

Communications in Computer and Information Science

794

Commenced Publication in 2007

Founding and Former Series Editors:

Phoebe Chen, Alfredo Cuzzocrea, Xiaoyong Du, Orhun Kara, Ting Liu,
Krishna M. Sivalingam, Dominik Ślęzak, Takashi Washio, Xiaokang Yang,
and Junsong Yuan

Editorial Board Members

Simone Diniz Junqueira Barbosa 

*Pontifical Catholic University of Rio de Janeiro (PUC-Rio),
Rio de Janeiro, Brazil*

Joaquim Filipe 

Polytechnic Institute of Setúbal, Setúbal, Portugal

Ashish Ghosh

Indian Statistical Institute, Kolkata, India

Igor Kotenko 

*St. Petersburg Institute for Informatics and Automation of the Russian
Academy of Sciences, St. Petersburg, Russia*

Lizhu Zhou

Tsinghua University, Beijing, China

More information about this series at <http://www.springer.com/series/7899>


Vadim V. Strijov · Dmitry I. Ignatov ·
Konstantin V. Vorontsov (Eds.)

Intelligent Data Processing

11th International Conference, IDP 2016
Barcelona, Spain, October 10–14, 2016
Revised Selected Papers

Editors

Vadim V. Strijov 
Moscow Institute of Physics and Technology
Dolgoprudny, Russia

Konstantin V. Vorontsov 
Yandex School of Data Analysis
Moscow, Russia

Dmitry I. Ignatov 
National Research University Higher School
of Economics
Moscow, Russia

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-030-35399-5 ISBN 978-3-030-35400-8 (eBook)
<https://doi.org/10.1007/978-3-030-35400-8>

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The International Conference on Intelligent Information Processing: Theory and Applications is a premier forum for the data science researchers and professionals to discuss, distribute, and advance the state of research and development of the data analysis field. The conference facilitates the exchange of insights and innovations between the industry and academia, each represented by leaders in their respective fields. The IDP Conference has a rich history, starting from 1989 up till now. The conference offers research and industry tracks in the areas of Data Mining, Machine Learning, Big Data Analytics, Deep Learning, Computer Vision, Text Mining, Social Networks Analysis, and Data Science for Biology and Medicine.

The conference gathers plenty of highly recognized speakers from large industrial companies and academic research and educational organizations. The goal is to encourage ideas, success stories, and challenges sharing them among big data research and industry communities and establish connections between them and a talent pool of 200 experienced data scientists in our community. The invited speakers were selected from a wide list of recognizable candidates. Wonderful speakers came such as Paolo Rosso (Associate Professor at Universidad Politécnica de Valencia), Kamran Elahian (Chairman of Global Catalyst Partners), Konstantin V. Vorontsov (Professor at the Moscow Institute of Physics and Technology), Michael Levin (Chief Data Scientist at Yandex Data Factory), and Victor Lempitsky (Leader of Computer Vision Group at Skolkovo Institute of Science and Technology), among the other. Talks, fireside chats, research, and industry sessions as well as exhibitions were held during the event.

The Editorial Board and the conference Program Committee accepted papers representing the latest achievements in the theory and practice of intelligent data processing. Particular attention is given to the innovative solutions of the industry and business problems. Each paper contains a well-developed computational experiment including analysis and comparison aspects. The Editorial Board received 52 submissions and carefully selected 11 papers out them. The paper acceptance rate was 25%. Each submitted manuscript was examined whether it is prepared according to the guidelines and fits the scope of the conference. All the submitted manuscripts were subjected to the single-blind peer review process, participated at least by three independent reviewers, who remain anonymous throughout the process. The reviewers were matched to the papers according to their expertise. No conflicts of interest were allowed. The reviewers were also asked to evaluate whether the manuscript is original and of sufficient weight and interest.

This event would not have been possible without the participation of many people. We would like to thank everyone who contributed to the success of the conference. We would like especially acknowledge the dedicated contribution of members of the Organizing Committee. The conference was organized and supported by the Russian Foundation for Basic Research, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, Moscow Institute of Physics and

Technology, Forecsys, Center of Forecasting Systems and Recognition, and Harbour. Space University Barcelona. We are grateful for the timely commitment of the Program Russian foundation for basic research and Reviewing Committee and all the external reviewers.

October 2016

Vadim V. Strijov
Dmitry Ignatov
Konstantin V. Vorontsov

Organization

Program Chairs

Konstantin Rudakov	Dorodnicyn Computing Centre of RAS, Moscow, Russia
Denis Zorin	New York University's Courant Institute of Mathematical Sciences, USA

Program Committee Secretary

Vadim V. Strijov	Moscow Institute of Physics and Technology, Russia
------------------	--

Program Committee

Mikhail Alexandrov	Autonomous University of Barcelona, Spain
Lorraine Baque	Autonomous University of Barcelona, Spain
Angels Catena	Autonomous University of Barcelona, Spain
Alexander Frey	Norment part University of Oslo, Finland
Ekaterrina Krymova	Universität Duisburg-Essen, Germany
Yuri Maximov	Los Alamos National Laboratory, USA
Konstantin Mertsalov	Harbour Space University Barcelona, Spain
Grigori Sidorov	National Polytechnic Institute of Mexico, Mexico
Roman Sologub	OBR Investments Limited, Cyprus
Iliya Tolstikhin	Max Planck Institute for Intelligent Systems, Germany
Liping Wang	Nanjing University of Aeronautics and Astronautics, China
Sergey Dvoenko	Tula State University, Russia
Edward Gimadi	Sobolev Institute of Mathematics of Siberian Branch of RAS (Novosibirsk) and Novosibirsk State University, Russia
Aleksandr Gornov	Institute of System Dynamics and Control Theory of Siberian Branch of RAS (Irkutsk), Russia
Olga Gromova	Ivanovo State Medical Academy, Russia
Aleksandr Kelmanov	Sobolev Institute of Mathematics of Siberian Branch of RAS (Novosibirsk), Russia
Mikhail Khachay	Yeltsin Ural Federal University and Institute of Mathematics and Mechanics of Ural Branch of RAS (Yekaterinburg), Russia
Leonid Mestetskiiy	Lomonosov Moscow State University and Moscow Institute of Physics and Technology, Russia
Vadim Mottl	Dorodnicyn Computing Centre of RAS (Moscow) and Tula State University, Russia

Aleksandr Shanenin	Moscow Institute of Physics and Technology, Russia
Viktor Soyfer	Samara State Aerospace University, Russia
Mikhail Ustinin	Institute of Mathematical Problems of Biology of RAS (Pushchino), Russia
Konstantin V. Vorontsov	Yandex School of Data Analysis and Dorodnicyn Computing Centre of RAS (Moscow), Russia

Reviewers

Aduenko Alexander	Loukachevitch Natalia
Bakhteev Oleg	Maksimov Yuri
Bunakova Vlada	Matrosov Mikhail
Chicheva Marina	Matveev Ivan
Chulichkov Alexey	Maysuradze Artchil
Dudarenko Marina	Mestetskiy Leonid
Dvoenko Sergey	Mikheeva Anna
D'yakonov Alexander	Mirkin Boris
Fedoryaka Dmitry	Mnukhin Valeriy
Frei Alexander	Motrenko Anastasia
Gasnikov Alexander	Murashov Dmirty
Genrikhov Ivan	Nedel'ko Victor
Gneushev Alexander	Neychev Radoslav
Goncharov Alexey	Odinokikh Gleb
Graboviy Andriy	Panov Alexander
Ignat'ev Vladimir	Panov Maxim
Ignatov Dmitry	Potapenko Anna
Inyakin Andrey	Pushnyakov Alexey
Isachenko Roman	Reyer Ivan
Ishkina Shaura	Riabenko Evgeniy
Ivakhnenko Andrey	Sen'ko Oleg
Karasikov Mikhail	Seredin Oleg
Karkishchenko Alexander	Stenina Maria
Katrutsa Alexandr	Strijov Vadim V.
Khachay Mikhail	Sulimova Valentina
Khritankov Anton	Torshin Ivan
Kochetov Yury	Trekin Anton
Kopylov Andrey	Turdakov Denis
Krasotkina Olga	Vetrov Dmitry
Krymova Ekaterina	Vladimirova Maria
Kulunchakov Andrey	Vorontsov Konstantin V.
Kushnir Olesya	Yanina Anastasia
Kuzmin Arseniy	Zajtsev Alexey
Kuznetsov Mikhail	Zharikov Ilya
Kuznetsova Rita	Zhivotovskiy Nikita
Lange Mikhail	

Organising Committee Chairs

Yury Zhuravlev	Dorodnicyn Computing Centre of Russian Academy of Sciences, Moscow, Russia
Svetlana Velikanova	Harbour.Space University, Barcelona, Spain

Organising Committee Secretary

Yury Chehovich	Moscow Institute of Physics and Technology, Russia
----------------	--

Organising Committee

Tatyana Borisova	Center of Forecasting Systems and Recognition, Moscow, Russia
Andrey Gromov	Dorodnicyn Computing Centre of RAS, Moscow, Russia
Andrey Inyakin	Dorodnicyn Computing Centre of RAS, Moscow, Russia
Shaura Ishkina	Moscow Institute of Physics and Technology, Russia
Andrey Ivakhnenko	Antiplagiat, Moscow, Russia
Evgeniya Pomazkova	Moscow Institute of Physics and Technology, Russia
Lina Romashkova	Harbour.Space University, Barcelona, Spain
Aleksandr Tatarchuk	Forecsys, Moscow, Russia

Technical Editor

Shaura Ishkina	Moscow Institute of Physics and Technology, Russia
----------------	--

Sponsors

Russian Foundation for Basic Research
Moscow Institute of Physics and Technology (State University)
Forecsys
Center of Forecasting and Recognition Systems

Invited Talks and Tutorials

Robust Principal Component Analysis

Boris Polyak

Institute for Control Sciences Russian Academy of Sciences, Russia
boris@ipu.ru

Abstract. The main trend of modern data analysis is to reduce huge databases to their low-dimensional approximations. Classical tool for this purpose is Principal Component Analysis (PCA). However, it is sensitive to outliers and other deviations from standard assumptions. There are numerous approaches to robust PCA. We propose two novel models. One is based on minimization of Huber-like distances from low-dimensional subspaces. A simple method for this nonconvex matrix optimization problem is proposed. The second is the robust version of maximum likelihood method for covariance and location estimation for contaminated multivariate Gaussian distribution; again we arrive at non-convex vector-matrix optimization. Both methods are based on Reweighted Least Squares Approximations. They demonstrated fast convergence in simulations, however, statistical validation, as well as convergence behavior of both approaches, remain open problems.

Keywords: Low-dimensional approximations • Principal Component Analysis

Mining Intelligent Information in Twitter: Detecting Irony and Sarcasm

Paolo Rosso

Universitat Politècnica de València, Spain
proso@dsic.upv.es

Abstract. There is a growing interest from the research community in investigating the impact of irony and sarcasm on sentiment analysis in social media. A sentiment analysis tool often processes information wrongly when dealing with ironic or sarcastic opinions. In fact, what is literally said is usually negated, moreover, in absence of an explicit negation marker. A task has been organised in 2015 at SemEval on sentiment analysis of figurative language in Twitter. In this talk I will describe how irony and sarcasm are employed in tweets and reviews in general and what are the recent state-of-the-art attempts for their automatic detection, for instance processing, explicit or implicit, affective information in tweets.

Keywords: Natural language processing · Sentiment analysis · Irony · Sarcasm

Big Data Platform and Analytics at Snapchat

Sina Sohangir

Snapchat, USA

sohangir@yahoo.com

Abstract. In this talk, I will go through our data platform at Snapchat which is based on Google Cloud. Then I present the type of problems we are trying to solve for ad targeting in general and for Snapchat in particular. I would look into Snapchat's social network and big data problems and solutions dealing with this specific social network.

Keywords: Snapchat · Ad targeting · Big data

Detecting Anomalies in the Real World

Alessandra Stagliano

The ASI, UK

`alessandra.stagliano@unige.it`

Abstract. Anomaly detection is a data science technique is used in many applications, from the IoT to finance. With the rise of the industrial Internet and the explosion of sensor data, businesses from transport to manufacturing are keen to develop predictive maintenance. Another key area where anomaly detection is important is identifying fraud in finance and within the social benefit system.

There are, however, a number of challenges when applying anomaly detection that are hindering progress. For a start, anomaly detection is a challenging problem by definition: defining and distinguishing between “normal” and an “anomaly” is often part of the problem statement. An anomaly is a relatively rare event and, hence, suffers from the accuracy paradox. Moreover, what is a good measure of success? Because of the nature of the problem, if the model misses all the anomalies, it will still be very accurate. The vastly different data types and preprocessing required, as well as the complex ensemble machine-learning methods needed, prove an additional challenge.

We illustrate these challenges and explain how to overcome them, with a practical application to credit card fraud detection.

Keywords: Anomaly detection · Fraud detection

Additively Regularized Topic Modeling for Searching Ethnical Discourse in Social Media

Konstantin V. Vorontsov

Yandex School of Data Analysis, Moscow Institute of Physics and Technology,
Lomonosov Moscow State University, Computing Centre of RAS, Russia
vokov@forecsys.ru

Abstract. Recently, social studies of the Internet have started to adopt various techniques of large-scale text mining for a variety of goals. One of such goals is the unsupervised discovery of topics related to ethnicity for early detection of ethnic conflicts emerging in social media. Probabilistic topic modeling used for such goals usually employs Bayesian inference for one of the numerous extensions of the Latent Dirichlet Allocation (LDA) model that has been widely popular over the last decade. However, recent research suggests that a non-Bayesian approach of additive regularization of topic models (ARTM) results in more control over the topics purity and interpretability, more flexibility for combining topic models, and faster inference.

In this work, we apply ARTM framework and BigARTM open-source software to a case study of mining ethnic content from the Russian-language blogosphere. We introduce a problem-specific combination of regularizers for ARTM and compare ARTM with LDA. The most important regularizer uses a vocabulary of a few hundred ethnonyms as seed words for ethnic-related topics. We conclude that ARTM is better suitable for mining rare topics, such as those on ethnicity, since it obviously finds larger numbers of relevant topics of higher or comparable quality.

Keywords: Topic modeling · Big ARTM · Ethnic discourse ·
Additive regularization

Image Synthesis with Deep Neural Networks

Victor Lempitsky

Skolkovo Institute of Science and Technology, Russia
lempitsky@skoltech.ru

Abstract. Using deep convolutional networks for pattern recognition in images has by now become a mature and well-known technology. More recently, there is a growing interest in using convolutional networks in a “reverse” mode, i.e. to synthesize images with certain properties rather than to recognize image content. In the talk, I will present several algorithmic results and application examples obtained for this very promising direction of research.

Keywords: Deep Learning · Convolutional networks · Image synthesis

Labelling Images Using Transfer Learning: An Application to Recommender Systems

Yannis Ghazouani

Dataiku, France

Abstract. Dataiku recently worked on an e-business vacation retailer recommender system. We created a meta model on top of classical recommender systems to optimise directly the probability of purchase. Along sales, clicks or descriptions of products data, we added image information in the recommender thanks to Deep Learning image recognition models. Since most companies cannot afford training neural networks, we followed a transfer learning approach.

The model we selected has been trained on the Places205 dataset and is based on the VGG16 architecture. We extracted last convolutional layer features from the mentioned VGG16 network and build a 2-layer TensorFlow model to learn the SUN397 labels from this intermediate features.

We then applied a Non-Negative Matrix Factorization (scikit-learn implementation) to reduce the dimension of the images representation space from around 600 categories to 30 variables and extracted some relevant topics from the images. Content-based image features were finally added to the meta model.

Keywords: Image labeling · Transfer learning · Feature extraction · Recommender systems

Contents

Machine Learning Theory with Applications

Overlapping Community Detection in Weighted Graphs: Matrix Factorization Approach	3
<i>Konstantin Slavnov and Maxim Panov</i>	
Reinforcement-Based Simultaneous Algorithm and Its Hyperparameters Selection.	15
<i>Valeria Efimova, Andrey Filchenkov, and Anatoly Shalyto</i>	
An Information Approach to Accuracy Comparison for Classification Schemes in an Ensemble of Data Sources	28
<i>Mikhail Lange, Sergey Ganebnykh, and Andrey Lange</i>	
On a Metric Kemeny's Median.	44
<i>Sergey Dvoenko and Denis Pshenichny</i>	

Intelligent Data Processing in Life and Social Sciences

Recognition of Herpes Viruses on the Basis of a New Metric for Protein Sequences	61
<i>Valentina Sulimova, Oleg Seredin, and Vadim Mottl</i>	
Learning Interpretable Prefix-Based Patterns from Demographic Sequences . . .	74
<i>Danil Gizdatullin, Jaume Baixeries, Dmitry I. Ignatov, Ekaterina Mitrofanova, Anna Muratova, and Thomas H. Espy</i>	
Population Health Assessment Based on Entropy Modeling of Multidimensional Stochastic Systems.	92
<i>Alexander N. Tyrsin and Garnik G. Gevorgyan</i>	
Students Learning Results Prediction with Usage of Mixed Diagnostic Tests and 2-Simplex Prism.	106
<i>Anna Yankovskaya, Yury Dementyev, Artem Yamshanov, and Danil Lyapunov</i>	

Morphological and Technological Approaches to Image Analysis

Application of Information Redundancy Measure To Image Segmentation . . .	125
<i>Dmitry Murashov</i>	

Eyelid Position Detection Method for Mobile Iris Recognition 140
*Gleb Odinokikh, Mikhail Korobkin, Vitaly Gnatyuk,
and Vladimir Ereemeev*

**Parametric Shape Descriptor Based on a Scalable
Boundary-Skeleton Model 151**
Ivan Reyer and Ksenia Aminova

Author Index 163