Object-Oriented & Component-based Software Development

Programming Style, classes, methods, UML

Frameworks 4. Multithreading, I/O, N/W

Design Patterns 6. Typing & Memory management

Software Engineering I & II

Component-based software

Java beans, EJB components

Web services, components

Fundamentals of Computing and Computer Systems

Digital Logic, Arithmetic

Sets, Theory of Computation

Data Structures

Algorithms

Computer Architecture - instruction sets, CPU, RISC, CISC

Supervisor

OS - threads, processes

OS - concurrency, memory management

OS - I/O, file systems

ADVANCES IN DATA MANAGEMENT

1. Relational model, SQL

- databases: collection of data; shared
  - DBMS: set of programs for managing databases
  - database system: database + DBMS

- 3 schema architecture: physical design + conceptual design + views

- physical & logical independence

- DDL + DML

- objectives: controlling redundancy
  - data structuring & efficient access
  - simultaneous access
  - indexes
  - security (integrity, backup)

- relational model: table, tuple, column, domain
  - key: rows
  - relational algebra
DATA INTEGRATION & WAREHOUSING

1. Introduction & basics, web data models

- Why? apps need to work with data from different sources

- Difference: data models, schema, entities, duplication, conflicts

- Proven & consolidation in order to provide:
  - correctness & completeness
  - single data model & schema
  - 1 entity rep
  - no conflicts

- Heterogeneity
  - Tech: means to access data
  - Syn: encoding

- New SQL
  - data fits into MS
  - logging or persistent DB
6. Data cleaning, quality & fusion

- Providers have different levels of knowledge, view, intentions.
- Info is wrong, biased, outdated, incomplete, inconsistent.
- Depends on consumer on how to use it.
- Data Provenance
  - Info on how data was produced.
  - Simple vs. full provenance.
  - Who processes.
  - Publish provenance on web.
  - Vocab.
  - Time stamp.
  - Metadata.

- Data quality
  - Fitness for use.
  - Enterprise vs. web content.
  - Control / no control.
  - Dimensions.

- Content tech.
  - Intellectual formatting.
DATA MINING

3. Data Mining Process, Preprocessing

- Extraction of useful patterns from data source, which are valid, novel etc

- Basic tasks:
  - Classification, clustering, association rule mining, regression

- Processes:
  - Understand application domain
  - Select sources, collect data
  - Preprocess data mining
  - Post-processing incorporates them

- Applications:
  - Marketing
  - Engineering
  - Tedd
  - Web

Tests

- Two-tailed
  - Upper-tailed
  - Lower-tailed

- F-score

- ANOVA: Analysis of Variance
  - Two or more sample
    - Between / within group variance
      - Independence of observations
      - Normality
- distance of 2 clusters
  - single, complete, avg, centroid

- distance function
  - euclidean distance
  - manhattan distance
  - chess board

- dissimilarity
- cosine

- data standardization
  - z-scores have a mean of zero
  - interval-scaled
  - ratio-scaled

- trial/ error
- evaluation
  - confusion matrix, precision, recall, F-1
  - entropy, purity
  - intra-cluster cohesion / inter-cluster separation

9. classification
- learn a target function that can predict values
- training / testing
- distance if training test data are same

- decision tree
  - NP hard, best tree
  - convert list of rules
  - recursive partitioning based on information gain
  - overfitting
  - preprocessing / postprocessing

- evaluation of classifiers
  - accuracy
  - holdout set, n-fold cross validation
  - precision, recall, F-1
  - score (prob that x belongs to i-th class)
2. Search Strategies
   - Concepts
     - State space
     - path
     - frontier (readable)
     - solution (good bit)
   - Algorithms
     - Tree search (exp)
     - Graph search (state sequence)
     - Uninformed search
     - informed search
     - strategy for expansion
     - evaluation - complexity, graph complexity, optimality (branching factor, depth)
   - Adversaries
     - BFS, DFS, Depth-limited, Uniform cost, Bidirectional
   - Game Life

   \( A^* \rightarrow g(n) + h(n) \)

   - Local search & optimization
     - hill climbing (gradient ascent decent) (local maxima)
     - Simulated annealing
     - expect bad move with prob
     - Tabu algorithm (random, mutate, crossover, iter)

   - Game playing AI
     - Minimax search
     - Alpha-Beta pruning
     - Evaluation function (for board position)
Knowledge representation and Logic

- Knowledge-based agent
- Knowledge base
- Not
- Propositional logic
- First-order logic
- Reasoning via inference

Logic
- Sentence
- Well-formed (meaning)
- Model, true/false
- Entailment
- Inference: e.g. sound, complete

Propositional logic
- Syntax
  - \( \land, \lor, \rightarrow, \leftrightarrow, T, F \)

- Inference
  - Rule: \( \forall x, \exists y (x \rightarrow y) \)
  - \( \forall x \forall y (x \leftrightarrow y) \)
  - Linear with hemi-clauses (almost first order)

- First-order Logic
  - Objects, relations (arity), constants, properties, functions
  - Quantifiers
    - \( \forall x \) (universal, restricted)
    - \( \exists x \) (existential)
  - Equality
  - Cloud world assumption
  - Inference
  - Convert to propositional logic (complete, inefficient)
    - Generalised resolution
    - Frege/Boolean chains
    - Resolution
      - Convert to LNF, search P1 (what to instantiate)
- new node in main memory
- data node
  - block server, block report
  - pipelining
    - block placement, data correctness, failures
- distributed access structures
  - indexing
    - tree, hash
- consistent hashing, distributed hash directories
- Amazon Dynamo, BTREE
  - Bigtable, BTREE + non-index index tables
- data model: Kvs + key (tablets)

5. Map Reduce & Hadoop
- divide & conquer
- scale-out, hide sys details from programmers
  - sync to others
    - comm + shared resource access
  - prg models lower, (mpi, threads)
  - DF: master/slave, prod ones
    - shared worker queues
- map, shuffle & sort, reduce
  - extract, group, aggregate
- MR: combine handlers
  - scheduling
  - data distribution
  - sync
  - error, failure
  - coordination/dealing with failures
- DFS + LFS
Beyond MR:

- BSP
  - processors & communication
  - supersteps - local map + count
  - barrier sync
  - pregel, graph, hama

7. Data Management, Link Analysis, information retrieval in the cloud

- KDBMS v/s MR
  - multipurpose
  - fault tolerant
  - native data

- DB workloads
  - OLTP
    - read-write, small queries, random reads & writes
  - OLAP
    - batches, less concurrency, complex query,
    - large scans for query
    - hadoop: cheaper than parallelized DBs
    - rel alg ops: proj, sel, join, cartesian, union, 
c      diff, group by, rename
      - proj: map over tuples
      - sel: map over tuples
      - group by: reduce
      - join: map-side, reduce-side, in-memory (hash)
      - Hive, Pig
Index construction
- map over all docs, emit term as key
  (doc, tf) as value
- sort/shuffle - group postings
- reduce - gather, sort, emit postings

 Retrieval
- MR optimized for throughput
  latency, mRstart is expensive
  real-time
- partition of terms/docs
- replication, caching
- ranking

Link analysis
- key: node
  value: adjacency list
  map: adjacency list
  sort/shuffle: group distances by reachable nodes
  reduce: set min distance for each reachable node

Semantic Web Technologies
1. Introduction, Layers
   - gobal vs foundations of semantic web
   - today: synthetic web
     - computers do presentation, humans do
     linking & interpretation
     impossible tasks
     - complex queries involving background
       knowledge
       - web service
       - agent
       - location info in data repositories
     - make implicit knowledge explicit
     - add extra info in a std way
       make it w/c readable
     - set is a set of technologies to realize
       web of data
     - novel w/c understandable language
       rules, ontologies
ontology: explicit specifications for a conceptualization

layers: unicode + uri
xml + namespace + xschema
rdfs + rdf
ontology
logic
proof
trust

HTML: is about presentation
XML: more structure (nesting)
content / formatting (data exchange)
tree model
DTD/XMLs
namespaces

XSL

XSLT xpm xsl-p c
not powerful enough for csw

2. RDF metadata

- subject, predicate, object (mathematical logic)
  ↓
  domain, range (unequal relation) (n-ary)

- resource description framework
- s, p, o all are uris
- triples, xml syntax
- blank node
- container elements
  - bag
  - seq
  - alt
  - verification
    - str
    - alt
    - str
6. Typing, Memory management

- For every operation, types of data on which it is applicable
- Weak/strong typing
  - Type-safe
  - Prevent value of one type to be treated as another

- Static/dynamic typing
  - Compile-time (no type for variable) during prog.
  - Run-time

- Activation records: allocated when proc is entered & deallocated when proc is exited
  - (Info needed by single of proc)
deadlock
- conditions
  - mutual exclusion
  - hold & wait
  - no preemption
  - circular wait
- resource allocation graph
- deadlock prevention/avoidance (future process request)
  - banker algorithm (claim matrix, resource matrix)

Memory Management
- prog relative address, actual address
- registers: base & band

- partitioning, placement: best fit, fixed, dynamic, next fit
- virtual memory
  - paging (replacement, locality, thrashing)

Operating System - I/O, File Systems
- devices
  - devices controller
  - I/O
    - prog, interrupt, DMA
  - device drivers
    - buffering: block, stream
  - disk performance parameters
    - seek, rot, access, latency, time
  - disk scheduling
    - FIFO, SSTF, SCAN
- RAID
- disk cache, LRU