Data-Mining in sozialen Online-Netzwerken

Bianca Böckelmann, Robotron Datenbank-Software GmbH

Soziale Online-Netzwerke wie Facebook bieten Nutzern zwar die Möglichkeit, Profil-Informationen durch Einstellungen vor Unbefugten zu verbergen, jedoch stellt sich die Frage nach einem ausreichenden Schutz der Privatsphäre. Inwieweit lassen sich zum Beispiel in solchen Netzwerken mittels Data-Mining durch Verknüpfung von öffentlichen Informationen der Freunde Rückschlüsse auf die privaten Daten eines Nutzers ziehen, ohne dass diese Daten explizit offengelegt wurden?

Im Rahmen einer Masterarbeit wurde diese Frage im Jahr 2012 an der Brandenburgischen Technischen Universität Cottbus in Kooperation mit der Robotron Datenbank-Software GmbH und der IHP GmbH untersucht. Ziel war es, die Risiken aufzuzeigen, die sich durch die Bekanntgabe von Freundschaften ergeben. Ist es also nötig, neben den eigenen Profil-Informationen auch die Freundschaftsbeziehungen privat zu halten, um die eigene Privatsphäre zu schützen?

Dieser Artikel gibt zunächst einen Überblick über die Schritte des zugrunde gelegten „Knowledge Discovery in Databases"-Prozesses (KDD). Nach einer kurzen Einführung in Oracle Data Mining wird der Prozess praxisnah am Beispiel des Forschungsthemas beschrieben. Hierbei kam der Oracle Data Miner (ODMr) mit Daten freiwilliger Facebook-Nutzer zum Einsatz. Neben der Anwendung der eigentlichen Data-Mining-Algorithmen für die Prognose beinhaltet der KDD-Prozess die vorherige Vorverarbeitung und Transformation der Nutzerdaten, die zeitlich einen Großteil des Gesamtprozesses einnahmen. Abschließend werden die Ergebnisse vorgestellt, die die Vermutungen bekräftigen, dass auch die Betrachtung der Freundes-Informationen wichtig ist, um seine Privatsphäre zu schützen.

Der KDD-Prozess

Knowledge Discovery in Databases (KDD) ist der (semi-)automatische [1], nicht triviale Prozess des Identifizierens von gültigen, bisher unbekannten, potenziell nützlichen und verständlichen Mustern in Daten [2]. Der KDD Prozess besteht aus mehreren Phasen, die iterativ durchlaufen werden können (siehe Abbildung 1). Zu Beginn werden bei der Selektion das Ziel festgelegt sowie die Daten für die Wissensextraktion herangezogen, bezüglich ihrer Qualität beurteilt [1] und im nächsten Schritt für das Erzielen von Integrität, Konsistenz und Vollständigkeit vorverarbeitet. Anschließend folgt die Transformation der vorverarbeiteten Daten in eine für den Analyse-Schritt geeignete Repräsentation. Dazu zählt unter anderem die Auswahl relevanter Attribute.

In der eigentlichen Data-Mining-Phase werden die Algorithmen für die Extraktion von Mustern angewendet. Abschließend werden die erzielten Ergebnisse interpretiert und hinsichtlich des festgelegten Ziels bewertet. Sind die gefundenen Modelle gut, kann das hieraus gewonnene Wissen beispielsweise gezielt zum Ableiten von Handlungsdirektiven eingesetzt werden.

Oracle Data Mining

Oracle bietet mit Oracle Data Mining (ODM) als Teil der Oracle-Advanced-Analytics-Option die Möglichkeit, Da-

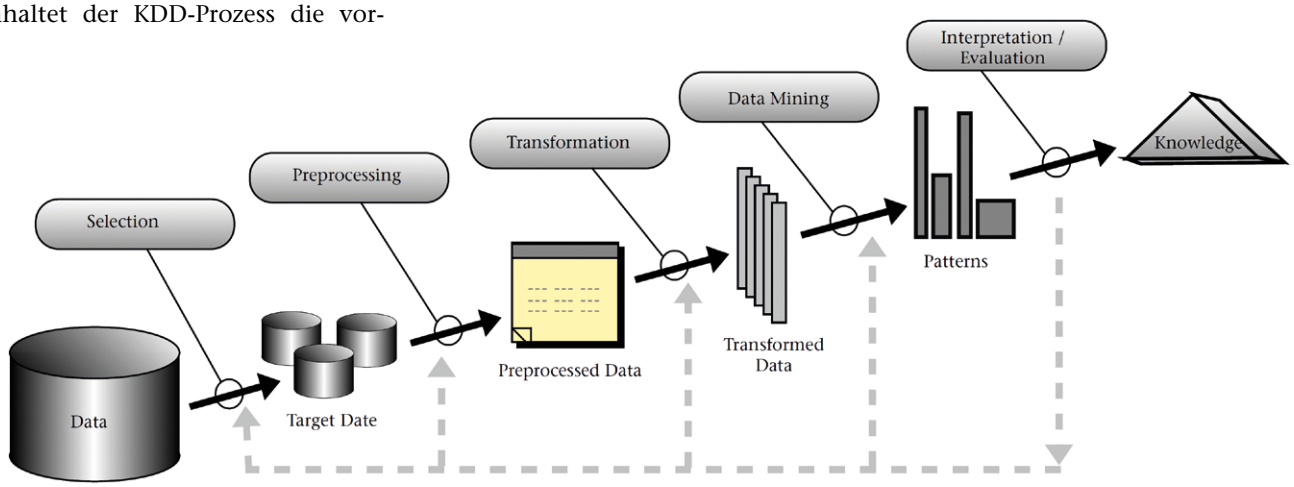

Abbildung 1: Schritte im KDD-Prozessmodell [3]

ta-Mining-Algorithmen direkt in der Oracle-Datenbank auszuführen. Für einen Zugriff auf die Quelldaten, die Modelle sowie die Analyse-Ergebnisse ist somit kein zusätzlicher Datentransport nötig. Die ODM-Funktionalitäten sind über PL/SQL, SQL und R im Zugriff [4]. Zudem steht mit dem Oracle Data Miner (ODMr) eine grafische Benutzeroberfläche bereit, die im Oracle SQL Developer eingebunden ist.

Mittels Workflows, die sich aus Knoten und Verbindungen zusammensetzen, lässt sich der gesamte KDD-Prozess abbilden. Abbildung 2 zeigt einen einfachen Workflow. Neben einfachen Statistiken, die einen Überblick über die Daten erlauben, können die Daten geeignet gefiltert, aggregiert und transformiert werden. Hierfür stehen beim ODMr verschiedene Knoten im Workflow-Editor zur Auswahl.

Grundlegend unterscheidet ODM drei Arten von Daten: numerische, kategorische und unstrukturierte. Die letzte Kategorie ist bei Texten relevant und erst in ODM 12.1 verfügbar [5, 6]. Anhand dieser Differenzierung werden die Daten in verschiedenen Data-Mining-Algorithmen anders behandelt. Eine Zuordnung findet entweder automatisch über den Datentyp oder manuell statt.

Für das Lernen von Data-Mining-Modellen unterstützt ODM die Klassifikation, die Regression, die Assoziations-Analyse, das Clustering, die Anomalie-Erkennung sowie die Feature-Extraktion und die Attribut-Wichtigkeit. Tabelle 1 gibt einen Überblick über die in ODM 12.1 zur Verfügung stehenden Data-Mining-Funktionen und zugehörigen Algorithmen. Die in blauer Schriftfarbe aufgeführten Algorithmen sind in der Version 12.1 neu hinzugekommen.

Da für die Masterarbeit ODM 11.2 zum Einsatz kam, wird im Folgenden vorwiegend Bezug auf diese Version genommen. Für eine Klassifikation können beispielsweise der Entscheidungsbaum, Naive Bayes, Support Vector Machine (SVM) sowie das generalisierte lineare Modell (GLM) in der logistischen Regression eingesetzt werden [4].

Bei ODM wird für das Anwenden der Algorithmen vorausgesetzt, dass alle Daten in genau einer Tabelle oder Sicht vorliegen, der sogenannten „Fall-Tabelle“. Einstellungen zu den Algorithmen können entweder standardmäßig belassen oder individuell zugeschnitten auf den jeweiligen Anwendungsfall vorgenommen werden. Für eine Bewertung der gelernten Modelle stehen in Abhängigkeit vom Verfahren diverse Evaluierungsmaße und Visualisierungen bereit. Beispielsweise kommt bei der Klassifikation die Gesamtklassifikationsgenauigkeit zum Einsatz. Diese ist der Anteil der korrekt klassifizierten Objekte in der gesamten Menge.

Indirekte Bestimmung privater Informationen

In der Masterarbeit wurde untersucht, ob es in sozialen Online-Netzwerken mittels Data-Mining möglich ist, nur anhand der öffentlichen Informationen der Freunde auf die privaten Daten einer Person selbst zu schließen. Müssen also neben den eigenen Profil-Informationen auch die Freundschafts-Beziehungen vor der breiten Öffentlichkeit privat gehalten werden, um personenbezogene Daten vor dem Ausspähen durch Fremde zu schützen?

Zu Beginn wurde dafür die Facebook-Anwendung „Hidden Profile“ in PHP entwickelt, um knapp 700 freiwillige Nutzer zu gewinnen. Neben Profil-Informationen wie Geburtstag, Geschlecht, Wohn- und Heimatort, „Gefällt mir“-Angaben (unter anderem TV, Film, Musik), Informationen zur Arbeit und Bildung wurden auch

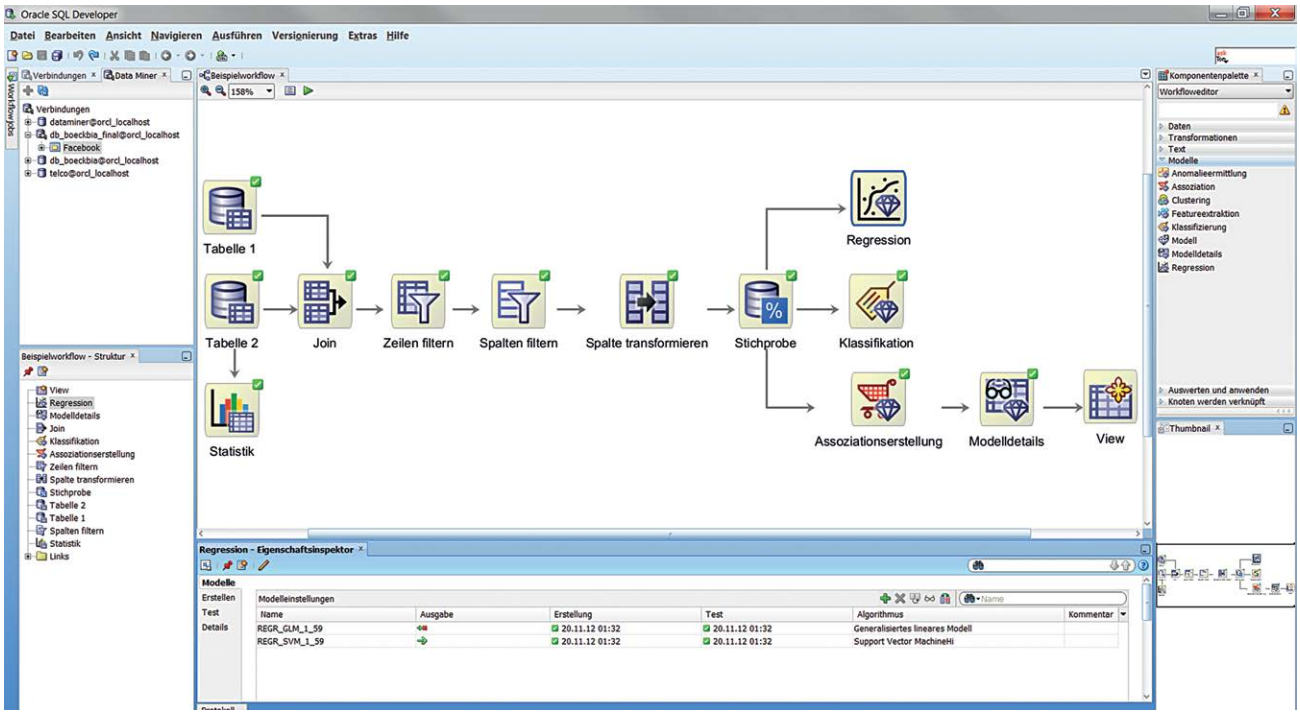

Abbildung 2: Benutzeroberfläche des Oracle Data Miner mit der Anzeige eines Beispiel-Workflows

Freundschafts-Beziehungen erfasst. Hierbei wurden nur diejenigen Freunde in der Datenbank gespeichert, die auch explizit der Anwendung zugestimmt hatten.

Im Schritt der Vorverarbeitung wurden zum einen fehlende Werte behandelt und zum anderen die Daten bereinigt. Während einige fehlende Daten wie der Wohn- und Heimatort anhand anderer im eigenen Profil zu findenden Informationen abgeleitet werden konnten, war dies für andere Attribute nicht möglich. Bei letzteren wurde der Median beziehungsweise der Modus unter allen Teilnehmern eingesetzt, um fehlende Werte zu behandeln.

Selten angegebene Attribute wurden jedoch von der weiteren Betrachtung ausgeschlossen, da zu viele künstliche Werte die Analyse verfälschen könnten. Zudem war eine Bereinigung der Daten notwendig, da in Facebook aufgrund von beispielsweise Rechtschreibfehlern und Mehrsprachigkeit viele unterschiedliche Bezeichnungen für das semantisch gleiche Objekt existieren.

Für die Erstellung eines hierfür notwendigen einheitlichen Vokabulars kamen DBpedia [8], Freebase [9] und andere frei verfügbare Datenquellen zum Einsatz. Die Profil-Informationen aus Facebook wurden automatisch jeweils dem Wort aus den zuvor genannten Datenquellen zugeordnet, das die maximale Jaro-Winkler-Ähnlichkeit besitzt und einen gewissen Schwellenwert überschreitet. Die Jaro-Winkler-Ähnlichkeit ist als Funktion im Oracle-Datenbank-Package „UTL_MATCH“ verfügbar. Im Anschluss daran wurde eine manuelle Nachkorrektur aufgrund von Synonymen und Homonymen vorgenommen.

In der Transformationsphase wurde neben der Normalisierung von metrischen Attributen und Datentyp-Konvertierungen auch eine Kategorisierung vorgenommen, um durch die Reduzierung der Kardinalität von Attributen möglicherweise kompaktere Modelle zu erzielen. Hierfür wurden vorwiegend die zuvor erwähnten Datenquellen genutzt. Beispielsweise hat man eine Orts-Hierarchie aufgebaut sowie TV-Sendungen ihren Genres beziehungsweise das Alter mittels Quantil-Binning Altersgruppen zugeordnet.

Darüber hinaus wurden mit dem ODMr Ausreißer anhand von One-Class SVM identifiziert und entsprechend behandelt. Als Ausreißer gelten Nutzer mit seltenen, von der Mehrheit abweichenden Informationen. Beispielsweise hat man unrealistische Altersangaben im Bereich von 100 Jahren geglättet.

Data-Mining-Funktionen	Data-Mining-Algorithmen
Klassifikation	• Entscheidungsbaum • Naive Bayes • Support Vector Machine (SVM) • Generalisiertes Lineares Modell (GLM) → *logistische Regression*
Regression	• GLM → lineare Regression • SVM
Assoziations-Analyse	• Apriori
Clustering	• k-Means • O-Cluster • Erwartungs-Maximierung
Anomalie-Erkennung	• One-Class SVM
Feature-Extraktion	• Nicht-negative Matrix-Faktorisierung • Singulärwert-Zerlegung mit Hauptkomponenten-analyse
Attribut-Wichtigkeit	• Minimum Description Length (MDL)

Tabelle 1: Überblick über die Data-Mining-Algorithmen in ODM 12.1; Algorithmen in blauer Schriftfarbe sind in ODM 12.1 neu hinzugekommen [7]

Ein weiterer wichtiger Punkt ist die probate Abbildung der Freundschaftsbeziehungen für die Prognose. Nur ein Teil der knapp 700 teilgenommenen Facebook-Nutzer konnte für die Vorhersage genutzt werden, da nur bei wenigen auch genügend Freunde der Anwendung zugestimmt hatten und somit deren Daten vorlagen. Während beispielsweise 241 Nutzer mindestens drei teilgenommene Freunde besaßen, waren es bei mindestens fünf Freunden nur noch 150 Facebook-Mitglieder.

Außerdem wurden solche Personen gänzlich von der Analyse ausgeschlossen, die mehr als 700 Leute in ihrer Facebook-Freundesliste pflegten. Dies liegt in der Vermutung begründet, dass neben engen Freundschaften auch viele Bekanntschaften enthalten sind, die eine Prognose wahrscheinlich verschlechtern würden. Zudem war es vorstellbar, neben den Personen aus der eigenen Freundesliste zusätzlich die indirekten Freunde einzubeziehen, also die Freunde der Freunde, um möglicherweise bessere Modelle zu erhalten.

Da die Anzahl der Freunde über verschiedene Personen hinweg variiert und ODM genau eine Fall-Tabelle voraussetzt, ist dazu eine Aggregation der Freundeswerte notwendig gewesen. Es wurde deshalb für jeden Nutzer ein Freundesvektor für jedes Attribut erstellt, der dem gewichteten relativen Vorkommen eines jeden Attributwerts unter seinen Freunden entspricht.

ODM 11.2 stellt für die Speicherung solcher verschachtelter Daten die Datentypen „DM_NESTED_NUMERICALS“ und „DM_NESTED_CATEGORICALS“ bereit [6], in ODM 12.1 stehen zusätzlich „DM_NESTED_BINARY_DOUBLES“ und „DM_NESTED_BINARY_FLOATS“ zur Verfügung [5]. Da die Anzahl der indirekten Freunde viel größer ist als die der direkten, wurde der relative Anteil für die zwei Freundespartitionen separat berechnet. Um den direkten Freunden eine größere Bedeutung zukommen zu lassen als den indirekten, wurde ihnen bei der anschließenden Summierung der Anteile ein größeres Gewicht zugewiesen.

Irrelevante und korrelierte Attribute verfälschen ein Modell. Aus diesem Grund wurden mit dem ODMr

irrelevante Attribute für die Vorhersage bestimmter Merkmale automatisch mittels Minimum Description Length (MDL) gefiltert.

Data-Mining

Das Ziel der Analyse war eine Prognose der Attributwerte eines Facebook-Nutzers ausschließlich anhand der Angaben seiner Freunde. Der vorherzusagende Nutzer stellte somit eine Blackbox dar. Die Klassifikation und die Regression sind für eine solche Aufgabenstellung geeignet. Für eine exemplarische Prognose von Geschlecht und Wohnort kamen die Klassifikations-Algorithmen SVM und Naive Bayes zur Anwendung. Aufgrund einer binären Ziel-Variablen kam beim Geschlecht zusätzlich die logistische Regression zum Einsatz.

Da der Entscheidungsbaum erst in ODM 12.1 verschachtelte Daten verarbeiten kann [5, 6], wurde deshalb beim Freundesvektor ein repräsentativer Wert ausgewählt. Für die Vorhersage des metrischen Alters wurden bei der Regressions-Analyse die lineare Regression und SVM angewendet. Eine Evaluierung der verschiedenen Modelle war durch eine Aufteilung der Objektmenge in zwei Drittel Trainings- und ein Drittel Test-Menge gegeben.

Darüber hinaus existiert neben den klassischen Prognose-Verfahren auch die Möglichkeit, anhand von Assoziationsregeln Gemeinsamkeiten zwischen allen Nutzern, also nicht nur zwischen seinen Freunden, für die Prognose genau eines Merkmals einzusetzen. Das setzt voraus, dass der Nutzer selbst für ein Attribut mindestens einen Wert öffentlich hält, damit anhand von Regeln weitere private Angaben vorhergesagt werden können.

Die Assoziations-Analyse wurde exemplarisch für TV-Sendungen durchgeführt. Gehört beispielsweise „The Big Bang Theory“ zu den Lieblings-Fernsehsendungen eines Nutzers, so kann man anhand der gefundenen Regel „The Big Bang Theory“ → „How I Met Your Mother“ die nicht in seinem Profil öffentlich einsehbare Serie „How I Met Your Mother“ zu seinen favorisierten Sendungen zählen.

Unter der Annahme, dass alle Informationen der Freunde eines Nutzers öffentlich zugänglich sind, konnte die Vermutung bestätigt werden, dass die Daten der Freunde gut geeignet sind, um mehr über einen Facebook-Nutzer zu erfahren. Beachtlich ist die gute Prognose des Alters anhand des Geburtstages der Freunde. Durchschnittlich wich das vorhergesagte Alter nur um 2,05 Jahre vom tatsächlichen ab. In sieben von zehn Fällen konnte der Wohnort korrekt anhand verschiedener Angaben der Freunde geschlussfolgert werden, das Geschlecht war in 65 Prozent aller Fälle richtig.

Neben der Untersuchung der prinzipiellen Ableitbarkeit wurden verschiedene Einflussfaktoren für die Güte der Ergebnisse analysiert. Es hat sich gezeigt, dass die Qualität der extrahierten Modelle unter anderem von der Wahl des Data-Mining-Algorithmus abhängt. Für die Vorhersage des Wohnorts war Naive Bayes am besten geeignet, beim Geschlecht war es hingegen SVM.

Der Entscheidungsbaum, der im Gegensatz zu anderen Modellen vom Menschen leichter interpretierbar ist, hatte im Allgemeinen schlechte Ergebnisse geliefert. Die lineare Regression war für das Ableiten des Alters besser geeignet als die SVM. Zudem stellte sich heraus, dass die zusätzliche Betrachtung indirekter Freunde zu keiner Verbesserung führt. Es war außerdem zu erkennen, dass die Ergebnisse beim Einbeziehen von mindestens fünf direkten Freunden besser waren als bei mindestens drei.

Es konnte jedoch keine allgemeine Aussage hieraus abgeleitet werden, da die Testmengen unterschiedlich groß waren. Die zuvor durchgeführte Bereinigung der Daten und die Einordnung in Kategorien führte nur beim Bestimmen des Wohnorts zu besseren Resultaten. Dieser Aufwand war bei der Vorhersage des Geschlechts und des Alters nicht nötig gewesen.

Es wurden außerdem Assoziationsregeln auf TV-Sendungen gelernt. Hierbei konnten 21 Regeln gefunden werden, die einen Support von mindestens 10 Prozent, eine Konfidenz von mindestens 40 Prozent und einen Lift größer als eins besitzen. Es konnten keine Regeln extrahiert werden, die sowohl hohe Support-, Konfidenz- als auch Liftwerte besitzen.

Fazit

Aus diesen Erkenntnissen lässt sich schlussfolgern, dass neben privaten Profil-Informationen auch die Freundschafts-Beziehungen vor der Öffentlichkeit verborgen werden sollten, um die eigene Privatsphäre zu schützen. Hierbei müssten jedoch beide Freunde ihre Privatsphäre-Einstellungen in Facebook ändern, um Fremden nicht die Möglichkeit zu geben, die Beziehung indirekt über die andere Person herausfinden zu können.

Literatur

[1] M. Ester und J. Sander, Knowledge Discovery in Databases - Techniken und Anwendungen, Springer, 2000.

[2] U. M. Fayyad, G. Piatetsky-Shapiro und P. Smyth, „From Data Mining to Knowledge Discovery: An Overview,“ in Advances in Knowledge Discovery and Data Mining, Menlo Park, CA, USA, American Association for Artificial Intelligence, 1996, pp. 1-34.

[3] U. M. Fayyad, G. Piatetsky-Shapiro und P. Smyth, „From Data Mining to Knowledge Discovery in Databases,“ AI Magazine, Bd. 17, Nr. 3, pp. 37-54, 1996.

[4] K. L. Taylor, „Oracle Data Mining Concepts, 11g Release 2 (11.2), E16808-06,“ Juli 2011. [Online]. Available: http://docs.oracle.com/cd/E11882_01/datamine.112/e16808.pdf.

[5] K. L. Taylor, „Oracle Data Mining User‘s Guide, 12c Release 1 (12.1), E17693-13,“ Mai 2013. [Online]. Available: http://docs.oracle.com/cd/E16655_01/datamine.121/e17693.pdf.

[6] K. L. Taylor, „Oracle Data Mining Application Developer‘s Guide, 11g Release 2 (11.2), E12218-07,“ Juli 2011. [Online]. Available: http://docs.oracle.com/cd/E11882_01/datamine.112/e12218.pdf.

[7] K. L. Taylor, „Oracle Data Mining Concepts, 12c Release 1 (12.1), E17692-13,“ Mai 2013. [Online]. Available: http://docs.oracle.com/cd/E16655_01/datamine.121/e17692.pdf.

[8] DBpedia. [Online]. Available: http://dbpedia.org.

[9] Freebase. [Online]. Available: http://www.freebase.com/.

Bianca Böckelmann
bianca.boeckelmann@robotron.de