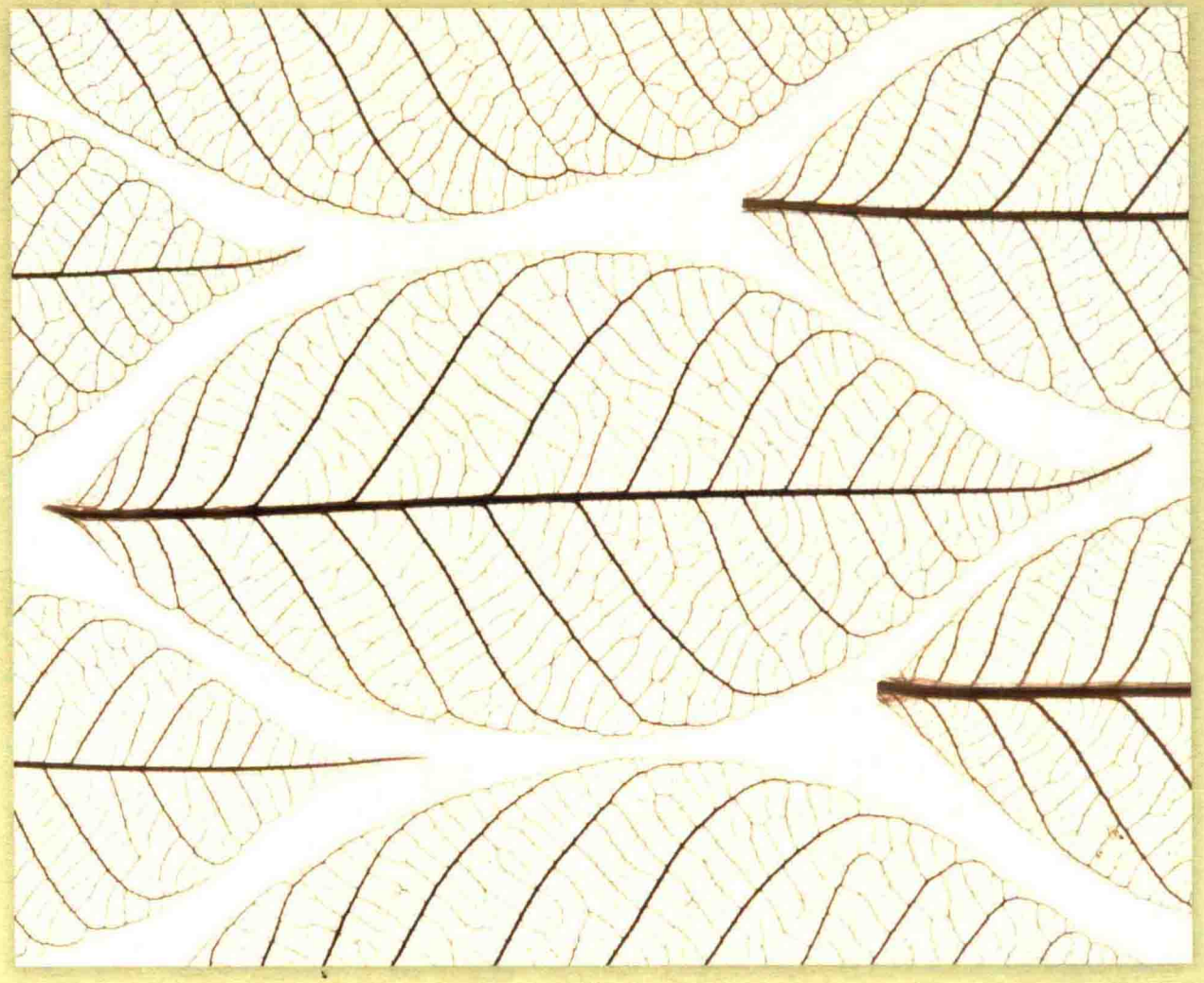
PREMIER REFERENCE SOURCE

Pattern Discovery Using Sequence Data Mining

Applications and Studies



Pradeep Kumar, P. Radha Krishna & S. Bapi Raju

Pattern Discovery Using Sequence Data Mining:

Applications and Studies

Pradeep Kumar
Indian Institute of Management Lucknow, India

P. Radha Krishna
Infosys Lab, Infosys Limited, India

S. Bapi Raju
University of Hyderabad, India

Information Science
REFERENCE

Senior Editorial Director:	Kristin Klinger
Director of Book Publications:	Julia Mosemann
Editorial Director:	Lindsay Johnston
Acquisitions Editor:	Erika Carter
Development Editor:	Joel Gamon
Production Editor:	Sean Woznicki
Typesetters:	Jennifer Romanchak, Lisandro Gonzalez
Print Coordinator:	Jamie Snavely
Cover Design:	Nick Newcomer

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Pattern discovery using sequence data mining : applications and studies / Pradeep Kumar, P. Radha Krishna, and S. Bapi Raju, editors.
p. cm.
Summary: "This book provides a comprehensive view of sequence mining techniques, and present current research and case studies in Pattern Discovery in Sequential data authored by researchers and practitioners"-- Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-61350-056-9 (hardcover) -- ISBN 978-1-61350-058-3 (print & perpetual access) -- ISBN 978-1-61350-057-6 (ebook) 1. Sequential pattern mining. 2. Sequential processing (Computer science) I. Kumar, Pradeep, 1977- II. Radha Krishna, P. III. Raju, S. Bapi, 1962-
QA76.9.D343P396 2012
006.3`12--dc22

2011008678

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

List of Reviewers

Manish Gupta, *University of Illinois at Urbana, USA*
Chandra Sekhar, *Indian Institute of Technology Madras, India*
Arnab Bhattacharya, *Indian Institute of Technology Kanpur, India*
Padmaja T Maruthi, *University of Hyderabad, India*
T. Ravindra Babu, *Infosys Technologies Ltd, India*
Pratibha Rani, *International Institute of Information Technology Hyderabad, India*
Nita Parekh, *International Institute of Information Technology Hyderabad, India*
Anass El-Haddadi, *IRIT, France*
Pinar Senkul, *Middle East Technical University, Turkey*
Jessica Lin, *George Mason University, USA*
Pradeep Kumar, *Indian Institute of Management Lucknow, India*
Raju S. Bapi, *University of Hyderabad, India*
P. Radha Krishna, *Infosys Lab, Infosys Limited, India*

Preface

A huge amount of data is collected every day in the form of sequences. These sequential data are valuable sources of information not only to search for a particular value or event at a specific time, but also to analyze the frequency of certain events or sets of events related by particular temporal/sequential relationship. For example, DNA sequences encode the genetic makeup of humans and all other species, and protein sequences describe the amino acid composition of proteins and encode the structure and function of proteins. Moreover, sequences can be used to capture how individual humans behave through various temporal activity histories such as weblog histories and customer purchase patterns. In general there are various methods to extract information and patterns from databases, such as time series approaches, association rule mining, and data mining techniques.

The objective of this book is to provide a concise state-of-the-art in the field of sequence data mining along with applications. The book consists of 14 chapters divided into 3 sections. The first section provides review of state-of-art in the field of sequence data mining. Section 2 presents relatively new techniques for sequence data mining. Finally, in section 3, various application areas of sequence data mining have been explored.

Chapter 1, *Approaches for Pattern Discovery Using Sequential Data Mining*, by Manish Gupta and Jiawei Han of University of Illinois at Urbana-Champaign, IL, USA, discusses different approaches for mining of patterns from sequence data. Apriori based methods and the pattern growth methods are the earliest and the most influential methods for sequential pattern mining. There is also a vertical format based method which works on a dual representation of the sequence database. Work has also been done for mining patterns with constraints, mining closed patterns, mining patterns from multi-dimensional databases, mining closed repetitive gapped subsequences, and other forms of sequential pattern mining. Some works also focus on mining incremental patterns and mining from stream data. In this chapter, the authors have presented at least one method of each of these types and discussed advantages and disadvantages.

Chapter 2, *A Review of Kernel Methods Based Approaches to Classification and Clustering of Sequential Patterns, Part I: Sequences of Continuous Feature Vectors*, was authored by Dileep A. D., Veena T., and C. Chandra Sekhar of Department of Computer Science and Engineering, Indian Institute of Technology Madras, India. They present a brief description of kernel methods for pattern classification and clustering. They also describe dynamic kernels for sequences of continuous feature vectors. The chapter also presents a review of approaches to sequential pattern classification and clustering using dynamic kernels.

Chapter 3 is *A Review of Kernel Methods Based Approaches to Classification and Clustering of Sequential Patterns, Part II: Sequences of Discrete Symbols* by Veena T., Dileep A. D., and C. Chandra Sekhar of Department of Computer Science and Engineering, Indian Institute of Technology Madras, India. The authors review methods to design dynamic kernels for sequences of discrete symbols. In their chapter they have also presented a review of approaches to classification and clustering of sequences of discrete symbols using the dynamic kernel based methods.

Chapter 4 is titled, *Mining Statistically Significant Substrings Based on the Chi-Square Measure*, contributed by Sourav Dutta of IBM Research India along with Arnab Bhattacharya Dept. of Computer Science and Engineering, Indian Institute of Technology, Kanpur, India. This chapter highlights the challenge of efficient mining of large string databases in the domains of intrusion detection systems, player statistics, texts, proteins, et cetera, and how these issues have emerged as challenges of practical nature. Searching for an unusual pattern within long strings of data is one of the foremost requirements for many diverse applications. The authors first present the current state-of-art in this area and then analyze the different statistical measures available to meet this end. Next, they argue that the most appropriate metric is the chi-square measure. Finally, they discuss different approaches and algorithms proposed for retrieving the top-k substrings with the largest chi-square measure. The local-maxima based algorithms maintain high quality while outperforming others with respect to the running time.

Chapter 5 is *Unbalanced Sequential Data Classification Using Extreme Outlier Elimination and Sampling Techniques*, by T. Maruthi Padmaja along with Raju S. Bapi from University of Hyderabad, Hyderabad, India and P. Radha Krishna, Infosys Lab, Infosys Technologies Ltd, Hyderabad, India. This chapter focuses on problem of predicting minority class sequence patterns from the noisy and unbalanced sequential datasets. To solve this problem, the atuhors proposed a new approach called extreme outlier elimination and hybrid sampling technique.

Chapter 6 is *Quantization Based Sequence Generation and Subsequence Pruning for Data Mining Applications* by T. Ravindra Babu and S. V. Subrahmanya of E-Comm. Research Lab, Education and Research, Infosys Technologies Limited, Bangalore, India, along with M. Narasimha Murty, Dept. of Computer Science and Automation, Indian Institute of Science, Bangalore, India. This chapter has highlighted the problem of combining data mining algorithms with data compaction used for data compression. Such combined techniques lead to superior performance. Approaches to deal with large data include working with a representative sample instead of the entire data. The representatives should preferably be generated with minimal data scans, methods like random projection, et cetera.

Chapter 7 is *Classification of Biological Sequences* by Pratibha Rani and Vikram Pudi of International Institute of Information Technology, Hyderabad, India, and it discusses the problem of classifying a newly discovered sequence like a protein or DNA sequence based on their important features and functions, using the collection of available sequences. In this chapter, the authors study this problem and present two techniques Bayesian classifiers: RBNBC and REBMEC. The algorithms used in these classifiers incorporate repeated occurrences of subsequences within each sequence. Specifically, RBNBC (Repeat Based Naive Bayes Classifier) uses a novel formulation of Naive Bayes, and the second classifier, REBMEC (Repeat Based Maximum Entropy Classifier) uses a novel framework based on the classical Generalized Iterative Scaling (GIS) algorithm.

Chapter 8, *Applications of Pattern Discovery Using Sequential Data Mining*, by Manish Gupta and Jiawei Han of University of Illinois at Urbana-Champaign, IL, USA, presents a comprehensive review of applications of sequence data mining algorithms in a variety of domains like healthcare, education, Web usage mining, text mining, bioinformatics, telecommunications, intrusion detection, et cetera.

Chapter 9, *Analysis of Kinase Inhibitors and Druggability of Kinase-Targets Using Machine Learning Techniques*, by S. Prashanthi, S. Durga Bhavani, T. Sobha Rani, and Raju S. Bapi of Department of Computer & Information Sciences, University of Hyderabad, Hyderabad, India, focuses on human kinase drug target sequences since kinases are known to be potential drug targets. The authors have also presented a preliminary analysis of kinase inhibitors in order to study the problem in the protein-ligand space in future. The identification of druggable kinases is treated as a classification problem in which druggable kinases are taken as positive data set and non-druggable kinases are chosen as negative data set.

Chapter 10, *Identification of Genomic Islands by Pattern Discovery*, by Nita Parekh of International Institute of Information Technology, Hyderabad, India addresses a pattern recognition problem at the genomic level involving identifying horizontally transferred regions, called genomic islands. A horizontally transferred event is defined as the movement of genetic material between phylogenetically unrelated organisms by mechanisms other than parent to progeny inheritance. Increasing evidence suggests the importance of horizontal transfer events in the evolution of bacteria, influencing traits such as antibiotic resistance, symbiosis and fitness, virulence, and adaptation in general. Considerable effort is being made in their identification and analysis, and in this chapter, a brief summary of various approaches used in the identification and validation of horizontally acquired regions is discussed.

Chapter 11, *Video Stream Mining for On-Road Traffic Density Analytics*, by Rudra Narayan Hota of Frankfurt Institute for Advanced Studies, Frankfurt, Germany along with Kishore Jonna and P. Radha Krishna, Infosys Lab, Infosys Technologies Limited, India, addresses the problem of estimating computer vision based traffic density using video stream mining. The authors present an efficient approach for traffic density estimation using texture analysis along with Support Vector Machine (SVM) classifier, and describe analyzing traffic density for on-road traffic congestion control with better flow management.

Chapter 12, *Discovering Patterns in Order to Detect Weak Signals and Define New Strategies*, by Anass El Haddadi of Université de Toulouse, IRIT UMR France Bernard Dousset, Ilham Berrada of Ensias, AL BIRONI team, Mohamed V University – Souissi, Rabat, Morocco presents four methods for discovering patterns in the competitive intelligence process: "correspondence analysis," "multiple correspondence analysis," "evolutionary graph," and "multi-term method." Competitive intelligence activities rely on collecting and analyzing data in order to discover patterns from data using sequence data mining. The discovered patterns are used to help decision-makers considering innovation and defining business strategy.

Chapter 13, *Discovering Patterns for Architecture Simulation by Using Sequence Mining,* by Pınar Senkul (Middle East Technical University, Computer Engineering Dept., Ankara, Turkey) along with Nilufer Onder (Michigan Technological University, Computer Science Dept., Michigan, USA), Soner Onder (Michigan Technological University, Computer Science Dept., Michigan, USA), Engin Maden (Middle East Technical University, Computer Engineering Dept., Ankara, Turkey) and Hui Meen Nyew (Michigan Technological University, Computer Science Dept., Michigan, USA), discusses the problem of designing and building high performance systems that make effective use of resources such as space and power. The design process typically involves a detailed simulation of the proposed architecture followed by corrections and improvements based on the simulation results. Both simulator development and result analysis are very challenging tasks due to the inherent complexity of the underlying systems. They present a tool called Episode Mining Tool (EMT), which includes three temporal sequence mining algorithms, a preprocessor, and a visual analyzer.

Chapter 14 is called *Sequence Pattern Mining for Web Logs* by Pradeep Kumar, Indian Institute of Management, Lucknow, India, Raju S. Bapi, University of Hyderabad, India and P. Radha Krishna,

Infosys Lab, Infosys Technologies Limited, India. In their work, the authors utilize a variation to the AprioriALL Algorithm, which is commonly used for the sequence pattern mining. The proposed variation adds up the measure Interest during every step of candidate generation to reduce the number of candidates thus resulting in reduced time and space cost.

This book can be useful to academic researchers and graduate students interested in data mining in general and in sequence data mining in particular, and to scientists and engineers working in fields where sequence data mining is involved, such as bioinformatics, genomics, Web services, security, and financial data analysis.

Sequence data mining is still a fairly young research field. Much more remains to be discovered in this exciting research domain in the aspects related to general concepts, techniques, and applications. Our fond wish is that this collection sparks fervent activity in sequence data mining, and we hope this is not the last word!

Pradeep Kumar
Indian Institute of Management Lucknow, India

P. Radha Krishna
Infosys Lab, Infosys Limited, India

S. Bapi Raju
University of Hyderabad, India

Section 1
Current State of Art

Table of Contents

Chapter 1
Applications of Pattern Discovery Using Sequential Data Mining

Manish Gupta
University of Illinois at Urbana-Champaign, USA

Jiawei Han
University of Illinois at Urbana-Champaign, USA

ABSTRACT

Sequential pattern mining methods have been found to be applicable in a large number of domains. Sequential data is omnipresent. Sequential pattern mining methods have been used to analyze this data and identify patterns. Such patterns have been used to implement efficient systems that can recommend based on previously observed patterns, help in making predictions, improve usability of systems, detect events, and in general help in making strategic product decisions. In this chapter, we discuss the applications of sequential data mining in a variety of domains like healthcare, education, Web usage mining, text mining, bioinformatics, telecommunications, intrusion detection, et cetera. We conclude with a summary of the work.

HEALTHCARE

Patterns in healthcare domain include the common patterns in paths followed by patients in hospitals, patterns observed in symptoms of a particular disease, patterns in daily activity and health data. Works related to these applications are discussed in this sub-section.

DOI: 10.4018/978-1-61350-056-9.ch001

Patterns in patient paths: The purpose of the French Diagnosis Related Group's information system is to describe hospital activity by focusing on hospital stays. (Nicolas, Herengt & Albuisson, 2004) propose usage of sequential pattern mining for patient path analysis across multiple healthcare institutions. The objective is to discover, to classify and to visualize frequent patterns among patient path. They view a patient path as a sequence of

sets. Each set in the sequence is a hospitalization instance. Each element in a hospitalization can be any symbolic data gathered by the PMSI (medical data source). They used the SLPMiner system (Seno & Karypis, 2002) for mining the patient path database in order to find frequent sequential patterns among the patient path. They tested the model on the 2002 year of PMSI data at the Nancy University Hospital and also propose an interactive tool to perform inter-institutional patient path analysis.

Patterns in dyspepsia symptoms: Consider a domain expert, who is an epidemiologist and is interested in finding relationships between symptoms of dyspepsia within and across time points. This can be done by first mining patterns from symptom data and then using patterns to define association rules. Rules could look like ANOREX2=0 VOMIT2=0 NAUSEA3=0 ANOREX3=0 VOMIT3=0 ⇒ DYSPH2=0 where each symptom is represented as <symptom>N=V (time=N and value=V). ANOREX (anorexia), VOMIT (vomiting), DYSPH (dysphagia) and NAUSEA (nausea) are the different symptoms. However, a better way of handling this is to define subgroups as a set of symptoms at a single time point. (Lau, Ong, Mahidadia, Hoffmann, Westbrook, & Zrimec, 2003) solve the problem of identifying symptom patterns by implementing a framework for constraint based association rule mining across subgroups. Their framework, Apriori with Subgroup and Constraint (ASC), is built on top of the existing Apriori framework. They have identified four different types of phase-wise constraints for subgroups: constraint across subgroups, constraint on subgroup, constraint on pattern content and constraint on rule. A constraint across subgroups specifies the order of subgroups in which they are to be mined. A constraint on subgroup describes the intra-subgroup criteria of the association rules. It describes a minimum support for subgroups and a set of constraints for each subgroup. A constraint on pattern content outlines the inter-subgroup criteria on association rules. It describes the criteria on the relationships between subgroups. A constraint on rule outlines the composition of an association rule; it describes the attributes that form the antecedents and the consequents, and calculates the confidence of an association rule. It also specifies the minimum support for a rule and prunes away item-sets that do not meet this support at the end of each subgroup-merging step. A typical user constraint can look like [1,2,3][1, a=A1&n<=2][2, a=B1&n<=2][3, v=1][rule, (s1 s2) ⇒s3]. This can be interpreted as: looking at subgroups 1, 2 and 3, from subgroup 1, extract patterns that contain the attribute A1 (a=A1) and contain no more than 2 attributes (n<=2); from subgroup 2, extract patterns that contain the attribute B1 (a=B1) and contain no more than 2 attributes (n<=2); then from subgroup 3, extract patterns with at least one attribute that has a value of 1 (v=1). Attributes from subgroups 1 and 2 form the antecedents in a rule, and attributes from subgroup 3 form the consequents ([rule, (s1 s2) ⇒ s3]). Such constraints are easily incorporated into the Apriori process by pruning away more candidates based on these constraints.

They experimented on a dataset with records of 303 patients treated for dyspepsia. Each record represented a patient, the absence or presence of 10 dyspepsia symptoms at three time points (initial presentation to a general practitioner, 18 months after endoscopy screening, and 8–9 years after endoscopy) and the endoscopic diagnosis for the patient. Each of these symptoms can have one of the following three values: symptom present, symptom absent, missing (unknown). At each of the three time points, a symptom can take any of these three possible values. They show that their approach leads to interesting symptom pattern discovery.

Patterns in daily activity data: There are also works, which investigate techniques for using agent-based smart home technologies to provide at-home automated assistance and health monitoring. These systems first learn patterns from at-home health and activity data. Further, for any

new test cases, they identify behaviors that do not conform to normal behavior and report them as predicted anomalous health problems.

EDUCATION

In the education domain, work has been done to extract patterns from source code and student teamwork data.

Patterns in source code: A coding pattern is a frequent sequence of method calls and control statements to implement a particular behavior. Coding patterns include copy-and-pasted code, crosscutting concerns (parts of a program which rely on or must affect many other parts of the system) and implementation idioms. Duplicated code fragments and crosscutting concerns that spread across modules are problematic in software maintenance. (Ishio, Date, Miyake, & Inoue, 2008) propose a sequential pattern mining approach to capture coding patterns in Java programs. They define a set of rules to translate Java source code into a sequence database for pattern mining, and apply PrefixSpan algorithm to the sequence database. They define constraints for mining source code patterns. A constraint for control statements could be: If a pattern includes a LOOP/IF element, the pattern must include its corresponding element generated from the same control statement. They classify sub-patterns into pattern groups. As a case study, they applied their tool to six open-source programs and manually investigated the resultant patterns.

They identify about 17 pattern groups which they classify into 5 categories:

1. A boolean method to insert an additional action: <Boolean method>, <IF>, <action-method>, <END-IF>
2. A boolean method to change the behavior of multiple methods: <Boolean method>, <IF>, <action-method>, <END-IF>
3. A pair of set-up and clean-up: <set-up method>, <misc action>, ..., <clean-up method>
4. Exception Handling: Every instance is included in a try-catch statement.
5. Other patterns.

They have made this technique available as a tool: Fung(http://sel.ist.osaka-u.ac.jp/~ishio/fung/)

Patterns in student team-work data: (Kay, Maisonneuve, Yacef, & Zaïane, 2006) describe data mining of student group interaction data to identify significant sequences of activity. The goal is to build tools that can flag interaction sequences indicative of problems, so that they can be used to assist student teams in early recognition of problems. They also want tools that can identify patterns that are markers of success so that these might indicate improvements during the learning process. They obtain their data using TRAC which is an open source tool designed for use in software development projects. Students collaborate by sharing tasks via the TRAC system. These tasks are managed by a "Ticket" system; source code writing tasks are managed by a version control system called "SVN"; students communicate by means of collaborative web page writing called "Wiki". Data consist of events where each event is represented as Event = {EventType, ResourceId, Author, Time} where: EventType is one of T (for Ticket), S (for SVN), W (for Wiki). One such sequence is generated for each of the group of students.

The original sequence obtained for each group was 285 to 1287 long. These event sequences were then broken down into several "sequences" of events using a per session approach or a per resource approach. In breakdown per session approach, date and the resourceId are omitted and a sequence is of form: (iXj) which captures the number of i consecutive times a medium X was used by j different authors, e.g., <(2T1), (5W3), (2S1),(1W1)>. In breakdown per resource ap-

proach, sequence is of form <iXj, t> which captures the number of i different events of type X, the number j of authors, and the number of days over which t the resource was modified, e.g., <10W5, 2>. In a follow-up paper (Perera, Kay, Yacef, & Koprinska, 2007), they have a third approach, breakdown by task where every sequence is of the form (i,X,A) which captures the number of consecutive events (i) occurring on a particular TRAC medium (X), and the role of the author (A).

Patterns observed in group sessions: Better groups had many alternations of SVN and Wiki events, and SVN and Ticket events whereas weaker groups had almost none. The best group also had the highest proportion of author sessions containing many consecutive ticket events (matching their high use of ticketing) and SVN events (suggesting they committed their work to the group repository more often).

A more detailed analysis of these patterns revealed that the best group used the Ticket more than the Wiki, whereas the weakest group displayed the opposite pattern. The data suggested group leaders in good groups were much less involved in technical work, suggesting work was being delegated properly and the leader was leading rather than simply doing all the work. In contrast, the leaders of the poorer groups either seemed to use the Wiki (a less focused medium) more than the tickets, or be involved in too much technical work.

Patterns observed in task sequences: The two best groups had the greatest percentage support for the pattern (1,t,L)(1,t,b), which were most likely tickets initiated by the leader and accepted by another team member. The fact this occurred more often than (1,t,L)(2,t,b), suggests that the better groups were distinguished by tasks being performed on the Wiki or SVN files before the ticket was closed by the second member. Notably, the weakest group had higher support for this latter pattern than the former. The best group was one of the only two to display the patterns (1,t,b)(1,s,b) and (1,s,b)(1,t,b) – the first likely being a ticket being accepted by a team member and then SVN work relating to that task being completed and the second likely being work being done followed by the ticket being closed. The close coupling of task-related SVN and Wiki activity and Ticket events for this group was also shown by relatively high support for the patterns (1,t,b)(1,t,b)(1,t,b), (1,t,b)(1,s,b)(1,t,b) and (1,t,b)(1,w,b)(1,t,b). The poorest group displayed the highest support for the last pattern, but no support for the former, again indicating their lack of SVN use in tasks.

Patterns observed in resource sequences: The best group had very high support for patterns where the leader interacted with group members on tickets, such as (L,1,t)(b,1,t)(L,1,t). The poorest group in contrast lacked these interaction patterns, and had more tickets which were created by the Tracker rather than the Leader, suggestive of weaker leadership. The best group displayed the highest support for patterns such as (b,3,t) and (b,4,t), suggestive of group members making at least one update on tickets before closing them. In contrast, the weaker groups showed support mainly for the pattern (b,2,t), most likely indicative of group members accepting and closing tickets with no update events in between.

Web Usage Mining

The complexity of tasks such as Web site design, Web server design, and of simply navigating through a Web site has been increasing continuously. An important input to these design tasks is the analysis of how a Web site is being used. Usage analysis includes straightforward statistics, such as page access frequency, as well as more sophisticated forms of analysis, such as finding the common traversal paths through a Web site. Web Usage Mining is the application of pattern mining techniques to usage logs of large Web data repositories in order to produce results that can be used in the design tasks mentioned above. However, there are several preprocessing tasks that

must be performed prior to applying data mining algorithms to the data collected from server logs.

Transaction identification from web usage data: (Cooley, Mobasher, & Srivastava, 1999) present several data preparation techniques in order to identify unique users and user sessions. Also, a method to divide user sessions into semantically meaningful transactions is defined. Each user session in a user session file can be thought of in two ways; either as a single transaction of many page references, or a set of many transactions each consisting of a single page reference. The goal of transaction identification is to create meaningful clusters of references for each user. Therefore, the task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. This process can be extended into multiple steps of merge or divide in order to create transactions appropriate for a given data mining task. Both types of approaches take a transaction list and possibly some parameters as input, and output a transaction list that has been operated on by the function in the approach in the same format as the input. They consider three different ways of identifying transactions based on: Reference Length (time spent when visiting a page), Maximal Forward Reference (set of pages in the path from the first page in a user session up to the page before a backward reference is made) and Time Window.

By analyzing this information, a Web Usage Mining system can determine temporal relationships among data items such as the following Olympics Web site examples:

- 9.81% of the site visitors accessed the Atlanta home page followed by the Sneakpeek main page.
- 0.42% of the site visitors accessed the Sports main page followed by the Schedules main page.

Patterns for customer acquisition: (Buchner & Mulvenna, 1998) propose an environment that allows the discovery of patterns from trading related web sites, which can be harnessed for electronic commerce activities, such as personalization, adaptation, customization, profiling, and recommendation.

The two essential parts of customer attraction are the selection of new prospective customers and the acquisition of the selected potential candidates. One marketing strategy to perform this exercise, among others, is to find common characteristics in already existing visitors' information and behavior for the classes of profitable and non-profitable customers. The authors discover these sequences by extending GSP so it can handle duplicates in sequences, which is relevant to discover navigational behavior.

```
A found sequence looks as the
  following:
{ecom.infm.ulst.ac.uk/, ecom.infm.
ulst.ac.uk/News_Resources.html, ecom.
infm.ulst.ac.uk/Journals.html, ecom.
infm.ulst.ac.uk/, ecom.infm.ulst.
ac.uk/search.htm} Support = 3.8%;
Confidence = 31.0%
```

The discovered sequence can then be used to display special offers dynamically to keep a customer interested in the site, after a certain page sequence with a threshold support and/or confidence value has been visited.

Patterns to Improve Web Site Design

For the analysis of visitor navigation behavior in web sites integrating multiple information systems (multiple underlying database servers or archives), (Berendt, 2000) proposed the web usage miner (WUM), which discovers navigation patterns subject to advanced statistical and structural constraints. Experiments with a real web site that integrates data from multiple databases,