UNSW

XUEMIN LIN
PROFESSOR
CSE
School of Computer
Science and Engineering

Thesis: Efficiently and Effectively Processing Probabilistic Queries on Uncertain Data

Significance. Driven by many recent applications including social networks, sensor networks, data cleaning and integration, moving objects, image processing, information retrieval, crime control, economic decision making and market surveillance, querying and analyzing uncertain data draws a great deal of research attention from database community. A number of system prototypes for managing uncertain data have been proposed by Stanford University, University of Washington, Oxford University, etc.

Originality. The PhD thesis mainly consists of five original works presented in Chapters 3, 4, 5, 6, 7; it systematically investigates a set of very important and fundamental problems. Each of these 5 chapters in the thesis is based on a publication in top venue (top conferences or top journals). Chapter 3 is based on a paper accepted by **Information Systems** (an A* ranked journal by ERA), Chapter 4 on an **ICDE** (IEEE International Conference in Data Engineering, a world premier DB conference) 2009 paper, Chapter 5 on a **VLDB Journal** (an A* ranked journal by ERA) 2010 paper, Chapter 6 on an **ICDE** 2010 paper, and Chapter 7 based on a paper in **TKDE** (IEEE Transactions on Knowledge and Data Engineering, an A ranked journal by ERA but arguably a top-tier DB journal) 2010.

Innovation. Novel techniques for probabilistic skyline queries and top-k dominating queries presented from Chapter 3 to Chapter 5 are essential tools to multi-criteria optimal decision making. A top-k nearest neighbour (KNN) query (Chapter 6) on multi-valued objects retrieves k objects from a dataset which are the nearest/most similar to a given object. KNN queries are widely utilized in location based services and recommending systems. All these queries are all firstly studied in the research works presented in this thesis. Chapter 7 proposes a novel index structure especially designed to support querying uncertain data and significantly outperforms all existing works. Detailed innovations and contributions of each chapter are listed below.

Chapter 3 studies the problem of probabilistic top-k skyline queries. A model for the top-k skyline operator is proposed combining the feature of top-k objects and that of skyline. Based on this model, an efficient exact algorithm and a randomized algorithm with ε -approximation guarantee are developed for discrete and continuous cases, respectively.

Chapter 4 extends skyline operator to streaming environment and studies the problem of *probabilistic skyline queries over sliding windows*. It firstly characterizes the minimum information needed in continuously computing probabilistic skyline against a sliding window. Then novel, efficient techniques are developed to process a continuous, probabilistic skyline query.

As the top-k dominating query is another important method for multi-criterion decision making, Chapter 5 studies *top-k dominating queries on uncertain data*. The problem is formally defined in a probability threshold fashion. Then, a threshold-based algorithm is developed to compute the exact solution. To overcome some inherent computational deficiency in an exact computation, an efficient randomized algorithm with an accuracy guarantee is developed

Chapter 6 explores the problem of *quantile-based KNN over multi-valued objects*. Two different quantile distances are proposed. While the first distance can be computed in polynomial time, the second problem is NP-hard. A set of efficient, novel algorithms have been proposed to give an exact solution for the first problem and an approximate solution for the second problem with the approximation ratio 2.

To overcome some deficiencies of existing uncertain index structures, Chapter 7 proposes *UI-tree* which can efficiently support various queries including range queries, similarity joins and their size estimation, as well as top-k range query, over multi-dimensional uncertain objects against continuous or discrete cases.

In the end, I warrant that all three examiners' reports are included in this application.

Professor
Head of Database Group
School of Computer Science and Engineering
The University of New South Wales
Email: lxue@cse.unsw.edu.au
URL: www.cse.unsw.edu.au/~lxue
Phone: 61-2-93856493