University of Macau
Department of Computer & Information Science
CISB363 – Information Retrieval and Web Mining
Syllabus
2nd Semester 2014/2015
Part A – Course Outline

Elective course in Computer Science

Catalog description:

3.0 credits / 3.5 weekly contact hours (lecture: 1.75 hours, tutorial: 1.75 hours). This course will introduce the latest development of information retrieval and web mining technologies. In the first part of the course, we will overview the fundamental concepts of information retrieval, such as crawling, parsing, indexing, searching, scoring, and compression. These techniques enable students to handle web scale datasets. In the second part, we will discuss how to extract knowledge from web scale datasets by link analysis, clustering, and recommendation techniques. Moreover, some latest implementation techniques (such as Apache *Hadoop*, *Pig*, and *Lucene*) will be studied thoroughly by the course project. The course is aimed at helping students to explore the latest techniques in information retrieve and web mining. Some research oriented projects will be given according to students' background knowledge. The contents of the course will mix with lectures, tutorials, and group discussions.

Course type:

Theoretical with substantial laboratory/practice content

Prerequisites:

- CISB120. Algorithms and Data Structures I

Textbook(s) and other required material:

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search, 2nd Edition*, ACM Press Books.

References:

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to information retrieval*, Singapore : Cambridge University Press, 2009.
2. Stefan Buttcher, Charles L. A. Clarke, Gordon V. Cormack, *Information retrieval: implementing and evaluating search engines*, Cambridge, MA : MIT Press, c2010.
3. Heting Chu, *Information Representation and Retrieval in the Digital Age, Second Edition*, Information Today, Inc.
4. Related International Journals
   - ACM Transactions on Information Systems (TOIS)
   - ACM Transactions on Knowledge Discovery on Data (TKDD)
   - International Journal of Cooperative Information Systems (IJCIS)
   - ACM Transactions on Database Systems (TODS)
   - The VLDB Journal
   - IEEE Transactions on Knowledge and Data Engineering (TKDE)
5. Related International Conferences
   - ACM Special Interest Group on Information Retrieval (SIGIR)
   - International World Wide Web Conference (WWW)
   - ACM International Conference on Web Search and Data Mining (WSDM)
   - ACM International Conference on Information and Knowledge Management (CIKM)
   - ACM Special Interest Group on Knowledge Discovery and Data Mining (KDD)
   - ACM Special Interest Group on Management Of Data (SIGMOD)
   - Very Large Data Bases (VLDB) Conference
   - International Conference on Data Engineering (ICDE)

- International Conference on Extending Database Technology (EDBT)

Major prerequisites by topic:

1. Fundamental algorithmic analysis and design of different programming constructs and data structures.
2. Language proficiency in at least one procedural language, e.g., C, C++, or Java.

Course objectives:

- Learn the fundamental concepts of information retrieval systems [a, c]
- Introduce the theory behind the major components of information retrieval, including crawling, parsing, scoring, indexing, and compression [a, c]
- Learn latest techniques of web mining [a, c]
- Design and implement parts of an information retrieval system [a, c, h]

Topics covered:

1. **Introduction of information retrieval (4 hours)**
   - Evolution of information retrieval in past decade
   - The challenges of information retrieval
2. **Fundamentals of information retrieval (12 hours)**
   - Term vocabulary and posting lists (inverted lists)
   - Index construction and compression
   - Scoring, term weighting, and vector space model
   - Evaluation in information retrieval
3. **Web mining (8 hours)**
   - Link analysis
   - Clustering and classification
   - Recommendation
4. **Data management in the cloud (4 hours)**
   - MapReduce (Apache Hadoop, Pig, Mahout)

Class/laboratory schedule:

Timetabled work in hours per week			No of teaching weeks	Total hours	Total credits	No/Duration of exam papers
Lecture	**Tutorial**	**Practice**				
2	2	0	14	49	3.0	1 / 3 hours

Student study effort required:

Class contact:	
Lecture	24.5 hours
Tutorial	24.5 hours
Practice	0 hours
Other study effort	
Self-study	20 hours
Homework assignment	12 hours
Project / Case study	25 hours
Total student study effort	106 hours

Student assessment:

Final assessment will be determined on the basis of:

Homework	20%	Mid-term	20%
Project	30%	Final Exam	30%

Course assessment:

The assessment of course objectives will be determined on the basis of:
Assignments, project, exams and, course evaluation

Course outline:

Weeks	Topic	Course work
1-2	**Introduction** Overview the history and development of information retrieval and web mining. Introduce the difficulty of handling web scale datasets. The necessities of information retrieval and web mining.	
3-4	**Fundamentals of information retrieval** Study fundamental concepts of information retrieval systems, including term vocabulary and posting lists (inverted lists), parsing and crawling.	Assignment#1
5-6	**Evaluation in information retrieval** Scoring, term weighting, and vector space model.	
7-8	**Indexing** Introduce index construction and compression. Discuss efficient query evaluation over web scale datasets	Assignment#2 Project
9	**Midterm**	
10	**Fundamentals of Web Mining** Overview the concept of web mining. Study the challenges of knowledge extraction from web scale datasets.	
11-13	**Web Mining** Link analysis, clustering, classification, and recommendation	Assignment#3
14	**Presentation**	

Contribution of course to meet the professional component:

This course prepares students to learn the state-of-the-art techniques in information retrieval and web mining.

Relationship to CIS program objectives and outcomes:

This course primarily contributes to CIS program outcomes that develop student abilities to:

(a) An ability to apply knowledge of computing and mathematics appropriate to the programme outcomes and to the discipline;

(c) An ability to analyse a problem, and identify and define the computing requirements appropriate to its solution;

The course secondarily contributes to CIS program outcomes that develop student abilities to:

(h) An ability to analyse the local and global impact of computing on individuals, organisations, and society;

Course content distribution:

Percentage content for			
Mathematics	Science and engineering subjects	Complementary electives	**Total**
0%	100%	0%	100%

Coordinator:

Zhiguo Gong, Associate Professor of CIS Program

Persons who prepared this description:

Leong Hou U, January 14, 2015

Part B General Course Information and Policies

1st/2nd Semester 201x/201y

Instructor: Leong Hou U
Office: B1-A704
Office Hour: (*to be announced*)
Phone: 83978469
Email: ryanlhu@umac.mo

Time/Venue: (*to be announced*)

Grading Distribution:

Percentage Grade	Final Grade	Percentage Grade	Final Grade
100 - 93	A	92 - 88	A–
87 - 83	B+	82 - 78	B
77 - 73	B–	72 - 68	C+
67 - 63	C	62 - 58	C–
57 - 53	D+	52 - 50	D
below 50	F		

Comment:
The objectives of the lectures are to explain and to supplement the text material. Students are responsible for the assigned material whether or not it is covered in the lecture. Students who wish to succeed in this course should read the assignments prior to the lecture and should work all homework and lab assignments. You are encouraged to look at other sources (other texts, etc.) to complement the lectures and text.

Homework Policy:
The completion and correction of homework is a powerful learning experience; therefore:

- There will be approximately 3 homework assignments.
- Homework is due one or two weeks after assignment unless otherwise noted, no late homework is accepted.
- Possible revision of homework grades may be discussed with the grader within one week from the return of the marked homework
- The course grade will partly be based on the average of the HW grades.

Project policy:

The project is probably the most exciting part of this course and provides students with meaningful experience to extend and enhance an existing compiler and interpreter:

- You will work with group of at most three students for the course project.
- The requirements will be announced and discussed in class.
- The project will be presented at the end of semester.

Quizzes
One mid-term exam will be held during the semester.

Note

- Recitation session is important part of this course and attendance is strongly recommended.
- Check UMMoodle for announcement, homework and lectures. Report any mistake on your grades within one week after posting.
- No make-up exam is given except for CLEAR medical proof.
- No exam is given if you are 15 minutes late in the midterm exams and 30 minutes late in the final exam. Even if you are late in the exam, you must turn in at the due time.
- Cheating is absolutely prohibited by the university, and is subject to University disciplinary actions.

Appendix:

Rubric for Program Outcomes

Rubric for (a)	5 (Excellent)	3 (Average)	1 (Poor)
Understand the theoretic background	Students understand theoretic background and the limitations of the respective applications.	Students have some confusion on some background or do not understand theoretic background completely.	Students do not understand the background or do not study at all.

Rubric for (c)	5 (Excellent)	3 (Average)	1 (Poor)
Design capability and design constraints	Student understands very clearly what needs to be designed and the realistic design constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability.	Student understands what needs to be designed and the design constraints, but may not fully understand the limitations of the design constraints.	Student does not understand what needs to be designed and the design constraints.

Rubric for (e)	5 (Excellent)	3 (Average)	1 (Poor)
Modeling, problem formulation and problem solving	Students choose and properly apply the correct techniques.	Students model correctly but cannot select proper technique or model incorrectly but solve correctly accordingly.	Students at loss as to how to solve a problem.

Rubric for (k)	5 (Excellent)	3 (Average)	1 (Poor)
Use modern principles, skills, and tools in engineering practice	Student applies the principles, skills and tools to correctly model and analyze engineering problems, and understands the limitations.	Student applies the principles, skills and tools to analyze and implement engineering problems.	Student does not apply principles and tools correctly and/or does not correctly interpret the results.