Knowledge and Data System
Comprehensive Exam Syllabus
Effective: January 2018

NOTE: This PhD comprehensive exam will have questions on

A. Database Systems
B. Information Retrieval
C. Data Mining

It will be sufficient for students to select questions from any 2 of the three topics, in order to score up to a maximum of 100 points.
A. Database Systems

Topics
1. **DBS and DBMS concepts**: technology, capabilities, design DBMS, DBS, data model, schema vs. instances. ANSI/SPARC three level schema architecture, data independence, logical data structure vs. physical storage structures.
2. **Database administration functions**: DBA responsibilities, DBMS performance measurement and evaluation.
3. **Data Models and DB Design**: Data modeling, structures (abstraction, sets, representation), constraints (domain, referential, functional dependencies, assertions, triggers), operations, primary data models (E-R, network, relational, semantic, object-oriented), design methodologies.
4. **Relational DB theory**: Normalization and normal forms (1, 2, 3 BCNF, 4) lossless join and dependency preserving decomposition.
5. **DBMS language and interfaces**: Data definition language, data manipulation language, query languages, view definition languages, storage definition languages, SQL, relational algebra, datalog, QBE, high-level and graphical user interfaces and interface building tools.
6. **Query Processing**: Logical and physical query plans, join techniques (NLF, SMJ, HF), cost-based query evaluation, concurrent queries, ACID properties, and Transactions.
7. **Database Normalization**: Informal design guidelines for Relation Schemas, functional dependencies, Normal Forms based on Primary Keys, Second and Third Normal Forms, and Boyce-Codd Normal Form.

References
B. Information Retrieval

Topics
1. *IR Models:* Boolean Retrieval, Fuzzy Set based, Vector Space and hybrid models.
2. *Evaluation of IR Systems:* Recall, Fall out and Precision, Performance Averaging, RB-Precision, Normalized Recall.

References
1. G. Salton, Automatic Text Processing, Addison-Wesley, 1989. (Ch. 8, sections 8.1-8.4, Ch. 9, Ch. 10).
2. C. D. Manning, P. Raghavan and M. Schuetz, Introduction to Information Retrieval, Cambridge University Press, 2008 (Ch. 1, Ch. 2.1, 2.2, Ch. 6.2, Ch. 6.3, Ch. 8.1 - 8.4, Ch.9, Ch. 9.1.1, Ch. 11.1 - 11.3, (exclude 11.3.1), 11.3.2).
C. Data Mining

Topics
1: Top 10 Algorithms in Data Mining
2: Decision Tree Construction
   • The Concept Learning System (CLS)
   • ID3/C4.5 and C4.5 software
   • CART
   • Backtracking vs greedy algorithms
   • Advanced topics and remaining issues:
     o c4.5rules: decompose a decision tree into rules
     o Cubist
3: Association Analysis
   • A mathematical model for association analysis
   • Large itemsets and association rules
   • Apriori: constructs large itemsets with minisup by iterations
   • Interestingness of Discovered Association Rules
   • Association analysis vs. classification
   • Machine Learning Software in Java at the University of Waikato
   • Additional topics:
     o Quantitative Association Rules
     o Multiple-Level Association Rules
   • Association Analysis with One Scan of Databases
4: Clustering
   • Clustering: unsupervised learning
   • k-means: iterative distance-based clustering
   • Incremental clustering/classification: pros and cons
   • Steps in COBWEB to construct a clustering tree
   • DBSCAN: Density-Based Clustering
   • How to combine clustering and classification?
   • How to measure the quality of clustering?
   • Outlier analysis
5: Rule Induction, kNN and GA
   • Classification rules
   • 1R ("1-rule")
   • c4.5rules vs c4.5
   • Rule Induction by Covering
   • PRISM: Constructing correct and "perfect" rules
   • Rule induction algorithms in Weka
   • Divide-and-Conquer vs Separate-and-Conquer
   • Lazy vs eager learning
   • The k-nearest neighbor algorithm
   • Genetic algorithms (GA)
- Feature Selection

6: Bayesian Methods
- Conditional probability
- Bayes theorem
- Maximum A Posteriori (MAP)
- Naive Bayes Classifier
- Belief networks
- Network topology
- The Naive Bayes algorithm in Weka
- Online streaming feature selection: features to arrive one by one

7: Dealing with Noise and Real-Valued Attributes (3 Lectures)
- Artificial vs. real-world databases
- The Monk's Problems: An example
- Sources of Noise
  - Erroneous values
  - Missing values (?)
  - Misclassifications
  - Contradictory data
  - Redundant data
  - Don't Care (#) values
  - Incomplete attributes and uneven data distribution
- Noise Handling
  - Preprocessing
  - Pre-pruning
  - Post-pruning
  - Dealing with unusual examples when deduction of induction results
- Cross validation
- Dealing with contradictions and redundancy
- Expansion of Don't Care values
- Handling of ? values
- Stopping criteria to avoid overfitting
- Overfitting vs underfitting
- Occam's Razor
- Truncation of rules - TRUNC
- "No match" and "multiple match" when deduction of induction results
- Measure of fit
- Estimate of probability
- Dealing with real-valued attributes: Discretization
- Random noise vs systematic noise
- Impact of noise handling

8: Data Mining from Very Large Databases
A. Why large databases?
B. Data partitioning
C. Sampling techniques
D. Subspacing
E. Windowing in C4.5  
F. Integrative windowing  
G. Bagging, boosting, and their differences  
H. Boosting in C5.0  
I. Random forest  
J. Incremental batch learning  
K. Aggregation of rules from different data sources

References: