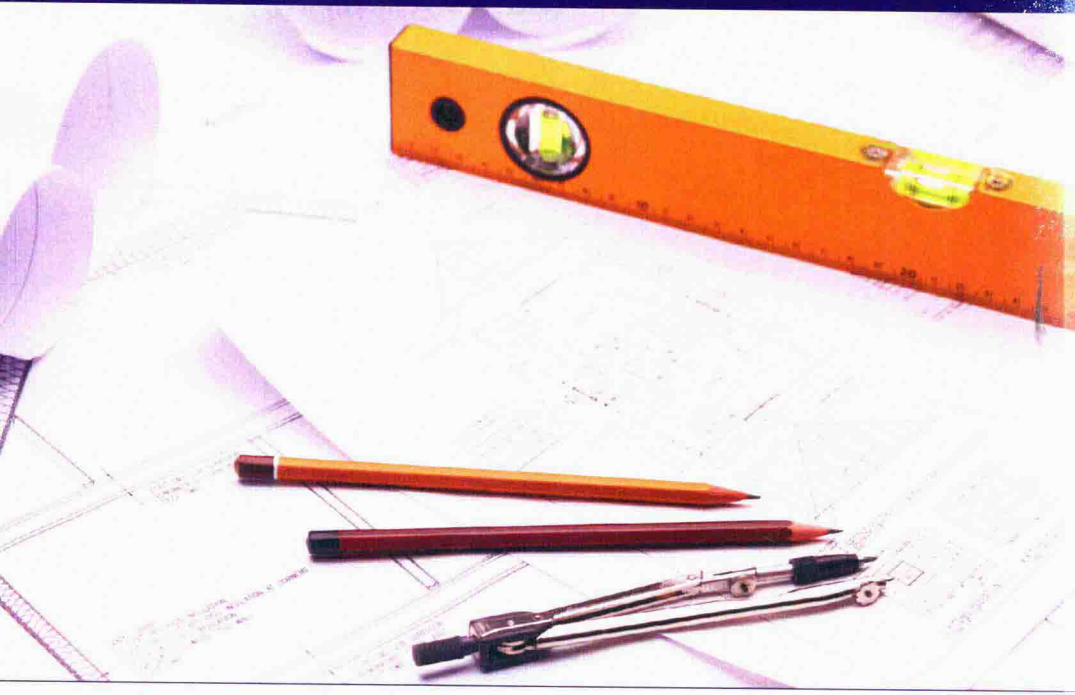

Jerry Chun-Wei Lin

Tree-based Algorithms for Incremental, Utility, and Fuzzy Data Mining

LAP LAMBERT Academic Publishing

Jerry Chun-Wei Lin

Tree-based Algorithms for Incremental, Utility, and Fuzzy Data Mining

LAP LAMBERT Academic Publishing

Impressum / Imprint
Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Coverbild / Cover image: www.ingimage.com

Verlag / Publisher:
LAP LAMBERT Academic Publishing
ist ein Imprint der / is a trademark of
OmniScriptum GmbH & Co. KG
Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Germany
Email: info@lap-publishing.com

Herstellung: siehe letzte Seite /
Printed at: see last page
ISBN: 978-3-659-67722-9

Zugl. / Approved by: Tainan, Taiwan, National Cheng Kung University, Diss., 2010

Jerry Chun-Wei Lin

Tree-based Algorithms for Incremental, Utility, and Fuzzy Data Mining

ABSTRACT

Data mining, also referred to as knowledge discovery, has recently emerged as an important research topic, and association rules mining is considered as one of the most referenced sub-topics in data mining. In the past, traditional algorithms process all records in a batch way for mining association rules. In real-world applications, records are constantly being inserted, deleted or modified in dynamic databases. Designing an algorithm that can efficiently maintain association rules in dynamic databases is critically important. In the first part of this dissertation, three Pre-FUFP maintenance algorithms are thus proposed to efficiently maintain and update the FUFP-tree structures regardless of whether records are inserted, deleted or modified in dynamic databases. Based on two support thresholds of pre-large concepts, it helps avoid the need to re-build the tree structure until after a number of records have been processed. The FP-growth-like algorithm is then implemented to mine the desired information for the updated FUFP trees.

In the association rules mining, it treats items as binary variables in databases, which considers whether an item is bought in a record or not. Utility mining was thus proposed to reflect any other implicit factors, such as prices or profits. In the second part of this dissertation, a novel HUP-tree algorithm is proposed to efficiently mine the high utility itemsets based on the downward closure property. A HUP tree is first designed to keep the related information for later mining process. A HUP-growth mining algorithm is then presented to efficiently mine high utility itemsets from it.

In the past, most association rules mining focused on processing binary variables in databases. In recent years, many fuzzy data mining algorithms have been proposed for managing quantitative data, and most of them are processed in the level-wise approaches. In the third part of this dissertation, we attempt to extend the FP-tree algorithm for handling quantitative data from the global values of fuzzy regions. Thus, the fuzzy FP-tree algorithm,

the compressed fuzzy frequent pattern tree (CFFP-tree) algorithm, and the upper-bound fuzzy frequent pattern tree (UBFFP-tree) algorithm are then proposed to efficiently mine the fuzzy frequent itemsets. The maximum cardinality is used to make the number of fuzzy regions processed equivalent to the number of the original items for reducing the processing time. Three mining algorithm are then proposed to mine the fuzzy frequent itemsets based on the designed tree structures, respectively.

Experimental results showed that the performance of the proposed algorithms in three parts of this dissertation for handling association rules mining, high utility mining and fuzzy data mining, respectively.

Keywords: data mining, fuzzy data mining, utility mining, tree structure, maintenance algorithm.

List of Figures

List of Tables

Contents

CHAPTER 1

INTRODUCTION

1.1 Motivation

Years of effort in data mining have been produced a variety of efficient techniques. Depending on the type of databases processed, the mining approaches may be classified as finding association rules [2, 4-7, 10-11, 46-48, 59], classification rules [29, 52], clustering rules [34, 39], sequential patterns [3, 50, 53], among others. Among them, association rules mining is the most commonly seen in data mining. It consists of two main steps to derive the association rules from the transaction databases. The first step is to discover the frequent itemsets from databases based on the minimum support threshold; and the second one is to create the association rules from the found frequent itemsets during the first step based on the minimum confidence threshold. That is, mining frequent itemsets from databases is a fundamental task of finding association rules.

Numerous methods were proposed in the past to discover frequent itemsets, such as level-wise approaches and pattern-growth ones. In the level-wise approaches, most of which were based on the Apriori algorithm [2, 4-5], which generated and tested candidate itemsets level-by-level. In the pattern-growth approaches [1, 16, 19, 35, 56], most of which were based on the Frequent-Pattern-tree (FP-tree) structure [20] for efficiently mining association rules without generation of candidate itemsets. Both of the Apriori and the FP-tree mining approaches, however, are processed in the batch way. Cheung *et al.* then proposed the noticeable Fast UPdate (FUP) algorithm and FUP2 algorithm to maintain the discovered rules for record insertion [12] and record deletion [13], respectively. Hong *et al.* then attempted to modify the batch procedure of the FP-tree algorithm based on the FUP concept and proposed

a Fast Updated FP-tree (FUFP-tree) structure for easily updating the tree. Three maintenance algorithms were also proposed to maintain the FUFP tree whether the records are inserted [12, 27], deleted [13, 28] or modified [13, 26] in dynamic databases.

Although the FUP and FUP2 algorithms could indeed improve mining performance for record insertion and record deletion, the original databases still needed to be re-scanned whenever necessary. Hong *et al.* thus proposed three pre-large algorithms for record insertion [24], record deletion [21] and record modification [22], respectively, to further reduce the need for rescanning the original databases based on two support thresholds of pre-large algorithms. Based on the pre-large concepts, the original databases are unnecessary to rescan until a number of records have been processed. Since rescanning the databases spent much computation time, the maintenance cost could thus be reduced in the pre-large algorithms.

In the first part of this dissertation, a maintenance framework is proposed for effectively updating the constructed FUFP-tree structures and then deriving the desired frequent itemsets from it. It consisted of three Pre-FUFP maintenance algorithms for record insertion, record deletion, and record modification in dynamic databases, respectively. The proposed three maintenance algorithms do not require rescanning the original databases to re-construct the FUFP tree until a number of records have been processed. The number is determined from the two support thresholds and the size of the original databases. In the experimental results, the proposed three Pre-FUFP maintenance algorithms ran faster than the batch FP tree and FUFP tree but generated nearly the same number of tree nodes. That is, the proposed algorithms can thus achieve a good trade-off between execution time and tree complexity.

In the association rules mining, it treats all items in the databases as binary variables. That is, they only consider whether an item is bought in a record or not. In this case, frequent itemsets just reveal the occurrence importance of the itemsets in the records, but do not reflect any other implicit factors, such as prices or profits. Utility mining was thus proposed to partially solve the above problem [9, 41, 55]. Liu *et al.* then presented the two-phase

algorithm for fast discovering all high utility itemsets based on the downward-closure property to generate and test candidate high utility itemsets in a level-wise way [42]. The databases-scanning time is, however, a bottleneck of the approach. In second part of this dissertation, a new high utility pattern tree (HUP-tree) algorithm with the aid of the HUP-tree structure is first designed for mining high utility itemsets. An array is then attached to each node for keeping the quantities of its super-items in the path for later mining process. The HUP-growth mining algorithm based on the proposed HUP-tree structure is then presented to efficiently mine the high utility itemsets. In the experimental results, the proposed algorithm for mining high utility itemsets can thus be efficiently than the two-phase algorithm in a level-wise way.

In addition to binary variables in databases of association rules mining, transaction data in real-world applications, however, usually consisted of quantitative values. In recent years, the fuzzy set theory [32, 58] has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning. Several fuzzy learning algorithms for inducing rules from given sets of data have been designed and used to good effect with specific domains [8, 17, 33, 49, 51, 57]. Hong *et al.* proposed the fuzzy mining algorithms [23, 25] for managing quantitative data in a level-wise approach of Apriori algorithm. Papadimitriou *et al.* then proposed the fuzzy frequent pattern tree (FFPT) algorithm to find fuzzy association rules [45] in the pattern-growth approach. In the third part of this dissertation, we attempt to extend the FP-tree mining process for handling quantitative data from the global values of fuzzy regions. A fuzzy data mining framework is proposed to efficiently mine the fuzzy frequent itemsets from quantitative databases. It consists of three fuzzy data mining algorithms called fuzzy FP-tree algorithm, the compressed fuzzy frequent pattern tree (CFFP-tree) algorithm, and the upper-bound fuzzy frequent pattern tree (UBFFP-tree) algorithm for constructing the tree structures and mining the fuzzy frequent itemsets from it, respectively. Experimental results also show that the performance of the