THE 18TH EUROPEAN CONFERENCE ON MACHINE LEARNING
AND
THE 11TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE
OF KNOWLEDGE DISCOVERY IN DATABASES

PROCEEDINGS OF THE

ECML/PKDD'2007
DISCOVERY CHALLENGE

September 17, 2007

Warsaw, Poland

Editors:
Hung Son Nguyen
Institute of Mathematics, Warsaw University

Typesetting:
Hung Son Nguyen

Preface

Knowledge discovery in real-world databases requires a broad scope of techniques and forms of knowledge. Both the knowledge and the applied methods should fit the discovery tasks and should adapt to knowledge hidden in the data. The Discovery Challenge will encourage a collaborative research effort, a broad and unified view of knowledge and methods of discovery, and emphasis on business problems and solutions to those problems.

The idea of Discovery Challenge came from Jan Żytkow, who suggested to organize such an event during PKDD'99 in Prague. The Discovery Challenge constitutes a collection of data and problems as a common ground for better comparisons and discussions of the applicability of KDD methods on a real-world problems with respect to both KDD and application viewpoints. The main goals of the Discovery Challenge are

- stimulate an open view of knowledge and discovery
- stimulate collaborative approach to KDD and research on unification of both different forms of knowledge and discovery
- integrate into KDD an emphasis on real-world problems and solutions to those problems

This year's **Discovery Challenge** was devoted to three problems: user behaviour ptrediction from web traffic logs, HTTP traffic classification, and Sumerian literature understanding. The Challenge was co-organized by Piotr Ejdys(Gemius SA), Hung Son Nguyen (Warsaw University), Pascal Poncelet (EMA-LGI2P) and Jerzy Tyszkiewicz (Warsaw University).

1. **User's behaviour prediction**
   This task is co-organized by Gemius, the leading Internet market research company in Poland. The problem objective is to predict user behaviour by characterising nature of user's visit, i.e., the list categories of the visited Internet portal and the number of page views in each category. The challenge is accomplished with use of web traffic data from Polish web sites employing gemiusTraffic study, grouped by appropriate categories.
2. **HTML traffic prediction**
   The task is co-organized by the LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, FRANCE) and the LGI2P (Ecole des Mines d'Alčs, FRANCE) and is based on a dataset of real-world web traffic in conjunction with Bee Ware, a leading provider of secure web enabled delivery solutions. The aim of this contest is to classify HTTP traffic into multiple classes, distinguishing between different attack types but also between anomalous and normal traffic. The procedure involves 3 tasks respectively handling classification, pattern isolation and performance issues.
3. **Sumerian Literature understanding**
   The challenge is related to a database of ca. 28'000 administrative documents from the kingdom of the III Dynasty of Ur, which existed in the 21st century b.C. in

Mesopotamia, present day southern Iraq. The documents were originally written in Sumerian on clay tablets, using cuneiform script.
The challenge is to discover useful information about water and spirit transport system using available transliterated documents in the text format.

There were 122 participants registering to the Discovery Challenge website. However, only the first challenge was attacked by more than five participants. The final classification for the first task is as follows:

The winner:

- "Auto-regressive and Score maximizing Approach" by Krzysztof Dembczynski and Wojciech Kotlowski from Poznan University of Technology, Poland and Marcin Sydow from Polish-Japanese Institute of Information Technology, Poland

Runners-up:

- "Bayesian Inference Approach" by Malik Tahir Hassan, Khurum Nazir Junejo and Asim Karim from Lahore University, Pakistan
- "Frequent Item Approach" by Tung-Ying Lee from National Tsing Hua University, Taiwan

I would like to express my gradtitute to Dr Petr Berka, Dr Steffen Bickel and Dr Bruno Cremilleux – the chairs of previous events of PKDD/ECML Discovery Challenge – for their help. I also indebted to Marcin Szczuka – the ECML/PKDD Local Chair – for his assistance on setting up the Discovery Challenge website. Last but not least I would like thank our sponsors for their great contributions for the success of ECML/PKDD Discovery Challenge.

Warsaw, August 2007 Hung Son Nguyen

ECML/PKDD'2007 Discovery Challenge Organization

Program Committee

Petr Berka
Steffen Bickel
Bruno Cremilleux
Piot Ejdys (co-chair)
Joanna Jaworska
Wojciech Jaworski
Hung Son Nguyen (chair)
Pascal Poncelet (co-chair)
Maguelonne Teisseire
Jerzy Tyszkiewicz (co-chair)

Table of Contents

Users' Behaviour Prediction Challenge

Joanna Jaworska[1], Hung Son Nguyen[2]

[1] Gemius SA,
Wołoska 7 Str., Mars Building, Staircase D, 2nd Floor
Warsaw 02-675, Poland
[2] Institute of Mathematics, Warsaw University
Banacha 2, 02-097 Warsaw, Poland

Abstract. This task is co-organized by Gemius SA, the leading Internet market research agency in Central and Eastern Europe. The problem objective is to predict user behaviour by characterising nature of user's visit, i.e., the list categories of the visited Internet portal and the number of page views in each category. The challenge is accomplished with use of web trafic data from Polish web sites employing gemiusTraffic study, grouped by appriopriate categories.
This will be accomplished with use of web traffic data from Polish web sites employing gemiusTraffic study, grouped by appropriate categories. The above defined objective has been divided into three separate challenge problems:
Problem 1 is related to the length of the visit. A visit – accordingly to the definition – is a sequence of page views by one user (cookie). As web pages are identified by their categories, during one visit user may view pages of one or more categories. The problem is to predict whether a given visit is short (1 category) or long (two or more categories). Problem 2 is related to the most probable categories. Solution of Problem 2 is a list of the most probable categories in a given visit of a given user. Problem 3 is to predict the most probable categories and ranges of numbers of page views. Solution of Problem 3 is a list of the most probable categories in a given visit of a given user with range of number of page views in each category.

Full article in PDF

Effective Prediction of Web User Behaviour with User-Level Models.

Krzysztof Dembczyński[1], Wojciech Kotłowski[1], Marcin Sydow[2]

[1] Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
[2] Polish-Japanese Institute of Information Technology, 02-008 Warsaw, Poland

Abstract. The paper presents our solution to the ECML/PKDD 2007 Challenge Task that concerns prediction of Internet user behaviour by characterising the nature of their Web page visits. Our solution has low time and space complexity, scales well with large datasets and, at the same time, produces high-quality results. Comparison of the performance of our ultimate approach with a suit of other approaches that we examined exhibits its superiority and some hardness of the given problem.

Full article in PDF

Bayesian Inference for Web Surfer Behavior Prediction

Malik Tahir Hassan, Khurum Nazir Junejo, Asim Karim

Dept. of Computer Science,
Lahore University of Management Sciences
Lahore, Pakistan
{mhassan, junejo, akarim}@lums.edu.pk

Abstract. The accurate prediction of user behavior on the Web has immense commercial value as the Web evolves into a primary medium for marketing and sales for many businesses. This broad and complex problem can be broken down into three more understandable problems: predicting (1) short and long visit sessions, (2) first three most probable categories of pages visited in a session, and (3) number of page views per category in a visit session. We present Bayesian solutions to these problems. The focus in our solutions is accuracy and computational efficiency rather than modeling the complex Web surfer behavior. We evaluate our solutions on four weeks of surfer data made available by the ECML/PKDD Discovery Challenge. Probabilities are estimated from the first three weeks of data and the resulting Bayesian models tested on last week's data. The results confirm the high accuracy and good efficiency of our solutions.

Predicting User's Behavior by the Frequent Items

Tung-Ying Lee

Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
tylee@cs.nthu.edu.tw

Abstract. Frequent items, frequent itemsets, frequent sequences, and graphical models have been used in predicting user's behavior. However, we found that frequent items are robust, time-efficient, and meaningful in the view of statistics. Our method is based on frequent items. The experiments are performed on a real dataset which is provided in ECML/PKDD Discovery Challenge. For most situations, we found that predicting from frequent items can obtain more convincing results than predicting from multiple Markov chains.
Keywords: Frequent Items, Markov Chains

Full article in PDF

Stacking Heterogeneous Data Resources for addressing the ECML-PKDD 2007 Discovery Challenge 1

Dimitrios Mavroeidis[1], Charis Brisagotis[1], Dimitris Drosos[1] and Michalis Vazirgiannis[1,2]

[1] Department of Informatics, Athens University of Economics and Business, Greece
[2] GEMO Team, INRIA/FUTURS, France

Abstract. In this paper we analyze and describe the approach undertaken by the DB-NET research group for addressing the ECML/PKDD Discovery Challenge 1. The Discovery Challenge was concerned with the construction of predictive models regarding several features of a user's browsing behavior. The training data consisted of attributes that described the users' browsing behavior (such as timestamp of visit and category-view paths) as well as information concerning their social and demographic profile (such as country, city, browser and operating system). The challenge that we faced in addressing this task, was the fact that the data resources, that could be employed for deriving the user behavior models, were heterogeneous. More precisely, for each user one could build predictive models based solely upon his personal browsing behavior, or incorporate features that described his social and demographic profile. In order to address this issue in a principled manner we have employed Stacking. Stacking was based on seven level-0 classifiers built upon seven heterogeneous datasets (constructed by the available training data). Consequently, for producing the user-specific models, we have compared the cross-validated accuracy of each level-0 and the level-1 classifiers (for each user), and finally selected the one that exhibited the highest accuracy estimate. The level-0 and level-1 classifiers were chosen after thorough experimentation to be C4 and Logistic Regression respectively. Apart from a detailed description and a discussion of the submitted model (which was solely for task 1 of the challenge), this paper contains our preliminary experimentations and experimental results for task 2.

Web Analyzing Traffic Challenge: Description and Results

Chedy Raïssi[1,2], Johan Brissaud[3],
Gérard Dray[2], Pascal Poncelet[2], Mathieu Roche[1], and Maguelonne Teisseire[1]

[1] LIRMM - UMR 5506, CNRS, Univ. Montpellier 2, France
[2] LGI2P, EMA Site EERIE – Parc Scientifique Georges Besse, Nîmes, France
[3] BEE WARE Company, 210, Avenue Frederic Joliot, 13852 Aix-en-Provence, France
raissi@lirmm.fr, jbrissaud@bee-ware.net
{gerard.dray,pascal.poncelet}@ema.fr,{mroche,teisseire}@lirmm.fr

Abstract. This paper describes the Web Analyzing Traffic Challenge (Discovery Challenge of ECML/PKDD'07) and the results. Using the data from query logs it is possible to recognize an attack and define its class. Then the aim of this challenge is the filtering of attacks in Web traffic.
Keywords: attack detection, classification.

Full article in PDF

ECML/PKDD Challenge: Analyzing Web Traffic
A Boundaries Signature Approach

Matthieu Exbrayat

Laboratoire d'Informatique Fondamentale d'Orléans
Université d'Orléans, B.P. 6759
45067 ORLEANS Cedex 2, France
Matthieu.Exbrayat@univ-orleans.fr

Abstract. To detect HTTP attacks (and the corresponding interval) we propose a signature detection method which concentrates on both ends of attack patterns. We also use a textual word / symbolic word model to describe attack patterns.

Full article in PDF

Feature Extraction from Web Traffic data for the Application of Data Mining Algorithms in Attack Identification

Konstantinos Pachopoulos[1], Dialekti Valsamou[1], Dimitrios Mavroeidis[1] and Michalis Vazirgiannis[1,2]

[1] Department of Informatics, Athens University of Economics and Business, Greece
[2] GEMO Team, INRIA/FUTURS, France

Abstract. In this paper we present and discuss the approach undertaken by the DB-NET research group for addressing the ECML/PKDD Discovery Challenge 2. The challenge was concerned with the analysis of web traffic data with the aim of constructing predictive models that can identify possible future attacks. The training data provided for the challenge consisted of a collection of pre-classified traffic data into 8 categories; one containing the valid (non-malicious communications), while the other 7 contained several types of web attacks. The attack-types were: Cross-Site Scripting, SQL Injection, LDAP Injection, XPATH Injection, Path traversal, Command execution and SSI. A challenge that we faced stemmed from the fact that the training data were provided in a preliminary HTTP protocol format, containing string representations of the HTTP packet fields (such as method, protocol, and uri). This information could not be directly incorporated in standard data mining algorithms, and significant preprocessing should be performed. In order to address this challenge we have identified several string patterns that could signify a malicious communication, and transformed the unstructured information to feature-vector format. This transformation allowed us to employ C4, a decision tree algorithm that exhibited an estimated accuracy of 77%.

Full article in PDF

Water transport in Sumer in the kingdom of the III Dynasty of Ur

Marek Stępień, Jerzy Tyszkiewicz, Wojciech Jaworski

Institute of Computer Science, Warsaw University
Banacha 2, 02-097 Warsaw, Poland

Full article in PDF

Using Semi-supervised Learning for Mining Sumerian Administrative Documents in the Kingdom of the III Dynasty of Ur

Dimitrios Mavroeidis[1], Dimitris Diamantis[1] and Michalis Vazirgiannis[1,2]

[1] Department of Informatics, Athens University of Economics and Business, Greece
[2] GEMO Team, INRIA/FUTURS, France

Abstract. In this paper we present and discuss the approach undertaken by the DB-NET research group for addressing the ECML/PKDD Discovery Challenge 3. The challenge was concerned with the analysis of a collection of administrative documents from the kingdom of the III Dynasty of Ur, which existed in the 21st century b.C. in Mesopotamia, located in present day southern Iraq. The water transport system of the era was very developed, undertaking a substantial part in the economic growth and development of the region. The task required that data mining techniques were applied on a collection of administrative documents, with the purpose of identifying and the dating the documents related to the water transport system. In the proposed solution, we have formulated the identification problem of the water transport-related documents as a semi-supervised clustering task. Our methodological approach was motivated by the fact that we could easily identify several documents that belonged in the water transport cluster, using simple keyword matching rules. These documents were consequently used as prior knowledge in the context of a 2-way semi-supervised clustering algorithm, where one cluster was defined as containing the water transport-related documents, while the other contained the rest of the documents. Concerning the document dating process, we have observed that the exact dating (defined through kingdom eras) could be extracted from the majority of the documents in the collection. This allowed us to train a Support Vector Machine and derive a document dating model that was consequently employed for dating the rest of the documents. In order to identify the main elements of the waterway transport system we have analyzed statistically the variables of the instances belonging in the waterway cluster and its centroid. Moreover, we have used Information Gain in order to identify the variables that can can be used for separating the two clusters.

Full article in PDF

Author Index