Data Science – Analytics and Applications

Peter Haber · Thomas Lampoltshammer · Manfred Mayr
(Eds.)

Data Science – Analytics and Applications

Proceedings of the 1st International Data Science Conference – iDSC2017

Springer Vieweg

Editors
Peter Haber
Informationstechnik & System-Management
Fachhochschule Salzburg Puch/Salzburg, Österreich

Manfred Mayr
Informationstechnik & System-Management
Fachhochschule Salzburg Puch/Salzburg, Österreich

Thomas Lampoltshammer
Department für E-Governance in Wirtschaft und Verwaltung
Donau-Universität Krems, Krems an der Donau / Österreich

ISBN 978-3-658-19286-0 **ISBN 978-3-658-19287-7 (eBook)**
https://doi.org/10.1007/978-3-658-19287-7

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

Springer Vieweg

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Vieweg ist Teil von Springer Nature
Die eingetragene Gesellschaft ist Springer Fachmedien Wiesbaden GmbH
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Preface

It is with deep satisfaction that we write this foreword for the Proceedings of the 1st International Data Science Conference (iDSC) held in Salzburg, Austria, June 12th - 13th 2017. The conference program and the resulting proceedings represent the efforts of many people. We want to express our gratitude towards the members of our program committee as well as towards our external reviewers for their hard work during the reviewing process.

iDSC proofed itself as an innovative conference, which gave its participants the opportunity to delve into state-of-the-art research and best practice in the fields of Data Science and data-driven business concepts. Our research track offered a series of presentations by Data Science researchers regarding their current work in the fields of Data Mining, Machine Learning, Data Management, and the entire spectrum of Data Science.

In our industry track, practitioners demonstrated showcases of data-driven business concepts and how they use Data Science to achieve organisational goals, with a focus on manufacturing, retail, and financial services. Within each of these areas, experts described their experience, demonstrated their practical solutions, and provided an outlook into the future of Data Science in the business domain.

Besides these two parallel tracks, a European symposium on Text and Data Mining has been integrated into the conference. This symposium highlighted the EU project FutureTDM, granting insights into the future of Text and Data Mining, and introducing overarching policy recommendations and sector-specific guidelines to help stakeholders overcome the legal and technical barriers, as well the lack of skills that have been identified.

Our sponsors had their own, special platform via workshops to provide hands-on interaction with tools or to learn approaches towards concrete solutions. In addition, an exhibition of products and services offered by our sponsors took place throughout the conference, with the opportunity for our participants to seek contact and advice.

Completing the picture of our program, we proudly presented keynote presentations from leaders in Data Science and data-driven business, both researchers and practitioners. These keynotes provided all participants the opportunity to come together and shared views on challenges and trends in Data Science.

In addition to the contributed papers, five invited keynote presentations were given by: Euro Beinat (CS Research, Salzburg University), Mario Meir-Huber (Microsoft Austria), Mike Olson (Cloudera), Ralf Klinkenberg (RapidMiner) and Janek Strycharz (Digital Center Poland). We thank the invited speakers for sharing their insights with our community.

The conference chair John Thompson has also helped us in many ways setting up the industry track, for which we are grateful. We would especially like to thank our two colleagues, Astrid Karnutsch and Maximilian Tschuchnig, for their enormous and constructive commitment to organizing and conducting the conference. The paper submission and reviewing process was managed using the EasyChair system.

These proceedings will provide scientists and practitioners with an excellent reference to current activities in the Data Science domain. We trust also that this will be an impetus to stimulate further studies, research activities and applications in all discussed areas ensured by the support of our publisher Springer / Vieweg Wiesbaden Germany.

Finally, again, the conference would not be possible without the excellent papers contributed by our authors. We thank them for their contributions and their participation at iDSC´17.

Peter Haber, Thomas Lampoltshammer and Manfred Mayr

Conference Chairs

Future TDM Symposium Recap

FutureTDM is a european project focusing on reducing barriers and increasing uptake of Text and Data Mining (TDM) for research environments in Europe. The outcomes of the project were presented in the Symposium which has also served to connect key actors and interest groups and promote open dialogue via discussion panels and informal workshops. The FTDM Symposium was scheduled alongside iDSC 2017, given that both events address similar target groups and share a common perspective: they both aimed at creating a communication network among the members of the TDM community, where experts can exchange ideas and share the most up-to-date research results, as well as legal and industrial advances relevant to TDM. The audience targeted by the iDSC conference was the broad community of researchers and industry practitioners as well as other practitioners and stakeholders, making it ideal for disseminating the project's results.

The project's objective has been to detect the barriers to TDM, reveal best practices and put together sets of recommendations for TDM practitioners through a collaborative knowledge and open information approach. The barriers recorded were grouped around four pillars: a) **legal**, b) **economic,** c) **skills**, d) **technical**. These categories emerged after discussions with respective stakeholders such as researchers, developers, publishers and SMEs during Knowledge Cafés run across Europe (the Netherlands, the United Kingdom, Italy, Slovenia, Germany, Poland etc) and two workshops held in Brussels[1] (on September, 27th 2016 and March, 29th 2017).

The Symposium[2] was a chance to invite experts from all over Europe to share their experience and expertise in different domains. It was also a great opportunity to announce the guidelines and recommendations formulated in order to increase TDM uptake. It started with a brief introduction by Bernhard Jäger (SYNYO)[3] underlying the need to bring together different groups of stakeholders, such as policy makers and legislators, developers and users who would benefit from the project's findings and the respective recommendations formed by the FTDM working groups. It continued with a keynote speech by Janek Strycharz (Projekt Polska Foundation) dedicated to the Economic Potential of Data Analytics. Janek Strycharz elaborated on different types of Big Data and the variety of possibilities they offer and explained how that at a global and european scale there could be a benefit from Big Data and TDM (the European GDP alone would be increased by USD 200 billion).

[1] FutureTDM Workshop I and II outcomes can be found at http://www.futuretdm.eu/knowledge-cafes/futuretdm-workshop/
http://www.futuretdm.eu/knowledge-cafes/futuretdm-workshop-2/

[2] All presentation slides are available online at www.slideshare.net/FutureTDM/presentations

[3] Presentation on Introduction to the FutureTDM project is available at https://www.slideshare.net/FutureTDM/introduction-to-the-future-tdm-project

The first session entitled "**Data Analytics and the Legal Landscape: Intellectual Property and Data Protection"** included Freyja van den Boom, researcher from Open Knowledge International/Content Mine who presented the legal barriers identified and the respective recommendations created under the subject "Dealing with the legal bumps on the road to further TDM uptake". The focus of the presentation was on the principles identified to counterbalance barriers: Awareness and Clarity, TDM Without Boundaries, and Equitable Access. The session was chaired by Ben White (Head of Intellectual Property at the British Library) and included the following panelists: i) Duncan Campbell (John Wiley & Sons, Inc.), representing the publisher's perspective, ii) Prodromos Tsiavos (Onassis Cultural Centre/IP Advisor), providing an organization's point of view, iii) Marie Timmermann (Science Europe), offering her point of view as the EU Legislation and Regulatory Affairs Officer and iv) Romy Sigl (AustrianStartups) sharing her experience from startUps. The discussion revolved around regulations which must address the implementation of the law and its exceptions, copyright issues, the distinction between commercial and noncommercial activities, the need for better communication between different groups of stakeholders and the importance and value of TDM for publishers.

During the following session the projects ContentMine (Stefan Kasberger), PLAZI (Donat Agosti), CORE (Petr Knoth), RapidMiner (Ralf Klinkenberg), clarin:el (Maria Gavrilidou) and ALCIDE (Alessio Palmero Aprosio) were introduced and the presenters were accessible for a more detailed presentation of their work to the attendees who would be interested in learning more. The researchers shared their experience on technical and legal problems they have encountered demonstrating the TDM applications and infrastructures they had created.

The next session offered an **overview of FTDM case studies from Startups to Multinationals**. The presentation entitled "Stakeholder consultations - The Highlights" was given by Freyja van den Boom (Open Knowledge International/Content Mine) who talked about the findings from continuous stakeholder consultations throughout the project. The session was chaired by Maria Eskevich (Radboud University) and included as panelists Donat Agosti (PLAZI), Petr Knoth (CORE), Kim Nilsson (PIVIGO), and Peter Murray-Rust (ContentMine). The issues raised during discussion pinpointed the need for realistic solutions to infrastructures, community engagement, and open source and data.

Kiera McNeice (British Library) was the presenter in the fourth session and her presentation was entitled "Supporting TDM in the Education Sector". The session focusing on "**Universities, TDM and the need for strategic thinking on educating researchers**" was chaired by Ben White (Head of Intellectual Property at the British Library) and panelists Claire Sewell (Cambridge University Library), Jonas Holm (Stockholm University Library), and Kim Nilsson (PIVIGO). The discussion which followed touched upon issues such as the future of Data Science and the nature of Data Scientists. Some of the key concepts which were discussed were that of inclusion and diversity, gender imbalance and nationality characteristics, which all affect access to Data Science and the ability to become a Data Scientist. Concerns were expressed as to whether anyone could become a Data Scientist, and whether the focus should be on becoming a Data Scientist or a more efficient TDM user.

The challenges and solutions regarding technologies and infrastructures supporting Text and Data Analytics was the topic of the fifth session, the main presenter of which was Maria Eskevich (Radboud University). She focused on "The TDM Landscape: Infrastructure and Technical Implementation" and touched upon the business and scientific perspectives on TDM by showing the investment made by the EU in the five economic sectors. She also talked about the barriers/challenges encountered in terms of accessibility and interoperability of infrastructures, sustainability of data and digital readiness of language resources. The following discussion, chaired by Stelios Piperidis (ARC) with Mihai Lupu (Data Market Austria), Maria Gavrilidou (clarin: el) and Nelson Silva (know-centre) revolved around real TDM problems and the solutions the researchers came up with and close with the requirements of an effective TDM infrastructure.

The final session of the Symposium was dedicated to the **Next Steps: A Roadmap to promoting greater uptake of Data Analytics in Europe**. A presentation was made by Kiera McNeice (British Library) who briefly summarised what the project has achieved so far and focussed on the key principles from the FutureTDM Policy Framework[4] which must underlie all the efforts to be made in the future in Legal Policies, Skills and Education, Economy and Incentives and Technical and Infrastructure.

The Symposium close with a presentation of Bernhard Jäger and Burcu Akinci (SYNYO) of the FutureTDM platform (http://www.futuretdm.eu/), which is populated with the project outcomes and findings.The platform will continue to exist after the end of the project and will be continuously revised and updated in order to maintain a coherent and up-to-date view on the TDM landscape open to the public.

Kornella Pouli

Athena RIC/ILSP, Athens

Burcu Akinci

SYNYO GmbH, Vienna

[4] http://www.futuretdm.eu/policy-framework/

Organisation

Organising Institutions

Salzburg University of Applied Sciences
Information Professionals GmbH

Conference Chairs

Peter Haber	Salzburg University of Applied Sciences
Thomas J. Lampoltshammer	Danube University Krems
Manfred Mayr	Salzburg University of Applied Sciences
John A. Thompson	Information Professionals GmbH

Organising Committee

Peter Haber	Salzburg University of Applied Sciences
Astrid Karnutsch	Salzburg University of Applied Sciences
Thomas J. Lampoltshammer	Danube University Krems
Manfred Mayr	Salzburg University of Applied Sciences
John A. Thompson	Information Professionals GmbH
Susanne Schnitzer	Information Professionals GmbH
Maximilian E. Tschuchnig	Salzburg University of Applied Sciences

Program Committee

David C. Anastasiu	San Jose State University
Vera Andrejcenko	University of Antwerp
Christian Bauckhage	University of Bonn
Markus Breunig	Rosenheim University of Applied Sciences
Stefanie Cox	IT Innovation Centre
Werner Dubitzky	University of Ulster, Coleraine
Günther Eibl	Salzburg University of Applied Sciences
Süleyman Eken	University Kocaeli
Karl Entacher	Salzburg University of Applied Sciences
Edison Pignaton de Freitas	Federal University of Rio Grande do Sul
Bernhard Geissler	Danube University Krems
Charlotte Gerritsen	Netherlands Institute for the Study of Crime and Law Enforcement (NSCR)
Mohammad Ghoniem	Luxembourg Institute of Science and Technology
Peter Haber	Salzburg University of Applied Sciences
Johann Höchtl	Danube University Krems
Martin Kaltenböck	Semantic Web Company
Astrid Karnutsch	Salzburg University of Applied Sciences
Elmar Kiesling	Vienna University of Technology
Robert Krimmer	University of Tallinn
Peer Kröger	Ludwig-Maximilians-Universität München
Thomas J. Lampoltshammer	Danube University Krems

Michael Leitner	Louisiana State University
Giuseppe Manco	University of Calabria
Manfred Mayr	Salzburg University of Applied Sciences
Mark-David McLaughlin	Bentley University
Robert Merz	Salzburg University of Applied Sciences
Elena Lloret Pastor	University of Alicante
Cody Ryan Peeples	Cisco
Gabriela Viale Pereira	Fundação Getúlio Vargas – EAESP
Peter Ranacher	University of Zurich
Siegfried Reich	Salzburg Research Forschungsgesellschaft mbH
Eric Rozier	Iowa State University
Johannes Scholz	Graz University of Technology
Maximilian E. Tschuchnig	Salzburg University of Applied Sciences
Jürgen Umbrich	Vienna University of Economics and Business
Andreas Unterweger	Salzburg University of Applied Sciences
Eveline Wandl-Vogt	Austrian Academy of Sciences
Stefan Wegenkittl	Salzburg University of Applied Sciences
Stefanie Wiegand	IT Innovation Centre / University of Southampton
Peter Wild	Austrian Institute of Technology
Radboud Winkels	University of Amsterdam
Anneke Zuiderwijk - van Eijk	Delft University of Technology

Reviewer

David C. Anastasiu	San Jose State University
Christian Bauckhage	University of Bonn
Markus Breunig	Rosenheim University of Applied Sciences
Cornelia Ferner	Salzburg University of Applied Sciences
Werner Dubitzky	University of Ulster, Coleraine
Günther Eibl	Salzburg University of Applied Sciences
Karl Entacher	Salzburg University of Applied Sciences
Bernhard Geissler	Danube University Krems Höchtl
Martin Kaltenböck	Semantic Web Company
Peer Kröger	Ludwig-Maximilians-Universität München
Thomas J. Lampoltshammer	Danube University Krems
Michael Leitner	Louisiana State University
Elena Lloret Pastor	University of Alicante
Manfred Mayr	Salzburg University of Applied Sciences
Robert Merz	Salzburg University of Applied Sciences
Edison Pignaton de Freitas	Federal University of Rio Grande do Sul
Siegfried Reich	Salzburg Research Forschungsgesellschaft mbH
Eric Rozier	Iowa State University
Johannes Scholz	Graz University of Technology
Maximilian E. Tschuchnig	Salzburg University of Applied Sciences
Jürgen Umbrich	Vienna University of Economics and Business
Andreas Unterweger	Salzburg University of Applied Sciences
Stefan Wegenkittl	Salzburg University of Applied Sciences

Sponsors of the conference

Platinum Sponsors

Cloudera GmbH
Apache Hadoop-based software, support and services, and training

www.cloudera.com

Silver Sponsors

The unbelievable Machine Company GmbH
Full-service provider for Big Data, cloud services & hosting

www.unbelievable-machine.com

F&F GmbH
IT consulting, solutions and Big Data Analytics

www.ff-muenchen.de

The MathWorks GmbH
Mathematical computing software

www.mathworks.com

RapidMiner GmbH
Data science software platform for data preparation, machine learning, deep learning, text mining, and predictive analytics

www.rapidminer.com

ITG: innovative consulting and location development
ITG is Salzburg's innovation centre

www.itg-salzburg.at

Table of Content

German Abstracts

Reasoning and Predictive Analytics

Circadian Cycles and Work Under Pressure: A Stochastic Process Model for E-learning Population Dynamics

Internetanalysetechniken, konzipiert zur Quantifizierung von Internetnutzungsmustern, erlauben ein tieferes Verständnis menschlichen Verhaltens. Neueste Modelle menschlicher Verhaltensdynamiken haben gezeigt, dass im Gegensatz zu zufällig verteilten Ereignissen, Menschen Tätigkeiten ausüben, die schubweises Verhalten aufweisen. Besonders die Teilnahme an Internetkursen zeigt häufig Zeiträume von Inaktivität und Prokrastination gefolgt von häufigen Besuchen kurz vor den Prüfungen. Hier empfehlen wir ein stochastisches Prozessmodell, welches solche Muster kennzeichnet und Tagesrhythmen menschlicher Aktivitäten einbezieht. Wir bewerten unser Modell anhand von realen Daten, die während einer Zeitspanne von zwei Jahren auf einer Plattform für Universitätskurse gesammelt wurden. Anschließend schlagen wir ein dynamisches Modell vor, welches sowohl Prokrastinationszeiträume als auch Zeiträume des Arbeitens unter Zeitdruck berücksichtigt. Da Tagesrhythmen und Prokrastination-Druck-Kreisläufe wesentlich für menschliches Verhalten sind, kann unsere Methode auf andere Tätigkeiten ausgeweitet werden, wie zum Beispiel die Auswertung von Surfgewohnheiten und Kaufverhalten von Kunden.

Investigating and Forecasting User Activities in Newsblogs: A Study of Seasonality, Volatility and Attention Burst

Das Studium allgemeiner Aufmerksamkeit ist ein Hauptthemengebiet im Bereich der Internetwissenschaft, da wir wissen wollen, wie die Beliebtheit eines bestimmten Nachrichtenthemas oder Memes im Laufe der Zeit zu- oder abnimmt. Neueste Forschungen konzentrierten sich auf die Entwicklung von Methoden zur Quantifizierung von Erfolg und Beliebtheit von Themen und untersuchten ihre Dynamiken im Laufe der Zeit. Allerdings wurde das gesamtheitliche Nutzerverhalten über Inhaltserstellungsplattformen größtenteils ignoriert, obwohl die Beliebtheit von Nachrichtenartikeln auch mit der Art verbunden ist, wie Nutzer Internetplattformen nutzen. In dieser Abhandlung zeigen wir ein neuartiges Framework, dass die Verlagerung der Aufmerksamkeit von Bevölkerungsgruppen in Hinblick auf Nachrichtenblogs untersucht. Wir konzentrieren uns auf das Kommentarverhalten von Nutzern bei Nachrichtenbeiträgen, was als Stellvertreter für die Aufmerksamkeit gegenüber Internetinhalten fungiert. Wir nutzen Methoden der Signalverarbeitung und Ökonometrie, um Verhaltensmuster von Nutzern aufzudecken, die es uns dann erlauben, das Verhalten einer Bevölkerungsgruppe zu simulieren und schlussendlich vorherzusagen, sobald eine Aufmerksamkeitsverlagerung auftritt. Nach der Untersuchung von Datenreihen von über 200 Blogs mit 14 Millionen Nachrichtenbeiträgen, haben wir zyklische Gesetzmäßigkeiten im Kommentarverhalten identifiziert: Aktivitätszyklen von 7 Tagen und 24 Tagen, die möglicherweise im Zusammenhang zu bekannten Dimensionen von Meme-Lebenszeiten stehen.

Knowledge-based Short-Term Load Forecasting for Maritime Container Terminals

Durch den Anstieg von Last- und Nachfragemanagement in modernen Energiesystemen erhält die Kurzzeitlastprognose für Industrieeinzelkunden immer mehr Aufmerksamkeit. Es scheint für Industriestandorte lohnenswert, dass Wissen zu geplanten Maßnahmen in den Lastprognoseprozess des nächsten Tages einzubeziehen. Im Fall eines Seecontainer-Terminals, basieren diese Betriebspläne auf der Liste ankommender und abfahrender Schiffe. In dieser Abhandlung werden zwei Ansätze vorgestellt, welche dieses Wissen auf verschiedene Weisen einbeziehen: Während fallbasiertes Schlussfolgern träge während des Prognoseprozesses lernt, müssen künstliche neurale Netzwerke erst trainiert werden, bevor ein Prognoseprozess durchgeführt werden kann. Es kann gezeigt werden, dass das Einbeziehung von mehr Wissen in den Prognoseprozess bessere Ergebnisse im Hinblick auf die Prognosegenauigkeit ermöglicht.

Data Analytics in Community Networks

Beyond Spectral Clustering: A Comparative Study of Community Detection for Document Clustering

Dokumenten-Clustering ist ein allgegenwärtiges Problem bei der Datengewinnung, da Textdaten eine der gebräuchlichsten Kommunikationsformen sind. Die Reichhaltigkeit der Daten erfordert Methoden, die – je nach den Eigenschaften der Informationen, die gewonnen werden sollen – auf verschiedene Aufgaben zugeschnitten sind. In letzter Zeit wurden graphenbasierte Methoden entwickelt, die es hierarchischen, unscharfen und nicht-gaußförmigen Dichtemerkmalen erlauben, Strukturen in komplizierten Datenreihen zu identifizieren. In dieser Abhandlung zeigen wir eine neue Methodologie für das Dokumenten-Clustering, das auf einem Graphen basiert, der durch ein Vektorraummodell definiert ist. Wir nutzen einen überlappenden hierarchischen Algorithmus und zeigen die Gleichwertigkeit unserer Qualitätsfunktion mit der von Ncut. Wir vergleichen unsere Methode mit spektralem Clustering und anderen graphenbasierten Modellen und stellen fest, dass unsere Methode eine gute und flexible Alternative für das Nachrichten-Clustering darstellt, wenn eingehende Details zwischen den Themen benötigt werden.

Third Party Effect: Community Based Spreading in Complex Networks

Ein wesentlicher Teil der Netzwerkforschung wurde dem Studium von Streuprozessen und Gemeinschaftserkennung gewidmet, ohne dabei die Rolle der Gemeinschaften bei den Merkmalen der Streuprozesse zu berücksichtigen. Hier verallgemeinern wir das SIR-Modell von Epidemien durch die Einführung einer Matrix von Gemeinschaftsansteckungsraten, um die heterogene Natur des Streuens zu erfassen, die durch die natürlichen Merkmale von Gemeinschaften definiert sind. Wir stellen fest, dass die Streufähigkeiten einer Gemeinschaft gegenüber einer anderen durch das interne Verhalten von Drittgemeinschaften beeinflusst wird. Unsere Ergebnisse bieten Einblicke in Systeme mit reichhaltigen Informationsstrukturen und in Populationen mit vielfältigen Immunreaktionen.

Cosine Approximate Nearest Neighbors

Kosinus-Ähnlichkeitsgraphenerstellung, oder All-Pairs-Ähnlichkeitssuche, ist ein wichtiger Systemkern vieler Methoden der Datengewinnung und des maschinellen Lernens. Die Graphenerstellung ist eine schwierige Aufgabe. Bis zu n2 Objektpaare sollten intuitiv verglichen werden, um das Problem für eine Reihe von n Objekten zu lösen. Für große Objektreihen wurden Näherungslösungen für dieses Problem vorgeschlagen, welche die Komplexität der Aufgabe thematisieren, indem die meisten, aber nicht unbedingt alle, nächsten Nachbarn abgefragt werden. Wir schlagen eine neue Näherungsgraphen-Erstellungsmethode vor, welche Eigenschaften der Objektvektoren kombiniert, um effektiv weniger Vergleichskandidaten auszuwählen, welche wahrscheinlich Nachbarn sind. Außerdem kombiniert unsere Methode Filterstrategien, welche vor kurzem entwickelt wurden, um Vergleichskandidaten, die nicht vielversprechend sind, schnell auszuschließen, was zu weniger allgemeinen Ähnlichkeitsberechnungen und erhöhter Effizienz führt. Wir vergleichen unsere Methode mit mehreren gängigen Annäherungs- und exakten Grundwerten von sechs Datensätzen aus der Praxis. Unsere Ergebnisse zeigen, dass unser Ansatz einen guten Kompromiss zwischen Effizienz und Effektivität darstellt, mit einer 35,81-fachen Effizienzsteigerung gegenüber der besten Alternative bei 0,9 Recall.

Data Analytics through Sentiment Analysis

Information Extraction Engine for Sentiment-Topic Matching in Product Intelligence Applications

Produktbewertungen sind eine wertvolle Informationsquelle sowohl für Unternehmen als auch für Kunden. Während Unternehmen diese Informationen dazu nutzen, ihre Produkte zu verbessern, benötigen Kunden sie als Unterstützung für die Entscheidungsfindung. Mit Bewertungen, Kommentaren und zusätzlichen Informationen versuchen viele Onlineshops potenzielle Kunden dazu zu animieren, auf ihrer Seite einzukaufen. Allerdings mangelt es aktuellen Online-Bewertungen an einer Kurzzusammenfassung, inwieweit bestimmte Produktbestandteile den Kundenwünschen entsprechen, wodurch der Produktvergleich erschwert wird. Daher haben wir ein Produktinformationswerkzeug entwickelt, dass gängige Technologien in einer Engine maschineller Sprachverarbeitung vereint. Die Engine ist in der Lage produktbezogene Online-Daten zu sammeln und zu sichern, Metadaten auszulesen und Meinungen. Die Engine wird auf technische Online-Produktbewertungen zur Stimmungsanalyse auf Bestandteilsebene angewendet. Der vollautomatisierte Prozess durchsucht das Internet nach Expertenbewertungen, die sich auf Produktbestandteile beziehen, und aggregiert die Stimmungswerte der Bewertungen.

Towards German Word Embeddings: A Use Case with Predictive Sentiment Analysis

Trotz des Forschungsbooms im Bereich Worteinbettungen und ihrer Textmininganwendungen der letzten Jahre, konzentriert sich der Großteil der Publikationen ausschließlich auf die englische Sprache. Außerdem ist die Hyperparameterabstimmung ein Prozess, der selten gut dokumentiert (speziell für nicht-englische Texte), jedoch sehr wichtig ist, um hochqualitative Wortwiedergaben zu erhalten. In dieser Arbeit zeigen wir, wie verschiedene Hyperparameterkombinationen Einfluss auf die resultierenden deutschen Wortvektoren haben und wie diese Wortwiedergaben Teil eines komplexeren Modells sein können. Im Einzelnen führen wir als erstes eine intrinsische Bewertung unserer deutschen Worteinbettungen durch, die später in einem vorausschauenden Stimmungsanalysemodell verwendet werden. Letzteres dient nicht nur einer intrinsischen Bewertung der deutschen Worteinbettungen, sondern zeigt außerdem, ob Kundenwünsche nur durch das Einbetten von Dokumenten vorhergesagt werden können.

User/Customer-centric Data Analytics

Feature Extraction and Large Activity-Set Recognition Using Mobile Phone Sensors

Diese Arbeit beschäftigt sich mit dem Problem der Aktivitätserkennung unter Verwendung von Daten, die vom Mobiltelefon des Benutzers erhoben wurden. Wir beginnen mit der Betrachtung und Bewertung der Beschränkungen der gängigen Aktivitätserkennungsansätze für Mobiltelefone. Danach stellen wir unseren Ansatz zur Erkennung einer großen Anzahl von Aktivitäten vor, welche die meisten Nutzeraktivitäten abdeckt. Außerdem werden verschiedene Umgebungen unterstützt, wie zum Beispiel zu Hause, auf Arbeit und unterwegs. Unser Ansatz empfiehlt ein einstufiges Klassifikationsmodell, dass die Aktivitäten genau klassifiziert, eine große Anzahl von Aktivitäten umfangreich abdeckt und in realen Umgebungen umsetzbar anzuwenden ist. In der Literatur gibt es keinen einzigen Ansatz, der alle drei Eigenschaften in sich vereint. In der Regel optimieren vorhandene Ansätze ihre Modelle entweder für einen oder maximal zwei der folgenden Eigenschaften: Genauigkeit, Umfang und Anwendbarkeit. Unsere Ergebnisse zeigen, dass unser Ansatz ausreichende Leistung im Hinblick auf Genauigkeit bei einem realistischen Datensatz erbringt, trotz deutlich erhöhter Aktivitätszahl im Vergleich zu gängigen Modellen, die auf Aktivitätserkennen basieren.

The Choice of Metric for Clustering of Electrical Power Distribution Consumers

Ein bedeutender Teil jedes Systemdatenmodells zur Energieverteilungsverwaltung ist ein Modell der Belastungsart. Eine Belastungsart stellt ein typisches Belastungsverhalten einer Gruppe gleicher Kunden dar, z. B. einer Gruppe von Haushalts-, Industrie- oder gewerblichen Kunden. Eine verbreitete Methode der Erstellung von Belastungsarten ist die Bündelung individueller Energieverbraucher auf der Basis ihres jährlichen Stromverbrauchs. Um ein zufriedenstellendes Maß an Belastungsartqualität zu erreichen, ist die Wahl des geeigneten Ähnlichkeitsmaßes zur Bündelung entscheidend. In dieser Abhandlung zeigen wir einen Vergleich verschiedener Metriken auf, die als Ähnlichkeitsmaß in unserem Prozess der Belastungsarterstellung eingesetzt werden. Zusätzlich zeigen wir eine neue Metrik, die auch im Vergleich enthalten ist. Die Metriken und die Qualität der damit erstellten Belastungsarten werden unter Verwendung von Realdatensätzen untersucht, die über intelligente Stromzähler des Verteilungsnetzes erhoben wurden.

Evolution of the Bitcoin Address Graph

Bitcoin ist eine dezentrale virtuelle Währung, die dafür genutzt werden kann, weltweit pseudo-anonymisierte Zahlungen innerhalt kurzer Zeit und mit vergleichsweise geringen Transaktionskosten auszuführen. In dieser Abhandlung zeigen wir die ersten Ergebnisse eine Langzeitstudie zur Bitcoinadressenkurve, die alle Adressen und Transaktionen seit dem Start von Bitcoin im Januar 2009 bis zum 31. August 2016 enthält. Unsere Untersuchung enthüllt eine stark verschobene Gradverteilung mit einer geringen Anzahl von Ausnahmen und zeigt, dass sich die gesamte Kurve stark ausdehnt. Außerdem zeigt sie die Macht der Adressbündelungsheuristik zur Identifikation von realen Akteuren, die es bevorzugen, Bitcoin für den Wertetransfer statt für die Wertespeicherung zu verwenden. Wir gehen davon aus, dass diese Abhandlung neue Einblicke in virtuelle Währungsökosysteme bietet und als Grundlage für das Design zukünftiger Untersuchungsmethoden und -infrastrukturen dienen kann.

Data Analytics in Industrial Application Scenarios

A Reference Architecture for Quality Improvement in Steel Production

Es gibt weltweit einen erhöhten Bedarf an Stahl, aber die Stahlherstellung ist ein enorm anspruchsvoller und kostenintensiver Prozess, bei dem gute Qualität schwer zu erreichen ist. Die Verbesserung der Qualität ist noch immer die größte Herausforderung, der sich die Stahlbranche gegenüber sieht. Das EU-Projekt PRESED (Predictive Sensor Data Mining for Product Quality Improvement) [Vorrausschauende Sensordatengewinnung zur Verbesserung der Produktqualität] stellt sich dieser Herausforderung durch die Fokussierung auf weitverbreitete, wiederkehrende Probleme. Die Vielfalt und Richtigkeit der Daten sowie die Veränderung der Eigenschaften des untersuchten Materials erschwert die Interpretation der Daten. In dieser Abhandlung stellen wir die Referenzarchitektur von PRESED vor, die speziell angefertigt wurde, um die zentralen Anliegen der Verwaltung und Operationalisierung von Daten zu thematisieren. Die Architektur kombiniert große und intelligente Datenkonzepte mit Datengewinnungsalgorithmen. Datenvorverarbeitung und vorausschauende Analyseaufgaben werden durch ein plastisches Datenmodell unterstützt. Der Ansatz erlaubt es den Nutzern, Prozesse zu gestalten und mehrere Algorithmen zu bewerten, die sich gezielt mit dem vorliegenden Problem befassen. Das Konzept umfasst die Sicherung und Nutzung vollständiger Produktionsdaten, anstatt sich auf aggregierte Werte zu verlassen. Erste Ergebnisse der Datenmodellierung zeigen, dass die detailgenaue Vorverarbeitung von Zeitreihendaten durch Merkmalserkennung und Prognosen im Vergleich zu traditionell verwendeter Aggregationsstatistik überlegene Erkenntnisse bietet.

Anomaly Detection and Structural Analysis in Industrial Production Environments

Das Erkennen von anormalem Verhalten kann im Kontext industrieller Anwendung von entscheidender Bedeutung sein. Während moderne Produktionsanlagen mit hochentwickelten Alarmsteuerungssytemen ausgestattet sind, reagieren diese hauptsächlich auf Einzelereignisse. Aufgrund der großen Anzahl und der verschiedenen Arten von Datenquellen ist ein einheitlicher Ansatz zur Anomalieerkennung nicht immer möglich. Eine weitverbreitete Datenart sind Logeinträge von Alarmmeldungen. Sie erlauben im Vergleich zu Sensorrohdaten einen höheren Abstraktionsgrad. In einem industriellen Produktionsszenario verwenden wir sequentielle Alarmdaten zur Anomalieerkennung und -auswertung, basierend auf erstrangigen Markov-Kettenmodellen. Wir umreißen hypothesegetriebene und beschreibungsorientierte Modellierungsoptionen. Außerdem stellen wir ein interaktives Dashboard zur Verfügung, um die Ergebnisse zu untersuchen und darzustellen.

Semantically Annotated Manufacturing Data to support Decision Making in Industry 4.0: A Use-Case Driven Approach

Intelligente Fertigung oder Industrie 4.0 ist ein Schlüsselkonzept, um die Produktivität und Qualität in industriellen Fertigungsunternehmen durch Automatisierung und datengetriebene Methoden zu erhöhen. Intelligente Fertigung nutzt Theorien cyber-physischer Systeme, dem Internet der Dinge sowie des Cloud-Computing. In dieser Abhandlung konzentrieren sich die Autoren auf Ontologie und (räumliche) Semantik, die als Technologie dienen, um semantische Kompatibilität der Fertigungsdaten sicherzustellen. Zusätzlich empfiehlt die Abhandlung, fertigungsrelevante Daten über die Einführung von Geografie und Semantik als Sortierformate zu strukturieren. Der in dieser Abhandlung verfolgte Ansatz sichert Fertigungsdaten verschiedener IT-Systeme in einer Graphdatenbank. Während des Datenintegrationsprozesses kommentiert das System systematisch die Daten – basierend auf einer Ontologie, die für diesen Zweck entwickelt wurde – und hängt räumliche Informationen an. Der in dieser Abhandlung vorgestellte Ansatz nutzt eine Analyse von Fertigungsdaten in Bezug auf Semantik und räumliche Abmessung. Die Methodologie wird auf zwei Anwendungsfälle für ein Halbleiterfertigungsunternehmen angewendet. Der erste Anwendungsfall behandelt die Datenanalyse zur Ereignisanalyse unter Verwendung von semantischen Ähnlichkeiten. Der zweite Anwendungsfall unterstützt die Entscheidungsfindung in der Fertigungsumgebung durch die Identifizierung potentieller Engpässe bei der Halbleiterfertigungslinie.