

PhD Thesis

Efficiently and Effectively Processing Probabilistic Queries on Uncertain Data

Candidate

Wenjie Zhang

Detailed Report:

Uncertainty is inherent in many real applications. Uncertain data analysis and query processing has become a critical issue and has attracted a great deal of attention in database research community recently. The thesis, therefore, targets an important and challenging topic - uncertain data management. It is a high quality and well-written PhD thesis. Five important and related aspects of uncertain data management are investigated in this thesis. They are probabilistic top- k skyline query (Chapter 3), probabilistic skyline over sliding windows (Chapter 4), probabilistic threshold based top- k dominating query (Chapter 5), KNN search over multi-valued objects (Chapter 6), and effective indexing structure (Chapter 7).

All the research problems identified in the thesis are all well motivated. The literature review is comprehensive, which covers the models for uncertain data, existing techniques on various probabilistic query types, representative DBMSs specially designed for supporting uncertainty management, and other research topics on uncertain data management. The research contributions reported in this thesis (Chapter 3 to Chapter 7) are significant and original. Some of them are the first attempt of the problem study. The work reported in each chapter follows the same methodology that is common in research in the area of Computer Science, including discussion about motivations, problem analysis and definition, framework and algorithm design, complexity analysis and experimental evaluation.

The results and findings of the thesis are clearly presented and technically sound. From the thesis, it is clear that the candidate has a solid background in both mathematics and computer science. She has demonstrated the strong capability of conducting innovative research. It is very impressive that each of 5 chapters has all been published in top venues in databases, including VLDB Journal, TKDE, Information Systems, ICDE'09 and ICDE'10. This is a remarkable achievement! Congratulations!

The thesis made the following contributions:

- (1) Probabilistic top- k skyline query: The problem of top- k skyline on uncertain data is formally defined and studied for both continuous and discrete uncertain cases. Efficient and effective exact and random algorithms are proposed with extensive experimental studies.

- (2) Probabilistic skyline operator over sliding windows: The problem of probabilistic skyline queries in the streaming environment is studied. A candidate set with minimum size is characterized and efficient techniques based on R-tree structures are developed to answer the skyline queries continuously.
- (3) Probabilistic threshold based top- k dominating query: The problem of efficiently computing top- k dominating queries on uncertain data is investigated. The problem is formally defined in a probability threshold fashion. Both exact and random algorithms are proposed to tackle the problem.
- (4) KNN search over multi-valued objects: The problem of KNN search over multi-valued objects is investigated. The quantile paradigm is used to retrieve KNN sensitive to the relative distribution among multi-valued objects. Two different problem definitions are introduced and corresponding techniques are developed.
- (5) Effective indexing structure - To overcome some deficiencies of existing uncertain index structures, a novel R-Tree based inverted index structure, named UI-tree is proposed which can efficiently support various queries including range queries, similarity joins, and top- k range query, over multi-dimensional uncertain objects against continuous or discrete cases.

The following typos can be corrected easily.

P3, -L5: so that are often used -> so that they are often used

P4, L4: for sake of -> for the sake of

P5, L9: wether -> whether

P14, L2: summary -> summarise

P18, L8: Sarma *et al* integrates -> Sarma *et al* integrate

P30, L2: three representatives or six representatives

P250, -L4: Two different problem definition -> Two different problem definitions



UNSW
THE UNIVERSITY OF NEW SOUTH WALES

GRADUATE RESEARCH SCHOOL

**EXAMINER'S REPORT FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY**

STRICTLY CONFIDENTIAL

Name of Candidate: W Zhang

School: School of Computer Science and
Engineering

Title of Thesis: Efficiently and effectively processing
probabilistic queries on uncertain data.

Date Report Due: 03/02/2010

A After examination of the thesis (and supporting papers) I recommend that:

Please circle the number next to the appropriate recommendation

1. The thesis merits the award of the degree.

This recommendation is appropriate if the thesis contains no faults that are apparent to the examiner. It is also appropriate where errors and omissions of an editorial nature are minor and, if left uncorrected, will not alter the conclusion that "the thesis merits the award of the degree".

or

2. The thesis merits the award of the degree subject to minor corrections as listed being made to the satisfaction of the Head of School.

The errors and omissions, which extend beyond those of an editorial nature, must be corrected if the thesis is to merit the award of the degree. The corrections are minor in that they do not change the structure or the conclusions of the relevant chapters of the thesis.

or

3. The thesis requires further work on matters detailed in my report. Should performance in this further work be to the satisfaction of the Faculty Higher Degree Committee, the thesis would merit the award of the degree.

The further work required should be sufficiently straightforward such that the examiner is happy to delegate approval of the revised thesis to the Higher Degree Committee. Examples of further work in this category could include: discussion and consideration of published work that is relevant to the conclusions of the thesis; consideration of alternative hypotheses that should reasonably be suggested by the candidate; presentation of additional experimental data that could be expected to be in the possession of the candidate; clearer specification of how the presented results/conclusions were arrived at.

or

4. The thesis does not merit the award of the degree in its present form and further work as described in my report is required. The revised thesis should be subject to re-examination.

Please indicate whether you are willing to review the resubmitted thesis Yes No

The further work involves a major revision of the thesis on the same topic. The examiner is assumed to be satisfied with the candidate's capability and demonstrated competence for this further work. The comments and suggestions in the detailed report should be clear and helpful to the candidate. As the thesis is to be revised along the lines suggested by the examiners, it would normally be re-examined by the same examiners. Examples of further work in this category could include: further analyses or experiments where the scientific method as presented in the thesis has significant flaws; performance of additional experiments that are deemed vital to the conclusions drawn in the thesis.

or

5. The thesis does not merit the award of the degree and does not demonstrate sufficient ability by the candidate for a resubmitted thesis to achieve this merit.

The examiner should provide the basis of this recommendation in the detailed report.

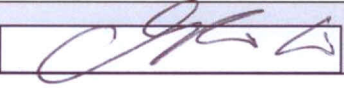
B I agree to my name being released to the candidate

Yes No

C An evaluation of the merits of the thesis is given overleaf
(Attach additional pages if required)

The examiner is requested to state concisely the grounds on which the recommendation is based, indicating where appropriate the strengths and weaknesses of the thesis. To assist the candidate with future work, this report may be released after a decision has been taken to award or not to award the degree. Any sections which are not to be released should be indicated by a vertical line in the margin.

Examiner: C Liu

Signature: 

Date: 15/10/2010

The work presented by the thesis explored many fundamental problems in managing uncertain data. It is based on a few papers by the author in top quality conferences including ICDE, VLDB, SIGMOD, etc. The scope of the work covers topics such as top-k skyline/dominating queries over uncertain data, KNN queries over uncertain data, uncertain data indexing, etc. The thesis not only serves as a summary of the author's work during her Ph.D. study, but because of its novelty and depth, also serves a good tutorial to this emerging field. It certainly merits the award of the degree.

Among the work described in the thesis, probabilistic top-k skyline queries are of fundamental importance to managing uncertain data. The author gave a very nice and comprehensive treatment to this problem. In particular, two algorithms have been presented: one exact algorithm that gives precise ranking to skyline objects, and one randomized algorithm that comes with accuracy guarantees. The randomized algorithm suits the problem particularly well, as it is able to provide an efficient solution without sacrificing accuracy. The extension of this work to uncertain stream data, top-k dominating queries, and KNN queries is natural and interesting.

Besides working on probabilistic top-k queries and their extensions, the author also worked on indexing of uncertain data. This is a field less explored, but of equally great importance. The author developed an indexing structure called UI-Tree, which summarizes uncertain data of arbitrary PDFs. Based on the UI-Tree, the author introduced methods to handle a variety of important queries, including the range query, which is probably the most used type of queries over uncertain data. The author also demonstrated the extensibility of the indexing framework for a large variety of uncertain data. The analysis of the UI-Tree in terms of its filtering capacity is rigorous, and the experiment results are convincing.

Overall, this is a high quality thesis representing high quality work.

The author may want to go over the writing as it has some typos and grammatical errors.



UNSW
THE UNIVERSITY OF NEW SOUTH WALES

GRADUATE RESEARCH SCHOOL

EXAMINER'S REPORT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

STRICTLY CONFIDENTIAL

Name of Candidate: W Zhang

School: School of Computer Science and
Engineering

Title of Thesis: Efficiently and effectively processing
probabilistic queries on uncertain data.

Date Report Due: 03/02/2010

A After examination of the thesis (and supporting papers) I recommend that:

Please circle the number next to the appropriate recommendation

1. The thesis merits the award of the degree.

This recommendation is appropriate if the thesis contains no faults that are apparent to the examiner. It is also appropriate where errors and omissions of an editorial nature are minor and, if left uncorrected, will not alter the conclusion that "the thesis merits the award of the degree".

or

2. The thesis merits the award of the degree subject to minor corrections as listed being made to the satisfaction of the Head of School.

The errors and omissions, which extend beyond those of an editorial nature, must be corrected if the thesis is to merit the award of the degree. The corrections are minor in that they do not change the structure or the conclusions of the relevant chapters of the thesis.

or

3. The thesis requires further work on matters detailed in my report. Should performance in this further work be to the satisfaction of the Faculty Higher Degree Committee, the thesis would merit the award of the degree.

The further work required should be sufficiently straightforward such that the examiner is happy to delegate approval of the revised thesis to the Higher Degree Committee. Examples of further work in this category could include: discussion and consideration of published work that is relevant to the conclusions of the thesis; consideration of alternative hypotheses that should reasonably be suggested by the candidate; presentation of additional experimental data that could be expected to be in the possession of the candidate; clearer specification of how the presented results/conclusions were arrived at.

or

4. The thesis does not merit the award of the degree in its present form and further work as described in my report is required. The revised thesis should be subject to re-examination.

Please indicate whether you are willing to review the resubmitted thesis Yes No

The further work involves a major revision of the thesis on the same topic. The examiner is assumed to be satisfied with the candidate's capability and demonstrated competence for this further work. The comments and suggestions in the detailed report should be clear and helpful to the candidate. As the thesis is to be revised along the lines suggested by the examiners, it would normally be re-examined by the same examiners. Examples of further work in this category could include: further analyses or experiments where the scientific method as presented in the thesis has significant flaws; performance of additional experiments that are deemed vital to the conclusions drawn in the thesis.

or

5. The thesis does not merit the award of the degree and does not demonstrate sufficient ability by the candidate for a resubmitted thesis to achieve this merit.

The examiner should provide the basis of this recommendation in the detailed report.

B I agree to my name being released to the candidate

Yes No

C An evaluation of the merits of the thesis is given overleaf
(Attach additional pages if required)

The examiner is requested to state concisely the grounds on which the recommendation is based, indicating where appropriate the strengths and weaknesses of the thesis. To assist the candidate with future work, this report may be released after a decision has been taken to award or not to award the degree. Any sections which are not to be released should be indicated by a vertical line in the margin.

Examiner's Report

The thesis is one of the best among those I have reviewed. It exhibits important problems and novel solutions for many practical applications. In real world, many applications deal with uncertain data. Uncertain data management has become an active and important research topic in recent years.

In this thesis, five specific problems related with the uncertain data management have been nicely addressed, including probabilistic top-k skyline queries, probabilistic top-k skyline queries over sliding windows, probabilistic top-k dominating queries, quantile-based top-k queries, and a new indexing method called UI-tree. All these problems have been properly motivated, defined, solved and proved with theory guarantees and experimental results.

This thesis's high quality is suggested from several aspects:

First, this thesis addresses important and practical problems. The applications which can benefit from this research are rich.

Second, the literature review has been extensively conducted. The previous research is concisely illustrated in Chapter 2.

Third, for each problem, the proposed idea is sound. Its technical depth is high, with theory proof.

Fourth, extensive experiments are conducted to compare with existing works, showing the superiority of the new proposals in proper settings.

Fifth, a large portion of this thesis has been published in several Tier-1 conferences and journals including ICDE, VLDB journal, etc, which also confirms the quality of this research.

Finally, the presentation of the thesis is excellent. It's easy to follow and understand. It is a pleasure to read.

In summary, this thesis has made significant and original contribution to the research field of database management and query processing. There is no clear errors/mistakes from my point of view.



UNSW
THE UNIVERSITY OF NEW SOUTH WALES

GRADUATE RESEARCH SCHOOL

EXAMINER'S REPORT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

STRICTLY CONFIDENTIAL

Name of Candidate: W Zhang

School: School of Computer Science and
Engineering

Title of Thesis: Efficiently and effectively processing
probabilistic queries on uncertain data.

Date Report Due: 03/02/2010

A After examination of the thesis (and supporting papers) I recommend that:
Please circle the number next to the appropriate recommendation

1. The thesis merits the award of the degree.

This recommendation is appropriate if the thesis contains no faults that are apparent to the examiner. It is also appropriate where errors and omissions of an editorial nature are minor and, if left uncorrected, will not alter the conclusion that "the thesis merits the award of the degree".

or

2. The thesis merits the award of the degree subject to minor corrections as listed being made to the satisfaction of the Head of School.

The errors and omissions, which extend beyond those of an editorial nature, must be corrected if the thesis is to merit the award of the degree. The corrections are minor in that they do not change the structure or the conclusions of the relevant chapters of the thesis.

or

3. The thesis requires further work on matters detailed in my report. Should performance in this further work be to the satisfaction of the Faculty Higher Degree Committee, the thesis would merit the award of the degree.

The further work required should be sufficiently straightforward such that the examiner is happy to delegate approval of the revised thesis to the Higher Degree Committee. Examples of further work in this category could include: discussion and consideration of published work that is relevant to the conclusions of the thesis; consideration of alternative hypotheses that should reasonably be suggested by the candidate; presentation of additional experimental data that could be expected to be in the possession of the candidate; clearer specification of how the presented results/conclusions were arrived at.

or

4. The thesis does not merit the award of the degree in its present form and further work as described in my report is required. The revised thesis should be subject to re-examination.

Please indicate whether you are willing to review the resubmitted thesis Yes No

The further work involves a major revision of the thesis on the same topic. The examiner is assumed to be satisfied with the candidate's capability and demonstrated competence for this further work. The comments and suggestions in the detailed report should be clear and helpful to the candidate. As the thesis is to be revised along the lines suggested by the examiners, it would normally be re-examined by the same examiners. Examples of further work in this category could include: further analyses or experiments where the scientific method as presented in the thesis has significant flaws; performance of additional experiments that are deemed vital to the conclusions drawn in the thesis.

or

5. The thesis does not merit the award of the degree and does not demonstrate sufficient ability by the candidate for a resubmitted thesis to achieve this merit.

The examiner should provide the basis of this recommendation in the detailed report.

B I agree to my name being released to the candidate Yes No

C An evaluation of the merits of the thesis is given overleaf
(Attach additional pages if required)

The examiner is requested to state concisely the grounds on which the recommendation is based, indicating where appropriate the strengths and weaknesses of the thesis. To assist the candidate with future work, this report may be released after a decision has been taken to award or not to award the degree. Any sections which are not to be released should be indicated by a vertical line in the margin.