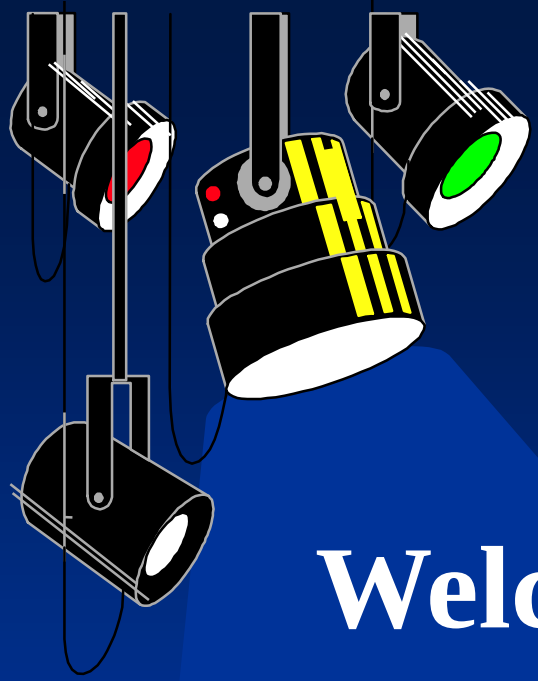

# Welcome to the Course of Web and Document Databases (7102-001)

Name: Yangjun Chen
Birthplace: China

P.h..D: University of Kaiserslautern, Germany, in 1995

Post Doctor: University of Chemnitz, Germany, 1995/7 - 1997/8

Senior engineer: Germany Research Center for Information Technology, 1997/9 - 2000/2

Post Doctor.: University of Alberta, 2000/2 - 2000/6

Assistant Prof.: University of Winnipeg, 2000/7 – 2004/6
Associate Prof.: University of Winnipeg, 2004/7 – 2009/6
Full Prof.: University of Winnipeg, from 2009/7

Professor: Dr. Y. Chen

Office: 3D27

home-page: http://www.acs.uwinnipeg.ca/ychen2

E-mail: y.chen@uwinnipeg.ca
phone: 204-786-9417

Meeting time: 11:30 – 12:45 pm, Monday and Wednesday

Meeting location: 3D03

Office hours: 4:00 pm - 5:00 pm Monday and Wednesday
11:00 – 16:00 Friday (except time for department meeting)

# Course Outline

**Intro. to the design of Web and document databases, analysis, and applications**

***Introduction to relational database theory***

Basic methods for designing relational databases, SQL and JDBC, Hashing, B+-trees, R-trees.

***Web and XML documents***

Web databases, Semistructured-data Model, Programming languages for XML, Web browser and search engines.

***Evaluation of tree pattern queries in document databases***

Tree searching and graph searching, unordered tree matching, ordered tree matching, XB-tree.

***Evaluation of reachability query in graph databases***

Transitive closure, matrix multiplication, Warren algorithm, tree encoding, Permanent linked-list storage of graphs

# Course Outline

**Intro. to the design of Web and document databases, analysis, and applications**

***Data Warehouse***

star schema, snowflake schema, galaxy schema,
OLAP (online analytical processing): drill down, roll up, pivot, slice, dice

***Data mining***

**Mining association rules**
transaction databases, a priori algorithm

**Find most popular packages**
most popular packages, *p*-graphs, *p**-graphs, trie-like graphs, layered representation of a trie-like graph

# Course Outline

***Evaluation of extended reachability queries***

Regular expressions, deterministic finite automata (DFA), non-deterministic finite automata (NDFA), intersection graphs, restricted regular expression, general reachability query evaluation algorithm. (A. O. Mendelzon, and P. T. Wood, Finding Regular Paths in Graph Databases, SIAM J. Comput. Vol. 24, No. 6, pp. 1235-1258, Dec. 1995)

***Reference books:***

- ***Database Systems (the complete book)*, 2nd Ed. by Carcia-Molina, Ullman, Widom, Pearson Prentice Hall, 2009.**
- **Fundamentals of Database Systems, 5th, 6th, 7th edition, Elmasri, Navathe, Addison-Wesley,**
- ***Introduction to Algorithms*, 2nd Ed. by Cormen, Leiserson, Rivest, & Stein (CLRS), McGraw Hill, 2002.**
- **Lecture slides online**
- **For evaluation of tree pattern queries:**

  - N. Bruno, N. Koudas, and D. Srivastava, Holistic Twig Joins: Optimal XML Pattern Matching, in *Proc. SIGMOD Int. Conf. on Management of Data*, Madison, Wisconsin, June 2002, pp. 310-321.

  - Y. Chen and L. Zou, Unordered tree matching: the evaluation of tree pattern queries, Int. J. Information Technology, Communications and Convergence, 2011.

***Reference books:***

- **For evaluation of reachability queries**

- Warren, “A Modification of Warshall’s Algorithm for the Transitive Closure of Binary Relations,” *Commun. ACM* 18, 4 (April 1975), 218 - 220.
- H. Wang, H. He, J. Yang, P.S. Yu, and J. X. Yu, Dual Labeling: Answering Graph Reachability Queries in Constant time, in *Proc. of Int. Conf. on Data Engineering*, Atlanta, USA, April 6-8, 2006.
- H. Yildirim, V. Chaoji, and M.J. Zaki, GRAIL: Scalable Reachability Index for Large Graphs, in *Proc. VLDB Endowment*, 3(1), 2010, pp. 276-284.
- Y. Chen and Y.B. Chen, Decomposing DAGs into spanning trees: A new way to compress transitive closures, in Proc. 27$^{th}$ Int. Conf. on Data Engineering (ICDE 2011), IEEE, April 2011, pp. 1007-1018.
- Chen, Yangjun: General Spanning Trees and Reachability Query Evaluation, in Proc. Canadian Conference on Computer Science and Software Engineering ($C^3S^2E$’09), Montreal, Canada, 2009, IEEE, pp. 243 – 252.

***Reference books:***

- **For evaluation of regular expressions**
  - A. O. Mendelzon, and P. T. Wood, Finding Regular Paths in Graph Databases, SIAM J. Comput. Vol. 24, No. 6, pp. 1235-1258, Dec. 1995)

# Course Roadmap

- Database basics

  EER-diagrams, SQL and JDBC, Indexes: Hashing, $B^+$-tree, R-tree, kd-trees, Quad-trees, …
- Web and document databases

  Web databases: PHP, Node.js

  Semi-structured data model

  Programming languages for XML

  Searching engines
- Evaluation of tree pattern queries

  - Tree searching and graph searching

  - Unordered and ordered tree matching

  - XB-trees

# Course Roadmap

- Graph databases and reachability query evaluation
  - Transitive closure
  - Matrix multiplication, Warren's algorithm
  - Methods based on tree encoding
  - Methods based on graph deduction and decomposition
  - permanent linked-list storage of graph databases
- Data warehouse
  - star schema, snowflake schema, galaxy schema
  - OLAP (online analytical processing): drill down, roll up, pivot, slice, dice

# Course Roadmap

- Data mining:

  mining association rules

  - transaction databases

  - A priori algorithm

  finding most popular packages

  - Questionnaire

  - *p*-graphs, *p**-graphs

  - trie-like graphs, layered representation of trie-like graphs

# Course Roadmap

- Evaluation of extended reachability queries
  - Regular expressions, DFA (deterministic finite automata), NDFA (non-deterministic finite automata)
  - Intersection graphs and restricted regular expressions
  - General reachability query evaluation algorithm

# Database Basics

- Basic method for designing relational databases
  - Database system architecture, Enhanced entity-relationship diagram
  - Rules for mapping EERD to relational schema
- SQL, database application: JDBC
- Hashing, $B^+$-tree, indexes over multi-dimensional data
  - Hashing and linear hashing
  - B+-tree, R-tree construction and maintenance
  - kd-trees, Quad-trees, Bit-map, inverted files
- ***Goals***
  - Background knowledge on database systems

# XML Document Databases

- Web databases: PHP,  Node.js
- Semistructured-data model
- Programming languages for XML
- Databases over internet
- ***Goals***
  - PHP, Node.js script languages
  - XML markup language
  - Data storage in document databases
  - Manipulation of data in document databases
  - Web browser and search engines

# Evaluation of Tree Pattern Queries

- Tree searching and graph searching
- Tree encoding
- Ordered tree matching
- Unordered tree matching
- XB-trees
- ***Goals***
  - Evaluation of queries against document databases
  - Indexing data to speed up query evaluation in document databases

# Reachability Query Evaluation

- Transitive closures and reachability checking
  - matrix multiplication
  - Warren's algorithm
- Method based on tree encoding
  - Tree encoding
  - Extension of tree encoding to DAGs
- Method based on graph deduction and decomposition
- permanent linked-list storage of graph databases
- ***Goals***
  - Efficient algorithm for checking reachability of nodes
  - Efficient storage of graphs on hard disk

# Data Warehouse

- Special relational schema:
  - *Star schema, snowflake schema, galaxy schema*
- OLAP: online analytical processing
  - *drill down, roll up, pivot, slice, dice*
- ***Goals***
  - Data analysis for supporting decision making
  - Built-in SQL queries for data analysis

# FINDING MOST POPULAR PACKAGES

- Popular package according to a questionnaire
- *p*-graphs, *p**-graphs
- Trie-like graphs, and layered representation of a trie-like graph
- ***Goals***
  - A kind of data mining to find customer patterns to enlarge sales
  - An efficient algorithm to find a most popular package according to a questionnaire

# Evaluation of Extended Reachability Queries

- Regular expressions, DFA, NDFA
- Intersection graphs
- Restricted regular expressions
- General reachability query evaluation algorithm

# Project

- Implementation of an algorithm for constructing an XB-tree
- Implementation of an algorithm for decomposing a DAG into a minimized set of chains
- Implementation of an algorithm for evaluating unordered tree pattern queries
- Implementing an algorithm for evaluating ordered tree pattern queries

**Important dates:**

| | |
|---|---|
| Wednesday, Jan. 06, 2025 | First class |
| Feb. 16 – 22, 2025 | Reading break |
| Mon., March 03, 2025 | Midterm examination |

March 14, 2025

Final date to withdraw without academic penalty from a course that begins in Jan. and ends in April of the 2025 Winter term

| | |
|---|---|
| April 02, 2025 | Last class |
| Final examination | no (replaced by projects) |

**Course Evaluation:**

| | |
|---|---|
| 3 assignments | 24% |
| 1 midterm examination | 26% |
| 1 project (or final exam.) | 50% |

- All assignments are handed in through email on the due date.
- All works must be prepared using a word processor (and placed in a folder).
- Late assignments are accepted (up to 1 day late) and receive a 25% penalty.
- Assignment sent to teaching assistant (Ms. Rasagnya Kondam, asagnya53@gmail.com)

**Academic dishonesty:**

- Academic dishonesty is a very serious offense and will be dealt with in accordance with the University's discipline bylaw. Be sure that you have read and understood Regulations and Policies, #8 in the 2024-2025 UW Calendar.