Preface

Since 1981, the British National Conferences on Databases (BNCOD) have provided a forum for database researchers to report the latest progress and explore new ideas. Over the last 28 years, BNCOD has evolved from a predominantly national conference into one that is truly international, attracting research contributions from all over the world.

This volume contains the proceedings of BNCOD 2008. We received 45 submissions from 22 countries. Each paper was reviewed by three referees, and 14 full papers and 7 posters were accepted. All the research papers and posters are included in this volume, and they are organized into five sections: data mining and privacy, data integration, stream and event data processing, query processing and optimization, and posters.

The keynote was delivered by Monica Marinucci, EMEA Programme Director for Oracle in R&D. She has been involved in various advanced developments concerning Oracle, and participated in EC-funded projects as an expert, especially the CHALLENGERS special support action to propose the future of grid computing. In her keynote presentation, she addressed the audience on the topic of the power of data, emphasizing that the ability to store, handle, manipulate, distribute and replicate data and information can provide a tremendous asset to organizations. She also explored some of the latest directions and developments in the database field, and described how Oracle contributes to them partnering up with other leading organizations in different sectors.

BNCOD 2008 marked a special occasion in the BNCOD history – it was the 25th conference in the BNCOD series (BNCOD was not held in 1987, 1997 and 1999 when VLDB and ICDE were held in the UK). To mark this 25th anniversary, an International Colloquium on Advances in Database Research was held as part of the BNCOD main conference. Leading researchers were invited to the colloquium, and they presented and shared their latest research with the audience. The invited papers or the abstracts of the talks presented at the colloquium are included in this volume.

Two workshops were held alongside BNCOD 2008. The Workshop on Teaching, Learning and Assessment of Databases continued to address, as in the previous years, the issues concerning the educational aspects of databases. The papers from this workshop were published separately, and were not included in this volume. The Workshop on Biodiversity Informatics: Challenges in Modelling and Managing Biodiversity Knowledge aimed at advancing understanding in the challenges involved in this important, multi-disciplinary research area. The two best papers from this workshop are included in this volume.

Finally, we would like to thank all the authors for contributing their papers to BNCOD 2008, the referees for their effort in reviewing the papers, the EPSRC for their support of the International Colloquium on Advances in Database
Research, the Welsh Assembly Government for hosting the reception, and the Organizing Committee at Cardiff University for making this conference possible.

July 2008

Alex Gray
Keith Jeffery
Jianhua Shao
Conference Committees

Programme Committee

Alex Gray (Chair)          Cardiff University, UK
David Bell               Queen’s University, Belfast, UK
Stefan Böttcher          University of Paderborn, Germany
Peter Buneman            University of Edinburgh, UK
Sharma Chakravarthy      The University of Texas at Arlington, USA
Richard Cooper           University of Glasgow, UK
Alfredo Cuzzocrea        Universita della Calabria, Italy
Barry Eaglestone         University of Sheffield, UK
Suzanne Embury           University of Manchester, UK
A. Fernandes             University of Manchester, UK
Mary Garvey              University of Wolverhampton, UK
Georg Gottlob            University of Oxford, UK
Jun Hong                 Queen’s University, Belfast, UK
Ela Hunt                 Swiss Federal Institute of Technology, Switzerland
Wendy Ivins              Cardiff University, UK
Mike Jackson             Birmingham City University, UK
Anne James               Coventry University, UK
Keith Jeffery            Rutherford Appleton Laboratory, UK
Andrew Jones             Cardiff University, UK
Graham Kemp              Chalmers University, Sweden
Jessie Kennedy           Napier University, UK
Kavin Lu                 Brunel University, UK
L MacKinnon              University of Abertay Dundee, UK
Nigel Martin             Birkbeck, University of London, UK
Ken Moody                University of Cambridge, UK
David Nelson             University of Sunderland, UK
M Norrie                 ETH Zurich, Switzerland
Werner Nutt              The Free University of Bozen-Bolzano, Italy
Norman Paton             University of Manchester, UK
Alex Poulouvasilis       Birkbeck, University of London, UK
Mark Roantree            Dublin City University, Ireland
Alan Sexton              University of Birmingham, UK
Jianhua Shao             Cardiff University, UK
Paul Watson              Newcastle University, UK
Richard White            Cardiff University, UK
John Wilson              University of Strathclyde, UK
Jian Yang                Macquarie University, Australia
Steering Committee

Lachlan Mackinnon (Chair)  University of Abertay Dundee, UK
Richard Cooper  University of Glasgow, UK
Barry Eaglestone  University of Sheffield, UK
Jun Hong  Queen’s University, Belfast, UK
Anne James  Coventry University, UK
Keith Jeffery  Rutherford Appleton Laboratory, UK
David Nelson  University of Sunderland, UK
Alex Poulovassilis  Birkbeck, University of London, UK
Jianhua Shao  Cardiff University, UK

Organizing Committee

Nick Fiddian (Chair)  Cardiff University
Mikhaila Burgess  Cardiff University
Wendy Ivins  Cardiff University
Andrew Jones  Cardiff University
Jianhua Shao  Cardiff University
Gareth Shercliff  Cardiff University
Alysia Skilton  Cardiff University
Richard White  Cardiff University

25th Anniversary Colloquium Organization

Keith Jeffery  Rutherford Appleton Laboratory, UK
Jianhua Shao  Cardiff University, UK

Additional Referees

Mikhaila Burgess  Iadh Ounis
Grigorios Loukides  Punitha Swarny
# Table of Contents

## Keynote
The Power of Data .................................................. 1

## Data Mining and Privacy
Efficient Mining of Frequent Itemsets from Data Streams .......... 2

An Empirical Study of Utility Measures for $k$-Anonymisation ...... 15

HLS: Tunable Mining of Approximate Functional Dependencies ..... 28

Sentence Ordering for Coherent Multi-document Summary Generation ........................................ 40

## Data Integration
Schema Matching across Query Interfaces on the Deep Web .......... 51

A Generic Data Level Implementation of ModelGen .................. 63

Reconciling Inconsistent Data in Probabilistic XML Data Integration ... 75

## Stream and Event Data Processing
A Semantics for a Query Language over Sensors, Streams and Relations .................................................. 87

Load Shedding in MavStream: Analysis, Implementation, and Evaluation .................................................. 100

Event-Driven Database Information Sharing ................................ 113
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query Processing and Optimisation</strong></td>
<td></td>
</tr>
<tr>
<td>Smooth Interpolating Histograms with Error Guarantees</td>
<td>126</td>
</tr>
<tr>
<td>Virtual Forced Splitting, Demotion and the BV-Tree</td>
<td>139</td>
</tr>
<tr>
<td>A Functional Data Model Approach to Querying RDF/RDFS Data</td>
<td>153</td>
</tr>
<tr>
<td>Ranking for Approximated XQuery Full-Text Queries</td>
<td>165</td>
</tr>
<tr>
<td><strong>Poster Papers</strong></td>
<td></td>
</tr>
<tr>
<td>Role Based Access to Support Collaboration in Healthcare</td>
<td>177</td>
</tr>
<tr>
<td>A Peer-to-Peer Database Server</td>
<td>181</td>
</tr>
<tr>
<td>Checking the Integrity Constraints of Mobile Databases with Three-Level Model</td>
<td>185</td>
</tr>
<tr>
<td>Finding Data Resources in a Virtual Observatory Using SKOS Vocabularies</td>
<td>189</td>
</tr>
<tr>
<td>Progressive Ranking for Efficient Keyword Search over Relational Databases</td>
<td>193</td>
</tr>
<tr>
<td>Semantic Matching for the Medical Domain</td>
<td>198</td>
</tr>
<tr>
<td>Towards the Automatic Generation of Analytical End-User Tools</td>
<td></td>
</tr>
<tr>
<td>Metadata for Data Warehouses</td>
<td></td>
</tr>
<tr>
<td><strong>25th Anniversary Colloquium on Advances in Database Research</strong></td>
<td></td>
</tr>
<tr>
<td>The Hyperdatabase Project – From the Vision to Realizations</td>
<td>207</td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>From Schema and Model Translation to a Model Management System</td>
<td>227</td>
</tr>
<tr>
<td>XtreemOS: Towards a Grid-Enabled Linux-Based Operating System</td>
<td>241</td>
</tr>
<tr>
<td>High-Assurance Integrity Techniques for Databases</td>
<td>244</td>
</tr>
<tr>
<td>Towards General Temporal Aggregation</td>
<td>257</td>
</tr>
<tr>
<td><strong>Best Papers from the Workshop on Biodiversity Informatics:</strong></td>
<td></td>
</tr>
<tr>
<td>Challenges in Modelling and Managing Biodiversity Knowledge</td>
<td></td>
</tr>
<tr>
<td>Distributed Systems and Automated Biodiversity Informatics: Genomic Analysis and Geographic Visualization of Disease Evolution</td>
<td>270</td>
</tr>
<tr>
<td>Visualisation to Aid Biodiversity Studies through Accurate Taxonomic Reconciliation</td>
<td>280</td>
</tr>
<tr>
<td><strong>Author Index</strong></td>
<td>293</td>
</tr>
</tbody>
</table>
The Power of Data

Monica Marinucci
EMEA Oracle in R&D Programme
ORACLE Corporation
monica.marinucci@oracle.com

Abstract. From raw data to knowledge discovery, the ability to store, handle, manipulate, distribute and replicate data and information provides a tremendous asset to organisations, both in the commercial and scientific and academic sector. As IT infrastructures become more and more reliable, dynamic and flexible, applications and services require a modular, distributed, functionality-driven and secure environment.

Continual improvements and advancements are needed to meet these requirements and Oracle constantly innovates and pushes the limits of the technology to make data, information and knowledge available and exploitable. For Oracle to continue leading the database sector and to contribute to the development of ICT, it is crucial to engage with the R&D sector, both industrial and scientific, where existing and new challenges can be explored and novel and ground-breaking solutions found.

The talk will explore some of the latest directions and developments in the database field and describe how Oracle contributes to them partnering up with other leading organisations in different sectors. Highlights from some of the R&D projects across Europe Oracle is part of will be also presented.
Efficient Mining of Frequent Itemsets from Data Streams

Carson Kai-Sang Leung* and Dale A. Brajczuk

The University of Manitoba, Winnipeg, MB, Canada
{kleung,umbrajcz}@cs.umanitoba.ca

Abstract. As technology advances, floods of data can be produced and shared in many applications such as wireless sensor networks or Web click streams. This calls for efficient mining techniques for extracting useful information and knowledge from streams of data. In this paper, we propose a novel algorithm for stream mining of frequent itemsets in a limited memory environment. This algorithm uses a compact tree structure to capture important contents from streams of data. By exploiting its nice properties, such a tree structure can be easily maintained and can be used for mining frequent itemsets, as well as other patterns like constrained itemsets, even when the available memory space is small.

Keywords: Data mining, frequent itemset mining, frequent patterns, tree structure, limited memory space.

1 Introduction

Data mining aims to search for implicit, previously unknown, and potentially useful information and knowledge—such as frequent itemsets—that might be embedded in data (within traditional static databases or continuous data streams). The mining of frequent itemsets from large traditional static databases has been the subject of numerous studies since its introduction [1]. These studies can be broadly divided into two categories focusing functionality and performance. Regarding functionality, the central question considered is what (kind of patterns) to mine. While some studies [3,10] in this category considered the data mining exercise in isolation, some others explored how data mining can best interact with the human user. Examples of the latter include constrained mining [4,16,19,20,21] as well as interactive and online mining [13].

Regarding performance, the central question considered is how to mine the frequent itemsets as efficiently as possible. Studies in this category focused on fast Apriori-based algorithms [2] and their performance enhancements. Note that these Apriori-based algorithms depend on a generate-and-test paradigm. They compute frequent itemsets by generating candidates and checking their frequencies (i.e., support counts) against the transaction database. To improve efficiency of the mining process, Han et al. [11] proposed an alternative framework, namely a tree-based framework. The algorithm they proposed in this framework (the FP-growth algorithm) constructs an extended prefix-tree structure, called the Frequent Pattern tree (FP-tree), to capture the content of the...
transaction database. Rather than employing the generate-and-test strategy of Apriori-based algorithms, the FP-growth algorithm focuses on frequent pattern growth—which is a restricted test-only approach (i.e., does not generate candidates, and only tests for frequency). In the past few years, some other structures—such as the Co-Occurrence Frequent-Item tree (COFI-tree) [6,7]—have been proposed to further reduce the memory consumption of the FP-tree.

Moreover, over the past decade, the automation of measurements and data collection has produced tremendously huge amounts of data in many application areas. The recent development and increasing use of a large number of sensors has added to this situation. Consequently, these advances in technology have led to a flood of shared data. We are now drowning in streams of data but starving for knowledge. In order to be able to make sense of the streams of data, algorithms for extracting useful information and knowledge from these streams of data are in demand. This calls for stream mining [8,12,18,19].

In recent years, several stream mining algorithms have been proposed, and they can be broadly categorized into exact algorithms and approximate algorithms. Exact algorithms (e.g., Moment [5]) find truly frequent itemsets (i.e., itemsets with frequency not lower than the user-defined minimum frequency/support threshold \( \minsup \)). However, these algorithms usually aim to mine some special subsets of frequent itemsets (e.g., maximal, closed, or “short” itemsets) instead of all frequent itemsets. On the contrary, approximate algorithms (e.g., FP-streaming [9], FDPM [22]) mine all “frequent” itemsets. They do so by using approximate procedures, which may lead to some false positives or false negatives. In other words, these algorithms may find some infrequent itemsets or may miss (certain frequency information of) some frequent itemsets.

When comparing with mining from traditional static databases [1,14,15,17,20,21], mining from data streams is more challenging due to the following properties of data streams:

1. **Data streams are continuous and unbounded.** To find frequent itemsets from streams, we no longer have the luxury of performing multiple data scans. Once the streams flow through, we lose them. Hence, we need some techniques to capture the important contents of the streams (e.g., recent data—because users are usually more interested in recent data than older ones) and ensure that the captured data can fit into memory.

2. **Data in the streams are not necessarily uniformly distributed, and their distributions are usually changing with time.** A currently infrequent itemset may become frequent in the future, and vice versa. We have to be careful not to prune infrequent itemsets too early; otherwise, we may not be able to get complete information such as frequencies of certain itemsets, as it is impossible to recall those pruned itemsets.

Hence, some natural questions are: How can we effectively capture the important contents of the streams? Can we design a data structure that helps finding frequent itemsets from streams of data? To this end, we previously proposed a tree structure—called the Data Stream Tree (DSTree) [18]—to capture the contents of the streaming data. Like much other existing work [9,23], we made the same realistic assumption about enough memory space that the tree can fit into the main memory. While this assumption holds in many real-life situations, it may not hold in some other situations due to various factors (e.g., the nature of the streams, the window size, the \( \minsup \) value). When the amount
Table 1. Our proposed DSP-tree vs. the most relevant work

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal</strong></td>
<td>For mining traditional static DBs</td>
<td>For mining traditional static DBs</td>
<td>For mining data streams</td>
<td>For mining data streams</td>
</tr>
<tr>
<td><strong>No. of components in each tree node</strong></td>
<td>2 (item &amp; its frequency)</td>
<td>3 (item, its frequency, &amp; a participation counter)</td>
<td>2 (item &amp; a list of frequencies)</td>
<td>2 (item &amp; a counter)</td>
</tr>
<tr>
<td><strong>Contents of the tree</strong></td>
<td>Entire DB or projected DBs</td>
<td>DB for an item</td>
<td>Current batches of transactions from data streams</td>
<td>Current batches of data stream content for an item</td>
</tr>
</tbody>
</table>

of available memory space is small, the DSTree may not be able to fit into the main memory. Hence, we need a data structure that can work in environments with sufficient memory as well as insufficient memory space. We also need an algorithm that can use such a data structure to mine frequent itemsets for data streams.

The key contributions of this work are (i) the proposal of a simple, yet powerful, tree structure for capturing and maintaining relevant data found in the data streams; and (ii) the development of an efficient novel algorithm, which makes use of the developed tree structure, for mining frequent itemsets from streams of data in a limited memory environment. Experimental results in Section 5 show the effectiveness of our developed algorithm using our proposed tree structure in mining frequent itemsets from data streams. Table 1 summarizes the salient differences between our proposed DSP-tree and the most relevant alternatives.

This paper is organized as follows. In the next section, related work is described. Section 3 introduces our DSP-tree for stream mining in a limited memory environment. In Section 4, we discuss the applicability of the DSP-tree. Section 5 shows experimental results. Finally, conclusions are presented in Section 6.

2 Related Work

In this section, we discuss two groups of existing structures that are relevant to our work: (i) the COFI-tree [6,7] for mining static databases, and (ii) the DSTree [18] for stream mining.

2.1 Mining with the COFI-Tree

El-Hajj and Zaïane [6,7] proposed a Co-Occurrence Frequent-Item tree (COFI-tree) for mining frequent itemsets from traditional static databases. The key idea of mining with the COFI-tree can be described as follows. First, we scan the static database twice. The first database scan is to find frequencies of all domain items. These items are then sorted in descending frequency order. (As a preview, this global order of domain items will determine how the items are arranged in the COFI-tree.) The second database scan is to build a global tree (more precisely, a global FP-tree) to capture the contents of the database. Once the global tree is built, it can be used for constructing a COFI-tree for each frequent domain item $x$, from which frequent itemsets containing $x$ can be mined. More precisely, given a user-defined minsup threshold, we construct a COFI-tree for each domain item with frequency greater than or equal to minsup, starting