Subject Description Form

Subject Code	COMP4433
Subject Title	Data Mining and Data Warehousing
Credit Value	3
Level	4
Pre-requisite / Co-requisite/ Exclusion	Pre-requisite: COMP2411
Objectives	This subject aims at equipping students with the latest knowledge and skills to: • Create a clean, consistent repository of data within a data warehouse for large corporations; • Utilize various techniques developed for data mining to discover interesting patterns in large databases; • Use existing commercial or public-domain tools to perform data mining tasks to solve real problems in business and commerce; • Expose students to new techniques and ideas that can be used to improve the effectiveness of current data mining tools.
Intended Learning Outcomes	Upon completion of the subject, students will be able to: *Professional/academic knowledge and skills* (a) understand why there is a need for data warehouse in addition to traditional operational database systems; (b) identify components in typical data warehouse architectures; (c) design a data warehouse and understand the process required to construct one; (d) understand why there is a need for data mining and in what ways it is different from traditional statistical techniques; (e) understand the details of different algorithms made available by popular commercial data mining software; (f) solve real data mining problems by using the right tools to find interesting patterns; (g) understand a typical knowledge discovery process such as CRISP-DM; (h) obtain hands-on experience with some popular data mining software. *Attributes for all-roundedness* (i) solve real-world problems in business and commerce using data mining and

	data warehousing tools; (j) learn independently and search for relevant information to write reports to recommend appropriate data warehousing and data mining tools. (k) Solve complex problems individually or in groups and develop group work skills directly and indirectly.
Subject Synopsis/ Indicative Syllabus	Topic 1. Introduction to data warehousing and data mining Introduction to data warehousing and data mining; possible application areas in business and finance; definitions and terminologies; types of data mining problems. 2. Data warehousing Data warehouse and data warehousing; data warehouse and the industry; definitions; operational databases vs. data warehouses. 3. Data warehouse architecture and design Data warehouse architecture and design; two-tier and three-tier architecture; star schema and snowflake schema; data characteristics; static and dynamic data; meta-data; data marts. 4. Data Replication and Online Analytical Processing Data replication, data capturing and indexing, data transformation and cleansing; replicated data and derived data; Online Analytical Processing (OLAP); multidimensional databases; data cube. 5. Data mining and knowledge discovery Data mining and knowledge discovery, the data mining lifecycle; pre-processing; data transformation; types of problems and applications. 6. Association rules Mining of association rules; the Apriori algorithm; binary, quantitative and generalized association rules; interestingness measures. 7. Classification Classification; decision tree based algorithms; Bayesian approach; statistical approaches, nearest neighbor approach; neural network based approach; genetic algorithms based technique; evaluation of classification model. 8. Clustering Clustering; k-means algorithm; hierarchical algorithm; Condorset; neural network and genetic algorithms based approach; evaluation of effectiveness. 9. Sequential data mining Sequential data mining; time dependent data and temporal data; time series analysis; sub-sequence matching; classification and clustering of temporal data; prediction. 10. Other techniques Computation intelligence techniques; fuzzy logic, genetic algorithms and neural networks for data mining. Laboratory Experiment: Topic

1. Knowledge discovery lifecycle using CRISP-DM
2. Discover Association rules and sequential patterns using Clementine
3. Discover Classification rules using Clementine
4. Discover Clusters using Clementine

Case Study:

- Application of data mining techniques to solve real business problems.
- Attributes leading to success and failure of data warehousing projects tutorials when appropriate.

Teaching/Learning Methodology

This subject consists mainly of class lectures and laboratory sessions. For the class lectures, various cases will be presented to help student understand why there is a need for data warehouse to be built and why data mining is important for modern day business intelligence. Students will be given time to participate in discussions when the cases are presented.

All assignments and projects will also be given in the form of different cases collected so as to allow students to learn more about how data warehouse and data mining can be and have been used in real business environment. For the projects and assignments, students are expected to learn independently and think critically with minimize guidance. They are expected to practice their writing kills through project documentations and report writing. As students will work in teams on the project, they are expected to also learn to work with each other collaboratively.

During laboratory sessions, students will be introduced to popular software products from Oracle and IBM that can support the building of data warehouses and the mining of them. Students are expected to solve real data mining problems by using the right tools to find interesting patterns.

Assessment Methods in Alignment with Intended Learning Outcomes

Specific assessment methods/tasks	% weighting	Intended subject learning outcomes to be assessed (Please tick as appropriate)										
		a	b	C	d	e	f	g	h	i	j	k
1. Assignments	55%	✓		✓	✓					✓	✓	
2. Project					✓	✓	✓	✓	✓	✓	✓	✓
3. Examination	45%	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Total	100 %											

The assessment consists of written assignments, a group project and an examination. For the assignments and projects, they are designed to ensure that students are able to achieve the learning outcomes intended for this subject. They are expected to tackle a number of cases drawn from different application areas in business and commerce so that they can understand why there is a need for data warehouse in addition to traditional operational database systems and why data mining is important for modern-day business intelligence. In addition, students will learn through the questions and cases, when a particular data warehouse architecture or when a particular data mining algorithm is useful and should be used. Questions in the assignments are expected to help students learning the details of the data mining algorithm and the use of popular data mining

<table>
<tr><td></td><td colspan="2">software. They are also expected to use such popular tool as Oracle Warehouse Builder to construct data warehouses. For the projects, students are expected to work in groups of three to four to tackle a real case involving the design of a data warehouse or the use of data mining to mine very large data bases. They are expected to learn how real-world problems in business and commerce should be tackled using real-world tools as Oracle's Warehouse Builder or IBM's Clementine data mining system. They are expected to learn independently and search for relevant information to write reports to recommend appropriate data warehousing and data mining tools. Students are expected to practice their writing skills with project document and report writing. They will learn to develop critical thinking and team work skills.</td></tr>
<tr><td rowspan="7">Student Study Effort Expected</td><td>Class contact:</td><td></td></tr>
<tr><td>▪ Lecture/Laboratory</td><td>39 Hrs.</td></tr>
<tr><td>▪ Tutorial</td><td>0 Hrs.</td></tr>
<tr><td>Other student study effort:</td><td></td></tr>
<tr><td>▪ Assignments and case studies</td><td>45 Hrs.</td></tr>
<tr><td>▪ Projects and research</td><td>25 Hrs.</td></tr>
<tr><td>Total student study effort</td><td>109 Hrs.</td></tr>
<tr><td>Reading List and References</td><td colspan="2">Reference Books:

1. Chan, K.C.C., Course Notes and Lab Manuals for COMP417, 2009.
2. Inmon, W.H., Strauss, D., and Neushloss, G., DW 2.0: The Architecture for the Next Generation of Data Warehousing, Morgan Kaufmann, 2008.
3. Golfarelli, M., and Rizzi, S., Data Warehouse Design: Modern Principles and Methodologies, McGraw-Hill, 2009.
4. Rokach, L., and Maimon, O., Data Mining with Decision Trees: Theory and Applications, World Scientific, 2008.
5. Witten, I.H., Frank, E., Hall, M.A., Data Mining, Third Edition: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011.
6. Westphal., C., Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies, CRC Press, 2008.
7. Cox, E., Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration, Morgan Kaufmann, 2005.
8. Liu, B., Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer, Berlin Heidelberg, 2009.
9. Tsiptsis, K., and Chorianopoulos, A., Data Mining Techniques in CRM: Inside Customer Segmentation, Wiley, 2010.
10. Han, J. and Kamber, M., Data Mining: Concepts and Techniques, 2nd Edition, Morgan Kaufmann, 2005.
11. Shapiro, A.F., and Jain, L.C., Intelligent and Other Computational Techniques in Insurance: Theory and Applications, World Scientific, 2003.</td></tr>
</table>