Web Information Retrieval and Data Mining

Department(s)	**Computer Science**
Career	**[ ] Undergraduate [x] Graduate**
Academic Level	**[x] Regular [ ] Compensatory [ ] Developmental [ ] Remedial**
Subject Area	
Course Number	**CSc 83060**
Course Title	**Web Information Retrieval and Data Mining**
Catalogue Description	Database Management Systems (DBMS) are vital components of modern information systems serving every type of organizations. We can hardly envision any computer application that does not utilize a DBMS. Database applications are pervasive and range in size from in-memory databases to terra bytes or even larger in various applications domains such as commercial, spatial, biological, scientific applications. The course is designed to develop an understanding of the fundamental concepts and issues in database research and extend it to knowledge representation.
Pre/ Co Requisites	**n/a**
Credits	**3**
Contact Hours	**3**
Liberal Arts	**[ ] Yes [x] No**

Rationale:

Information retrieval, data mining, as well as Web information processing are important driving forces for both research and industrial development in not only Computer Science, but also our economy at large in the past two decades, and remain this way in the foreseeable future. A course that provides state of the art knowledge, in depth discussions, and challenging research topics in these areas help to get our students better prepared in both academia and industry in their future endeavor.

Course Description: Information Retrieval (IR) is the process of extracting relevant documents or their parts from larger quantities of documents based on a query presented by the user. Data Mining (DM) is the process of analyzing large volumes of data using pattern recognition or knowledge discovery techniques to find meaningful information that is hidden within available data (such as trends and implicit relationships). Traditional IR and DM focus more on structured data stored in databases. However, databases are not the only means for the storage of information. The World Wide Web (WWW), as a global distributed information repository, has become the largest data sources in today's world. With the great impact of the WWW, Web information processing has become one of the hottest research topics in both academic and industry.

The major differences in between normal databases and the WWW is that Web information are semi-structured, which makes IR and DM more difficult. However, the Web also provides some useful information such as hyperlinks, presentation structures, and user visiting patterns, which are unavailable from normal databases. All these make Web IR and DM quite different from traditional IR and DM. Consisting of five parts, this course mainly discusses technologies of IR and DM on Web information. The first part is about Web information storage and presentation schemes. The second and third parts discuss basic IR and DM technologies. The fourth and fifth parts discuss how to make use of the semi-

structured/heterogeneous data, hyperlinks information, and user visiting patterns on the World Wide Web for Web IR and DM. In addition, this course will also cover the topic of Web Information Extraction.

Learning Goals/Outcomes:

After successful completion of this course,

Theoretically, you are expected to understand

- Web information storage and presentation schemes
- Basic IR technologies: modeling, indexing, searching, etc.
- Basic DM technologies: Association mining, Classification, Clustering, etc.
- Web IR technologies: content structure/visual based Web IR, link based Web IR, User centered Web IR, etc.
- Web DM technologies: Web usage mining, Web structure mining, Web content mining, etc.
- Web Information Extraction: HMM based Web IE

Practically, you are expected to produce

- A demonstration prototype for a selected topic on Web IR/DM
- A paper of publishable quality on Web IR/DM

Assessment:

❖ **Small projects**

There will be four small projects for this course. For each small project you are expected to design a small software prototype in each area of basic/Web IR/DM, and analyze the testing result of your prototype.
Unless otherwise specified, for each project, you need only submit electronic version via Blackboard.

❖ **Big project**

Each student is expected to select a special research topic in Web IR/DM and design a new algorithm for the area selected. You are encouraged to work in group (generally two members in one group) and discuss with the instructor about the topics that you are interested in earlier in the semester. A software prototype should be implemented to demonstrate your algorithm. Based on your research work, you are also expected to produce a research paper of publishable quality.

❖ **Quizzes**

Quizzes are designed to help you better understand what you have learned. The lowest quiz score will NOT count towards the course grade. This allows for sickness, emergencies etc. Therefore please do not ask for remedy if you miss one or more quizzes.

❖ **Grading**

Your final score for the course will be determined as follows:
Small Projects (50%) + Quizzes/participation (15%) + Big Project (35%)