UNSW Course Outline

COMP9319 Web Data Compression and Search - 2024

Published on the 21 May 2024

General Course Information

Course Code : COMP9319
Year : 2024
Term : Term 2
Teaching Period : T2
Is a multi-term course? : No
Faculty : Faculty of Engineering
Academic Unit : School of Computer Science and Engineering
Delivery Mode : Multimodal
Delivery Format : Standard
Delivery Location : Kensington
Campus : Sydney
Study Level : Postgraduate, Undergraduate
Units of Credit : 6

Useful Links

Handbook Class Timetable

Course Details & Outcomes

Course Description

As the amount of Web data increases, it is becoming vital to not only be able to search and retrieve this information quickly, but also to store it in a compact manner. This is especially important for mobile devices which are becoming increasingly popular. Without loss of generality, within this course, we assume Web data (excluding media content) will be in XML and its like (e.g., HTML, JSON).

If time allows, we may cover optional topics such as: streaming algorithms, text analytics, Web data optimization for mobile devices. The lecture materials will be complemented by two programming assignments and numerous tutorial-type, written exercises.

Course Aims

This course aims to introduce the concepts, theories, and algorithmic issues important to Web data compression and search. The course will also introduce the most recent development in various areas of Web data optimization topics, common practice, and its applications. The course is composed of the following parts:

- Adaptive coding, information theory
- Text compression (zip, gzip, bzip, etc)
- Burrows-Wheeler Transform and backward search
- XML compression
- Indexing
- Pattern matching and regular expression search
- Distributed querying
- Fast index construction
- Implementation

Course Learning Outcomes

Course Learning Outcomes
CLO1 : Apply the fundamentals of text compression
CLO2 : Apply advanced data compression techniques such as those based on Burrows Wheeler Transform
CLO3 : Write computer programs for Web data compression and search with optimization
CLO4 : Use selected XML processing and optimization techniques
CLO5 : Analyze the advantages and disadvantages of data compression for Web search
CLO6 : Apply basic techniques from XML distributed query processing
CLO7 : Discuss the past, present, and future of data compression and Web data optimization

Course Learning Outcomes	Assessment Item
CLO1 : Apply the fundamentals of text compression	• Assignment 1 • Assignment 2 • Final Examination
CLO2 : Apply advanced data compression techniques such as those based on Burrows Wheeler Transform	• Assignment 1 • Assignment 2 • Final Examination
CLO3 : Write computer programs for Web data compression and search with optimization	• Assignment 1 • Assignment 2
CLO4 : Use selected XML processing and optimization techniques	• Final Examination • Assignment 1 • Assignment 2
CLO5 : Analyze the advantages and disadvantages of data compression for Web search	• Final Examination • Assignment 1 • Assignment 2
CLO6 : Apply basic techniques from XML distributed query processing	• Final Examination
CLO7 : Discuss the past, present, and future of data compression and Web data optimization	• Final Examination

Learning and Teaching Technologies

Moodle - Learning Management System | Echo 360 | EdStem | Blackboard Collaborate

Assessments

Assessment Structure

Assessment Item	Weight	Relevant Dates
Assignment 1 Assessment Format: Individual	15%	Due Date: Week 5: 24 June - 30 June
Assignment 2 Assessment Format: Individual	35%	Due Date: Week 9: 22 July - 28 July
Final Examination Assessment Format: Individual	50%	Start Date: Not Applicable Due Date: During Exam Period

Assessment Details

Assignment 1

Assessment Overview

This is a warm-up programming assignment for the course. Hence it will be relatively lightweight (students are expected to be able to finish the assignment in a few hours).

Assessment of assignments will be primarily based on how accurately they satisfy the requirements; this means that most of the marks will be based on automatic marking. However, we may also manually examine submitted assignments to determine (a) whether they are written with good style, (b) how closely they satisfied the requirements, if time allows.

Individual graded results with optional comments will be emailed to each student. Overall feedbacks will be discussed in the lectures, and students may discuss with the tutors in consultation sessions for further assessment feedbacks.

Course Learning Outcomes

- CLO1 : Apply the fundamentals of text compression
- CLO2 : Apply advanced data compression techniques such as those based on Burrows Wheeler Transform
- CLO3 : Write computer programs for Web data compression and search with optimization
- CLO4 : Use selected XML processing and optimization techniques
- CLO5 : Analyze the advantages and disadvantages of data compression for Web search

Assignment 2

Assessment Overview

This is the second programming assignment for the course. Hence it will be relatively heavier weight since it involves more advanced techniques that students have learnt from the course (students are expected to be able to finish the assignment in a few days).

Assessment of assignments will be primarily based on how accurately they satisfy the requirements; this means that most of the marks will be based on automatic marking. However, we may also manually examine submitted assignments to determine (a) whether they are written with good style, (b) how closely they satisfied the requirements, if time allows.

Individual graded results with optional comments will be emailed to each student. Overall feedbacks will be discussed in the lectures, and students may discuss with the tutors in consultation sessions for further assessment feedbacks.

Course Learning Outcomes

- CLO1 : Apply the fundamentals of text compression
- CLO2 : Apply advanced data compression techniques such as those based on Burrows Wheeler Transform
- CLO3 : Write computer programs for Web data compression and search with optimization
- CLO4 : Use selected XML processing and optimization techniques
- CLO5 : Analyze the advantages and disadvantages of data compression for Web search

Final Examination

Assessment Overview

The final exam will be a major assessment in this course and aims to test what students learned about data compression and search during the course of the semester. To pass this course, students are required to have satisfactory performance on the final exam even if they do very well on the assignments. In order to meet the hurdle requirement, students must score better than 40% on the final exam. Note that the hurdle will be enforced after any required scaling.

Course Learning Outcomes

- CLO1 : Apply the fundamentals of text compression
- CLO2 : Apply advanced data compression techniques such as those based on Burrows Wheeler Transform
- CLO4 : Use selected XML processing and optimization techniques
- CLO5 : Analyze the advantages and disadvantages of data compression for Web search
- CLO6 : Apply basic techniques from XML distributed query processing
- CLO7 : Discuss the past, present, and future of data compression and Web data optimization

Assignment submission Turnitin type

Not Applicable

Hurdle rules

To pass this course, students are required to have satisfactory performance on the final exam even if they do very well on the assignments. In order to meet the hurdle requirement, students must score better than 40% on the final exam. Note that the hurdle will be enforced after any required scaling.

General Assessment Information

Assignments will be completed *individually* ; this means that you should do them *yourself* without assistance from others, except for asking advice from the Lecturer or Tutor. As noted above, assignments are the primary vehicle for learning the material in this course. If you don't do them, or simply copy and submit someone else's work, you have wasted a valuable learning opportunity.

Assignments are to be submitted via "give" before the specified time on the due date. Assessment of assignments will be primarily based on how accurately they satisfy the requirements; this means that most of the marks will be based on automatic marking. However, we may also manually examine submitted assignments to determine (a) whether they are written with good style, (b) how closely they satisfied the requirements, if time allows.

The penalty for late submission of assignments will be 5% (of the worth of the assignment) subtracted from the raw mark per day of being late. In other words, earned marks will be lost. For example, assume an assignment worth 20 marks is marked as 18, but had been submitted two days late. The late penalty will be 2 marks, resulting in a mark of 16 being awarded. **No assignments will be accepted later than 5 days after the original deadline.** For example, if you have your special consideration granted by UNSW for a one-week extension, there will be no late penalty if the assignment is submitted within 7 days after the original deadline. However, no further late submissions will be accepted after these 7 days.

Grading Basis

Standard

Course Schedule

Teaching Week/Module	Activity Type	Content
Week 1 : 27 May - 2 June	Lecture	Introduction, basic information theory, basic compression
Week 2 : 3 June - 9 June	Lecture	More basic compression algorithms
Week 3 : 10 June - 16 June	Lecture	Adaptive Huffman; Overview of BWT
Week 4 : 17 June - 23 June	Lecture	Pattern matching and regular expression
Week 5 : 24 June - 30 June	Lecture	FM index, backward search, compressed BWT
Week 7 : 8 July - 14 July	Lecture	Suffix tree, suffix array, the linear time algorithm
Week 8 : 15 July - 21 July	Lecture	XML overview; XML compression
Week 9 : 22 July - 28 July	Lecture	Graph compression; Distributed Web query processing
Week 10 : 29 July - 4 August	Lecture	Optional advanced topics; Course Revision

Attendance Requirements

Students are strongly encouraged to attend all classes and review lecture recordings.

General Schedule Information

The course schedule is an **approximate** guide to the sequence of topics in this course. It is subject to change as the term progresses.

Course Resources

Recommended Resources

There will be no textbook used in this course. Lecture slides and supplementary readings will be provided and used.

You may find the readings below useful as reference materials:

- Managing Gigabytes: Compressing and Indexing. Documents and Images, Second Edition. Ian H. Witten, Alistair Moffat, Timothy C. Bell, Morgan Kaufmann, 1999. (recommended reference, available at the university bookstore)
- Search Engines: Information Retrieval in Practice. W. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.
- http://www.data-compression.info contains lots of valuable resources on data compression (especially links to readings and useful advice), despite the website's pink color!
- Data on the Web: from relations to semistructured data and XML. Serge Abiteboul, Peter Buneman, Dan Suciu. Morgan Kaufmann, 2000.

You will also find your previous textbooks on data structures and/or algorithms useful, in case you need to refer to the fundamentals of data structures and algorithms for text processing.

Course Evaluation and Development

This course is evaluated each session using MyExperience.

The MyExperience evaluation from the last time I taught this course showed that students were overall satisfied with all aspects of the course. Thus we maintain a similar style and structure for this term. Since this is the second time that we run this course after the pandemic (from totally online back to hybrid mode), we will go through the in-depth topics in the recorded lectures and discuss more examples and/or practical considerations in the live lectures (mixed online & in person. Please note that your feedback is important and will be considered to improve future offerings of this course (e.g., how much content can remain online).

Students are also encouraged to provide informal feedback during the term and let the lecturer know of any problems, as soon as they arise. Suggestions will be listened to very openly, positively, constructively, and thankfully, and every reasonable effort will be made to address them as soon as possible.

Staff Details

Position	Name	Email	Location	Phone	Availability	Equitable Learning Services Contact	Primary Contact
Convenor	Raymond Wong					Yes	Yes

Other Useful Information

Academic Information

I. Special consideration and supplementary assessment

If you have experienced an illness or misadventure beyond your control that will interfere with your assessment performance, you are eligible to apply for Special Consideration prior to, or within 3 working days of, submitting an assessment or sitting an exam.

Please note that UNSW has a Fit to Sit rule, which means that if you sit an exam, you are declaring yourself fit enough to do so and cannot later apply for Special Consideration.

For details of applying for Special Consideration and conditions for the award of supplementary assessment, please see the information on UNSW's Special Consideration page.

II. Administrative matters and links

All students are expected to read and be familiar with UNSW guidelines and polices. In particular, students should be familiar with the following:

- Attendance
- UNSW Email Address
- Special Consideration
- Exams
- Approved Calculators
- Academic Honesty and Plagiarism
- Equitable Learning Services

III. Equity and diversity

Those students who have a disability that requires some adjustment in their teaching or learning environment are encouraged to discuss their study needs with the course convener prior to, or at the commencement of, their course, or with the Equity Officer (Disability) in the Equitable Learning Services. Issues to be discussed may include access to materials, signers or note-takers, the provision of services and additional exam and assessment arrangements. Early notification is essential to enable any necessary adjustments to be made.

IV. Professional Outcomes and Program Design

Students are able to review the relevant professional outcomes and program designs for their streams by going to the following link: https://www.unsw.edu.au/engineering/student-life/student-resources/program-design.

Note: This course outline sets out the description of classes at the date the Course Outline is published. The nature of classes may change during the Term after the Course Outline is published. Moodle or your primary learning management system (LMS) should be consulted for the up-to-date class descriptions. If there is any inconsistency in the description of activities between the University timetable and the Course Outline/Moodle/LMS, the description in the Course Outline/Moodle/LMS applies.

Academic Honesty and Plagiarism

UNSW has an ongoing commitment to fostering a culture of learning informed by academic integrity. All UNSW students have a responsibility to adhere to this principle of academic integrity. Plagiarism undermines academic integrity and is not tolerated at UNSW. *Plagiarism at*

UNSW is defined as using the words or ideas of others and passing them off as your own.

Plagiarism is a type of intellectual theft. It can take many forms, from deliberate cheating to accidentally copying from a source without acknowledgement. UNSW has produced a website with a wealth of resources to support students to understand and avoid plagiarism, visit: student.unsw.edu.au/plagiarism. The Learning Centre assists students with understanding academic integrity and how not to plagiarise. They also hold workshops and can help students one-on-one.

You are also reminded that careful time management is an important part of study and one of the identified causes of plagiarism is poor time management. Students should allow sufficient time for research, drafting and the proper referencing of sources in preparing all assessment tasks.

Repeated plagiarism (even in first year), plagiarism after first year, or serious instances, may also be investigated under the Student Misconduct Procedures. The penalties under the procedures can include a reduction in marks, failing a course or for the most serious matters (like plagiarism in an honours thesis or contract cheating) even suspension from the university. The Student Misconduct Procedures are available here:

www.gs.unsw.edu.au/policy/documents/studentmisconductprocedures.pdf

Submission of Assessment Tasks

Work submitted late without an approved extension by the course coordinator or delegated authority is subject to a late penalty of five percent (5%) of the maximum mark possible for that assessment item, per calendar day.

The late penalty is applied per calendar day (including weekends and public holidays) that the assessment is overdue. There is no pro-rata of the late penalty for submissions made part way through a day. This is for all assessments where a penalty applies.

Work submitted after five days (120 hours) will not be accepted and a mark of zero will be awarded for that assessment item.

For some assessment items, a late penalty may not be appropriate. These will be clearly indicated in the course outline, and such assessments will receive a mark of zero if not completed by the specified date. Examples include:

- Weekly online tests or laboratory work worth a small proportion of the subject mark;
- Exams, peer feedback and team evaluation surveys;
- Online quizzes where answers are released to students on completion;
- Professional assessment tasks, where the intention is to create an authentic assessment that has an absolute submission date; and,
- Pass/Fail assessment tasks.

Faculty-specific Information

Engineering Student Support Services – The Nucleus - enrolment, progression checks, clash requests, course issues or program-related queries

Engineering Industrial Training – Industrial training questions

UNSW Study Abroad – study abroad student enquiries (for inbound students)

UNSW Exchange – student exchange enquiries (for inbound students)

UNSW Future Students – potential student enquiries e.g. admissions, fees, programs, credit transfer

Phone

(+61 2) 9385 8500 – Nucleus Student Hub

(+61 2) 9385 7661 – Engineering Industrial Training

(+61 2) 9385 3179 – UNSW Study Abroad and UNSW Exchange (for inbound students)

School Contact Information

CSE Help! - on the Ground Floor of K17

- For assistance with coursework assessments.

The Nucleus Student Hub - https://nucleus.unsw.edu.au/en/contact-us

- Course enrolment queries.

Grievance Officer - grievance-officer@cse.unsw.edu.au

- If the course convenor gives an inadequate response to a query or when the courses convenor does not respond to a query about assessment.

Student Reps - stureps@cse.unsw.edu.au

- If some aspect of a course needs urgent improvement. (e.g. Nobody responding to forum queries, cannot understand the lecturer)

You should **never** contact any of the following people directly:

- Vice Chancellor

- Pro-vice Chancellor Education (PVCE)

- Head of School

- CSE administrative staff

- CSE teaching support staff

They will simply bounce the email to one of the above, thereby creating an unnecessary level of indirection and a delay in the response.