.1

Attempt exercises 21.16 and 21.18 from the textbook on pages 681–682.

Module 4

Transaction management

Objectives

In this module, you will learn

- the purpose of concurrency control and database recovery
- the function, importance, and properties of transactions
- the meaning and importance of serializability
- how the two-phase locking protocol (2PL) works
- what optimistic concurrency control is, and how it works
- how different levels of locking affects transaction throughput.

Learning resources

Material from the textbook that you should read:

- 20.1 Transaction support
- 20.2 Concurrent control
- 20.3 Database recovery
- 20.5.1 Oracle's Isolation Levels

Key concepts

- the purpose of concurrency control and database recovery
- the function, importance, and properties of transactions
- the meaning of serializability and how it applies to concurrency control
- how locks can be used to ensure serializability
- how the two-phase locking (2PL) protocol works
- the meaning of deadlock and how it can be resolved
- how timestamps can be used to ensure serializability
- how optimistic concurrency control techniques work
- how different levels of locking may affect concurrency
- the ACID properties
- the purpose of the transaction log file
- the purpose of checkpoints during transaction logging
- how to recover following database failure

Introduction

The aim of this module is to introduce three functions that a DBMS should provide: transaction management, concurrency control and recovery control. These functions are intended to ensure that the database is reliable and remains in a consistent state, when multiple users are accessing the database and in the presence of failures of both hardware and software components.

4.1 Time allocation

It is recommended that you devote a maximum of two weeks to this module.

4.2 Relevance

Web Information Systems frequently let thousands of users access a relational database almost simultaneously. Two things must be ensured in the presence of such behaviour: (1) the database must ensure that it remains in a consistent state at all times, giving users correct information and writing correct values to the database, and (2) all this is done as efficiently as possible, because undue overhead means losing business!

Both goals are somewhat in conflict with each other. It is important to understand that a database administrator can alter the concurrency behaviour of the database: either relaxing the correctness constraints to improve response times, or demand guaranteed correctness at the price of efficiency. He or she can also choose to select 'optimistic' concurrency control, which is both safe and fast, but may not be appropriate in all circumstances.

As a developer of Web Information Systems, you need to know about the various possibilities, and make educated decisions.

4.3 Possible difficulties

The transaction is fundamental to an understanding of concurrency and recovery control. However, some students are uncertain about what constitutes a transaction. There are a couple of examples in Section 20.1, but a few more examples will be included in the slides and selected readings to supplement these to reinforce the learning.

Next, it is important to understand that serializability is a theoretical ideal. It defines when the interweaving of the actions of different transactions is correct. In practise, however, it is expensive to enforce serializability, therefore often the condition is relaxed.

Different locking protocols achieve different levels of correctness – 2PL ensures serializability.

4.4 Learning hints

Make sure you understand what a transaction is and how it is perceived from users and DBMS perspectives.

Work through the overheads on the three concurrency problems: the **lost update problem**, the **uncommitted dependency problem**, and the **inconsistent analysis problem**. Then, study 2PL, go through how 2PL prevents these problems. And it is also useful to go through how the protocols for timestamping and optimistic techniques would prevent these problems for occurring, similar to Examples 20.6–20.8.

Make sure you understand that serializability is not itself a protocol for concurrency control.

When covering the material on recovery, it is useful to look at other examples similar to Example 20.11/20.12 to reinforce the learning.

4.5 Activities

Reading activity 4.1

Read the relevant sections of Connolly & Begg 2005, Chapter 20 ‘Transaction management’.

Also read Zhang, Y and Jia, X 1999, ‘Transaction processing’, in John Webster (ed.) *Wiley encyclopedia of electrical and electronics engineering*, John Wiley & Son, Inc.

Module 5

Distributed DBMSs

Objectives

In this module, you will learn

- the need for distributed databases, and the advantages and disadvantages of distributed DBMSs
- the functions that should be provided by DDBMSs
- issues in distributed database design: fragmentation, replication, and allocation.

Learning resources

Material from the textbook that you should read:

- 22.1 Introduction
- 22.3 Functions and architectures of a DDBMS
- 22.4 Distributed relational database design
- 22.5 Transparencies in a DDBMS
- 22.6 Date's twelve rules for a DDBMS

The other sections in Chapter 22 are optional.

Key concepts

- the differences between distributed database systems, distributed processing, and parallel database systems
- the advantages and disadvantages of DDBMSs
- the problems of heterogeneity in a distributed environment
- what fragmentation is, and how it should be carried out
- what the four main types of transparency in a DDBMS involve

Introduction

The aim of this module/chapter is to examine the concept of a physically distributed relational database. We first describe what it is in an informal way, and then look at the advantages and disadvantages of DDBMSs. In Sections 22.3 and 22.4 we delve a little deeper in some of the issues in DDBMSs: its functions and possible architectures, and importantly how to design a distributed relational database. Here the concept of fragmentation takes centre stage. We finish with a look at various transparencies that a DDBMS should offer, and with Date's 12 rules on DDBMSs.

5.1 Time allocation

It is recommended that a maximum of one week should be devoted to this module.

5.2 Relevance

Most web information systems only require a central RDBMS provided that it resides on a powerful server with a very large storage space, large memory, and fast processor(s). However, the largest web information systems require use of a distributed database. Hence it is useful for you to learn about them, albeit not in the level of detail that Chapter 23 of the textbook offers. Therefore, we only look at the material in Chapter 22.

5.3 Possible difficulties

Most of the material in this module is relatively easy to understand. It is not necessary to study the possible architectures listed in Section 22.3 in great detail.

Sections 22.4 (especially fragmentation) and 22.5 require a bit more attention.

5.4 Learning hints

It may be useful for you to revisit the material on relational algebra and on centralized relational database design before you attempt Section 22.4.

5.5 Activities

Reading activity 5.1

Read the relevant sections in Connolly & Begg 2005, Chapter 22 ‘Distributed DBMSs – concepts and design’.

Also read the lecture notes posted on the course website.

Module 6

Object-oriented and object-relational DBMSs

Objectives

In this module, you will learn:

- what the weaknesses are of traditional relational database systems
- the requirements for advanced database application
- a review of concepts in the Object-Oriented paradigm
- the problems associated with storing objects in a RDBMS.

Learning resources

Material from the textbook that you should read:

- 25.1 Advanced database applications
- 25.2 Weaknesses of traditional RDBMSs
- 25.3 Object-oriented concepts
- 25.4 Storing objects in a relational database
- 28.1 Introduction to ORDBMSs
- 28.3 Postgres – an early ORDBMS

All other sections of Chapters 25 and 28 are optional. We do not study Chapters 26 and 27.

Key concepts

- understand why traditional RDBMSs are not suitable for some advanced database applications
- learn what impedance mismatch is, and why it is a real problem
- understand why object-orientation in a database could solve these problems
- understand why OODBMS have not been successful from a commercial point of view
- understand what ORDBMSs are, and why and how they combine the strengths of RDBMSs and OODBMSs.

Introduction

In the previous modules (with the exception of module 2 on XML) we have only looked at advanced issues related to relational databases. This class of databases has been the most important for more than three decades now. However, current mainstream databases are not purely relational anymore. Instead, they are often so-called Object-Relational databases. In this module we first take a look at pure Object-Oriented databases, before turning to ORDBMSs. In the textbook, four chapters are used for this discussion. In this module, however, we deal mostly with the first part of Chapter 25 and two short sections in Chapter 28. The objective is to understand why ORDBMSs are currently the most popular databases, rather than knowing all details of them.

6.1 Time allocation

It is recommended that a maximum of one week should be devoted to this module.

6.2 Relevance

As web information systems typically store not only tabular information, but also multimedia and other complex data, traditional relational database systems will often be insufficient for its purposes. Most WISs therefore use an Object-Relational database such as Oracle and PostgreSQL instead. It is good to know what such systems offer, so that you build better web applications.

6.3 Possible difficulties

Most concepts covered in this module are not hard to understand, especially if you already know about Object-Oriented programming languages.

Some students wonder what the difference is between XML databases and OODBMSs. First, XML databases are only in their infancy — many issues need to be resolved before they become commercially viable. On the other hand, while OODBMSs are mature, they have not done very well commercially. Secondly, an XML database stores semi-structured data having no schema, or a fast-changing and irregular one. Object-Oriented databases, like relational ones, however, do impose a relatively static schema on the data they store. Thirdly, impedance mismatch is again a problem in the context of XML databases, but was solved in OODBMSs.

6.4 Learning hints

For those wanting to understand ORDBMSs better, it is recommended to experiment with one. If you do not have access to Oracle or another big commercial system, PostgreSQL is an excellent and free alternative. It can be downloaded from their website.

6.5 Activities

Reading activity 6.1

Read the relevant sections in Connolly & Begg 2005, Chapters 25 and 28.

Also read the lecture notes posted on the course website.

Module 7

Data warehousing

Objectives

In this module, you will learn

- the main concepts, benefits, and problems associated with data warehousing
- the difference between On-Line Transaction Processing (OLTP) and data warehousing
- the architecture and main components of a data warehouse
- the issues associated with designing a data warehouse
- the differences between the Dimensional Model and the Entity-Relationship Model.

Learning resources

Material from the textbook that you should read:

- 31.1 Introduction to data warehousing
- 31.2 Data warehouse architecture
- 32.1 Designing a data warehouse database
- 32.2 Dimensionality modelling
- 32.3 Design methodology

The other sections of Chapters 31 and 32 are optional.

Key concepts

- concepts and benefits of data warehousing
- comparison of OLTP systems and data warehousing
- problems of data warehousing
- main components of the data warehouse architecture
- Star schema, Snowflake schema, and Starflake schema
- nine steps of designing a data warehouse.

Introduction

After we briefly ventured out of the traditional relational database systems in the previous module, we now look at some relatively new applications of relational databases. Decision support is very important to large organizations – ‘we have all this data, but what is it telling me about my business processes?’ In addition to regular queries, users should be able to easily analyse the data in the DBMS and create reports suitable to support the decision-making process of managers. Hence, the focus shifts from a transactional database to a data store that can be analysed. This is especially important when data (e.g. information about sales) is highly dimensional, and can thus be viewed from many different perspectives.

7.1 Time allocation

It is recommended that you devote a maximum of one week to this module.

7.2 Relevance

There is no immediate connection between Web Information Systems and data warehouses—it is not common for analysts to use a web interface to query the underlying database for highly dimensional data-analysis purposes. However, you should be aware that such systems exist, because your company will want to use your database for these purposes as well.

7.3 Possible difficulties

Since we are only taking an introductory look at data warehouses, most concepts in this module are easy to understand. It is recommended to spend some time studying the diverse schema in Section 32.1, and understand the differences between them.

7.4 Learning hints

High-end commercial database systems such as Oracle and DB2 have built-in data warehousing capabilities. Unfortunately I do not know of any open source or otherwise free products that have this. If you do have access to one, it may be worthwhile to investigate how it works.

7.5 Activities

Reading activity 7.1

Read the relevant sections of Connolly & Begg 2005, Chapters 31 and 32.

Module 8

OLAP and data mining

Objectives

In this module, you will learn

- the purpose of On-Line Analytical Processing (OLAP)
- the relationship between OLAP and data warehousing
- the key features and benefits of OLAP
- how to represent multi-dimensional data
- rules and categories of OLAP tools
- the concepts associated with data mining
- the main types of data mining operations.

Learning resources

Material from the textbook that you should read:

- 33.1 Online analytical processing
- 33.2 OLAP applications
- 33.3 Representation of multi-dimensional data
- 33.4 OLAP tools
- 34.1 Data mining
- 34.2 Data mining techniques
- 34.5 Data mining and data warehousing

The other sections of Chapters 33 and 34 are optional.

Key concepts

- what OLAP is and what its benefits are
- how Multi-Dimensional Data can be represented
- what drilling-down and slicing means
- the differences between OLAP, MOLAP, and ROLAP
- the four main operations associated with data mining techniques

Introduction

The previous module introduced the concept of the data warehouse as a large collection of data possibly from diverse sources, intended for analysis rather than transactional use. In this module we look at two different but complementary decision-support techniques applied to data warehouses:

OLAP and data mining. Again we only introduce the concepts. The course *CSC3417 Data mining discovers knowledge* looks at data mining techniques in much greater detail.

8.1 Time allocation

It is recommended that you devote a maximum of one week to this module.

8.2 Relevance

There is no immediate connection between Web Information Systems, data warehouses, OLAP, and data mining. However, you should be aware that such systems exist, because your company will want to use your database for these purposes as well.

8.3 Possible difficulties

Since we are only taking an introductory look at OLAP and Data Mining, most concepts in this module are easy to understand.

8.4 Learning hints

High-end commercial database systems such as Oracle and DB2 have built-in OLAP and Data Mining capabilities. Unfortunately I do not know of any open source or otherwise free products that have this. If you do have access to one, it may be worthwhile to investigate how it works.

8.5 Activities

Reading activity 8.1

Read the relevant sections of Connolly & Begg 2005, Chapters 33 and 34.

Module 9

Information retrieval and web search

Objectives

In this module you will learn:

- what Information Retrieval is, and what some of its properties are
- how IR fundamentally differs from querying in (relational and other) databases
- why IR is so important in the context of the World Wide Web
- what Web Search is
- an overview of how Google works
- how the Page Rank algorithm works.

Learning resources

The textbook does not present material relevant to this module.

The study materials are some slides and research papers to be provided on the course home page. A good reference for this module is *Modern information retrieval* by Baeza-Yates and Riberiro-Neto. In particular, the materials from the following Chapters will be covered: Chapters 1, 2, 3 and 13. As you may not have this book, we will put the presentation slides and some related materials on the Web.

The slides and additional materials for this book can be found at: <http://www.sims.berkeley.edu/~hearst/irbook/>

Key concepts

- IR retrieves information from data sources (usually text, but possibly also multimedia) that are highly unstructured; hence not from databases
- IR queries return approximate results only, not exact results as returned by database queries
- IR searches and indexes text sources based on their content, not their schema
- The page rank algorithms plays a key role in how Google works

Introduction

The aim of this module is to introduce the topics on web-based information retrieval and web search.

9.1 Time allocation

It is recommended that a maximum of two weeks should be devoted to this module.

9.2 Relevance

Good Web Information Systems have the ability to let users search the site (or other sources) for material that is not stored in a database, but in web pages or other text files. Often commercial or free web search tools are used for this. They all amount to retrieval of information from highly unstructured sources. It is therefore necessary to understand how this works, and how it is fundamentally different from the querying in relational databases.

9.3 Possible difficulties

Some concepts in this module are not easy to understand right away, but require some focused studying. For database students, the concepts are also quite foreign which may initially complicate understanding. The Page Rank algorithm is quite complex if you go into details, but is rather understandable if you first try to understand what it tries to do in general.

9.4 Learning hints

To consolidate your understanding of web search, it may be worthwhile to experiment with your favourite Web search engine: try out different types of queries, with varying degrees of complexity, and see how the results change.

9.5 Activities

Reading activity 9.1

If you can find the recommended reference book in the Library, read at least the introductory sections of Chapters 1, 2, 3, 8, and 13. Also, take a look at the websites for both that book and this course, and read the lecture material.