Second Semester Examination
2016/2017 Academic Session

June 2017

CIT553 – Business Intelligence and Data Mining
[Kecerdasan Perniagaan dan Perlombongan Data]

Duration : 2 hours
[Masa : 2 jam]

INSTRUCTIONS TO CANDIDATE:
[ARAHAN KEPADA CALON:]

- Please ensure that this examination paper contains **FOUR** questions in **SEVEN** printed pages before you begin the examination.

  *[Sila pastikan bahawa kertas peperiksaan ini mengandungi **EMPAT** soalan di dalam **TUJUH BELAS** muka surat yang bercetak sebelum anda memulakan peperiksaan ini.]*

- Answer **ALL** questions.

  *[Jawab **SEMUA** soalan.]*

- You may answer the questions either in English or in Bahasa Malaysia.

  [Anda dibenarkan menjawab soalan sama ada dalam bahasa Inggeris atau bahasa Malaysia.]

- In the event of any discrepancies, the English version shall be used.

  [Sekiranya terdapat sebarang percanggahan pada soalan peperiksaan, versi bahasa Inggeris hendaklah diguna pakai.]

1. (a) List and briefly describe the structure, relations of the components of business intelligence.

Senarai dan terangkan secara ringkas struktur, hubungan komponen kecerdasan perniagaan.

(25/100)

(b) Explain the Enterprise Data Warehousing Architecture (major components of a data warehousing process). Name any one type of the Enterprise Data Warehousing Architecture.

Terangkan seni bina gudang data perusahaan (komponen utama proses pergudangan data). Namakan mana-mana dua jenis Seni Bina Gudang Data Perusahaan.

(25/100)

(c) Explain any one type of Data Mart. Can a data mart can replace a data warehouse or complement it? Compare and discuss these options for any airline company.

T*erangkan mana-mana satu jenis mart data. Bolehkah mart data menggantikan gudang data atau melengkapkan ia? Bandingkan dan bincangkan pilihan ini untuk sesebuah syarikat penerbangan.*

(25/100)

(d) Explain why it is important for any airline company to use a real-time data warehouse?

Terangkan mengapa ia adalah penting bagi sesebuah syarikat penerbangan untuk menggunakan gudang data masa sebenar?

(25/100)

2. (a) How can banks use Geographical Information System (GIS)?. Why is the combination of GIS and Global Positioning Systems (GPS) so useful?

Bagaimanakah bank-bank boleh menggunakan Geographical Information System (GIS)? Kenapakah kombinasi GIS dan Global Positioning Systems (GPS) berguna?

(25/100)

(b) Summarize and explain briefly five (5) ways in which BI for BPM differs from traditional BI.

Sila rumuskan dan terangkan secara ringkas lima (5) cara BI untuk BPM berbeza daripada BI tradisional.

(25/100)

(c) What are some of the drawbacks of relying solely on financial metrics for measuring performance?

Apakah kelemahan-kelemahan sekiranya bergantung semata-mata kepada metrik kewangan untuk mengukur prestasi?

(25/100)

(d) What does the term balanced refer to in Balanced Scorecard (BSC)? Explain how does BSC overcome the limitations of systems that are financially focused?

Apakah istilah seimbang merujuk kepada Balanced Scorecard *(BSC)? Terangkan bagaimanakah BSC mengatasi batasan sistem yang memberi tumpuan dari segi kewangan?*

(25/100)

3. (a) Discuss (briefly) whether or not each of the following activities is a data mining task.

Bincangkan (dengan ringkas) sama ada atau tidak setiap satu daripada aktiviti berikut adalah perlombongan data.

(i) Computing the total sales of a company.

Komputasi jumlah penjualan sesebuah syarikat.

(ii) Monitoring seismic waves for earthquake activities.

Pemantauan ombak seismic bagi aktiviti gemba bumi.

(iii) Predicting the future security attack based on malwares using malware patterns.

Meramalkan serangan malware berdasarkan corak malware.

(iv) Monitoring the elderly daily movement for abnormalities.

Pemantauan aktiviti harian warga tua untuk keabnormalan.

(v) Sorting a student database based on student identification numbers.

Pengisihan pangkalan data pelajar berdasarkan nombor pengenalan pelajar.

(15/100)

(b) Suppose you are employed as a data mining consultant for Yahoo. Describe how data mining can help the company by giving specific examples of how techniques such as clustering, classification, association rule mining and anomaly detection can be applied?

Katakan anda bekerja sebagai perunding perlombongan data untuk Yahoo. Huraikan bagaimana perlombongan data boleh membantu syarikat dengan memberi contoh-contoh spesifik bagaimana teknik-teknik seperti kelompok, klasifikasi, perlombongan persatuan peraturan dan anomali pengesanan?

(20/100)

(c) (i) Assume you own an garment store which sells clothes over the Internet. How can your business benefit from data mining?

Andaikan anda memiliki sebuah kedai pakaian yang menjual pakaian melalui Internet. Bagaimana perlombongan data dapat memberi manfaat pada perniagaan anda?

(15/100)

(ii) Give a business intelligence application that benefits from data mining solutions instead of OLAP queries. Be specific with the data mining method you recommend and justify your answer.

Beri aplikasi perisikan perniagaan yang memberi manfaat dari penyelesaian perlombongan data dan bukannya pertanyaan OLAP. Beri kaedah perlombongan data secara spesifik dan beri justifikasi untuk jawapan anda.

(10/100)

(iii) Consider a document-term matrix, where tfij is the frequency of the ith word (term) in the jth document and m is the number of documents. Consider the variable transformation that is defined by

Pertimbangkan sebuah matriks dokumen-istilah, di mana tfij adalah kekerapan perkataan ith (istilah) di dalam dokumen jth dan m adalah nombor dokumen. Pertimbangkan penjelmaan pemboleh ubah yang ditakrifkan oleh

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i}$$

where *dfi* is the number of documents in which the ith term appears and is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation.

di mana dfi adalah nombor dokumen di mana istilah ith muncul dan dikenali sebagai dokumen kekerapan penggunaan istilah. Transformasi ini dikenali sebagai transformasi dokumen kekerapan kekerapan penggunaan istilah songsang.

(A) What is the effect of this transformation if a term occurs in one document? In every document?

Apakah kesan daripada transformasi ini jika satu istilah berlaku dalam satu dokumen? Dalam setiap dokumen yang ada?

(10/100)

(B) What might be the purpose of this transformation?

Apakah tujuan transformasi ini?

(10/100)

(d) (i) What is overfitting? Briefly describe **one** (**1**) method to prevent overfitting in classification trees.

*Apa itu overfitting? Terangkan secara ringkas **satu** (**1**) kaedah untuk mengelakkan overfitting dalam klasifikasi pepohon..*

(8/100)

(ii) List **four** (**4**) goals of dimensionality reduction techniques, such as PCA - for what are they used in practice?

*Senaraikan **empat** (**4**) matlamat teknik pengurangan kematraan, seperti PCA - untuk apa yang mereka digunakan dalam amalan?*

(12/100)

4. (a) (i) Distinguish the difference between Hierarchical and Partitional clustering algorithm. List an example for each of the type

Nyatakan perbezaan algoritma kelompok Hierarchical Partitional. Senaraikan satu contoh bagi setiap algorithm ini.

(10/100)

(ii) Compare the pros and cons of Support Vector Machine and neural network classification methods.

Bandingkan kebaikan dan keburukan pokok keputusan dan kaedah klasifikasi rangkaian neural.

(10/100)

(b) (i) Suppose the fraction of undergraduate students who play football is 15% and the fraction of graduate students who play is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who play football is a graduate student?

Katakan sebahagian daripada pra-siswazah yang bermain bola sepak adalah 15% dan pecahan pelajar siswazah yang bermain bola sepak adalah 23%. Jika satu perlima daripada pelajar kolej adalah pelajar Siswazah dan selebihnya adalah Pra- Siswazah, apakah kebarangkalian bahawa seseorang pelajar yang bermain bola sepak adalah pelajar siswazah?

(5/100)

(ii) Given the information in part 4(b)(i), is a randomly chosen college student more likely to be a graduate or undergraduate student?

Berdasarkan maklumat yang diberikan dalam bahagian 4(b)(i), adakah seorang pelajar kolej yang dipilih secara rambang lebih cenderung untuk menjadi seorang pelajar siswazah atau prasiswazah?

(5/100)

(iii) Repeat part 4(b)(i) assuming that the student is a football player.

Ulang bahagian 4(b)(i) dengan anggapan pelajar itu adalah pemain bola sepak.

(5/100)

(iv) Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who play football.

Katakan 30% pelajar siswazah tinggal di asrama, tetapi hanya 10% pelajar yang tinggal di sebuah bilik asrama. Jika seorang pelajar yang bermain bola sepak dan tinggal di asrama, adakah dia lebih cenderung untuk menjadi seorang pelajar siswazah atau prasiswazah? Anda boleh menganggap secara bebas bahawa pelajar yang tinggal di asrama dan orang-orang yang bermain bola sepak.

(5/100)

(c) An Internet marketer is interested in segmenting Internet with a clustering tool using the input attributes – top ten search key words used, top 10 URLs, recent 10 online purchases (vendor, product, quantity, amount), Internet usage level, heaviest access hour, and heaviest access day of a week. Answer the following questions:

Pemasar Internet berminat untuk mengsegmen Internet dengan alat pengelompokan menggunakan input atribut - sepuluh perkataan utama carian yang digunakan, 10 URL teratas, 10 pembelian dalam talian (vendor, produk, kuantiti, jumlah) yang terbaru , tahap penggunaan Internet, jam akses yang terbanyak, dan hari akses seminggu yang terbanyak. Jawab soalan-soalan berikut:

(i) Can we find users with different income level? Why or why not.

Bolehkah kita mencari pengguna dengan tahap pendapatan yang berbeza? Mengapa atau mengapa tidak.

(5/100)

(ii) Can we expect to find clusters differentiated based on Internet usage level? Why or why not.

Bolehkah kita harapkan untuk mencari kelompok yang berbeza berdasarkan tahap penggunaan Internet? Mengapa atau mengapa tidak.

(5/100)

(d) (i) Compare the major distinction of Unsupervised and Semi-supervised approaches for anomaly detection.

Bandingkan perbezaan utama antara pendekatan Unsupervised dan pendekatan Semi-supervised untuk teknik pengesanan anomali.

(10/100)

(ii) As a data mining supervisor in a Credit Card Company, devise a plan in detecting fraud credit card attack. Your plan must be according to KDD approach which begins with preparation of test data, pre-processing, detection mechanism and ends with the evaluation output. To show your work, an assumption of 1000 instances which includes 5% of fraud instances can be used.

Sebagai seorang penyelia perlombongan data dalam sebuah syarikat kad kredit, rangkakan satu pelan untuk mengesan serangan penipuan kad kredit. Rancangan anda mestilah mengikut pendekatan KDD yang bermula dengan penyediaan data ujian pra pemproses, mekanisme pengesanan dan berakhir dengan output penilaian. Untuk menunjukkan kerja anda, andaikan data anda mempunyai 1000 contoh dan 5% contoh data penipuan .

(40/100)