Klassifizierung für E-Commerce Nutzergruppen auf Basis von Data Mining unter Verwendung von künstlicher Intelligenz.

Untersuchung der Nutzerklassifizierung durch Datenanalyse

vorgelegte Masterarbeit
zum Erlangen des akademischen Grades
Master of Science (M.Sc.)

eingereicht von: Wilfried Pahl
Matrikelnummer: 901932
Studiengang: Online Medieninformatik
Berliner Hochschule für Technik

Betreuender Prüfer Prof. Dr. S. Edlich Berliner Hochschule für Technik
Zweitgutachter noch nicht bekannt auch von einer Hochschule

Temmen-Ringenwalde, der 14. November 2022

Stichworte

Data-Mining, Big Data, künstliche Intelligenz, Clusterung, Nutzergruppen, k-Means-Algorithm.

Kurzzusammenfassung

Hier kommt das Abstract auf Deutsch.

Keywords

Data-Mining, Big Data, künstliche Intelligenz, Clusterung, Nutzergruppen, k-Means-Algorithm.

Abstract

Here comes later the abstract on German.

Danksagung

Hier sage ich auch mal zu irgendjemand Danke.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe, dass ich sie zuvor an keiner anderen Hochschule und in keinem anderen Studiengang als Prüfungsleistung eingereicht habe und dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die wörtlich oder sinngemäß aus Veröffentlichungen oder aus anderweitigen fremden Äußerungen entnommen wurden, sind als solche kenntlich gemacht.

Datum Unterschrift

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einführung

Das Volumen der bestellten Waren im E-Commerce Bereich wächst stetig. Die Corona-Krise hat diesen Prozess noch beschleunigt. Die meisten Umsätze verzeichnen virtuelle Marktplätze, wie Amazon und Ebay. Aber auch kleine Onlineshops können durchaus bestehen. Im E-Commerce nutzen Onlineshop Betreiber Kundenklassifizierungen, um Kunden einzuteilen und ihnen besser Angebote unterbreiten zu können. Oft sind diese Klassifizierungen in den ersten Schritten bei der Shop-Planung oder bei einem Relaunch erstellt wurden, basierend auf einer erwarteten und anvisierten Kundengruppe. In vielen Fällen werden keine Kundenklassifizierungen erstellt. Es wird unterstellt das alle Menschen zur Zielgruppe gehören. Mithilfe von Machine Learning können aus vorhandenen Daten, Nutzergruppe bestimmt werden.

1.1 Motivation

Hier kommt auch die Problembeschreibung.

1.2 Ziele der Arbeit

1.3 Inhaltlicher Aufbau der Arbeit

Kapitel 2

Grundlagen

2.1 Grundbegriffe

Big Data

Eine Abgrenzung des Begriff Big Data ist nicht eindeutig, da der Begriff sehr heterogen verwendet wird. Somit gibt es viele Definitionen vom Begriff Big Data.

Als Big Data werden Daten bezeichnet, die entweder zu groß, zu komplex, zu schnelllebig oder zu schwach strukturiert sind, um diese mit herkömmlichen Methoden auszuwerten. Big bezieht sich in der Definition auf die vier Dimensionen. In Big Data sind Technologien die richtigen Informationen dem richtigen Adressaten zur richtigen Zeit in der richtigen Menge am richtigen Ort und in der erforderlichen Qualität bereitstellen.

Auch strukturierte Daten. Verschiedene (autonome) Datenquellen (Datenbanken oder Anwendungen).

Auf volume (Umfang, Datenvolumen), velocity (Geschwindigkeit, mit der die Datenmengen generiert und transferiert werden), variety (Bandbreite der Datentypen und -quellen) und veracity (Echtheit von Daten).

Knowledge Discovery in Data (-base)

Dies hat das Ziel aus vorhandenen meist großen Datenbeständen, fachliche Zusammenhänge zu erkennen. Zu den Teilschritten des KDD Prozesses gehören 1. Bereitstellung von Hintergrundwissen, 2. Definition der Ziele, 3. Datenauswahl, 4. Datenbereinigung, 5. Datenreduktion, 6 Auswahl eines Modells, 7. Data-Mining, die eigentliche Datenanalyse, 8. Interpretation der gewonnenen Erkenntnisse.

Data-Mining

Data-Mining die systematische Anwendung von statischen Methoden auf große Datenbestände, um neue Querverbindungen zu erkennen. Data-Mining ist ein Teilprozess KDD Prozesses. Mit Data-Mining findet ein Informationsgewinn oder -erweiterung, aus den Big Data statt. Hier werden Algorithmen aus dem Bereich ML angewandt.
k-Means Algorithmus

Machine Learning

Maschinelles Lernen (eng. Machine Learning) ist nach [https://datasolut.com/was-ist-

machine-learning] ein Teilbereich der künstlichen Intelligenz, der System in die Lage versetzt, automatisch aus Erfahrungen (Daten) zu lernen und sich zu verbessern. Aufgaben die Machine Learning erledigen kann, ist Berechnung von Wahrscheinlichkeiten für bestimmte Ereignisse, Erkennen von Gruppen und Clustern in Datensätzen, Erkennen von Zusammenhängen in Sequenzen, Reduktion von Dimensionen ohne großen Informationsverlust und Optimierung von Geschäftsprozessen.

Künstliche Intelligenz

Eine einheitliche gemeingültige Definition von künstlicher Intelligenz zu geben ist nicht einfach. Zuvor muss Intelligenz definiert werden. Aber was ist Intelligenz? In der Literatur werden kognitive Fähigkeiten oft mit Intelligenz in Verbindung gebracht. In der Definition von künstlicher Intelligenz gibt es schwache oder enge künstliche Intelligenz, die auf die Lösung bestimmter Aufgaben beschränkt ist und menschliche Intelligenz nicht imitieren kann. Starke oder allgemeine künstliche Intelligenz hingegen ist in der Lage die kognitiven Fähigkeiten des Menschen zu erzielen.

Weitere Definitionen werden im Verlauf der Arbeit ergänzt.

2.2 Verwandte Arbeiten

Mit weiteren Anwendungsfällen beschäftigt sich das ,,Fraunhofer Institut BIG DATA" in ihrem Paper [Big Data Frauenhofer].

Um den Einsatz künstlicher Intelligenz im werte-orientierten Marketing zu bewerten, befassten sich in [EConster] einige Professoren des ,,Leibniz-Informationszentrum Wirtschaft" mit diesem Thema.

Die Masterarbeit [Bitstream] von Eduard Weigandt befasst sich mit der Personalisierung im E-Commerce basierend auf Data-Mining. Interessante Grundlagen zum Einstieg in E-Commerce und künstlicher Intelligenz sind auf [Eqop] zu finden. Die Firma Kobold AI befasst sich in ihrem Artikel [Kobold.ai] ,,Optimale Segmentierung von Bestandskunden durch KI" und erläutert Methoden zur Clustering von Bestandskunden. ,,Datasolut" ist ein weiteres Unternehmen das sich in [Datasolut1] mit Kundenklassifizierung, Clusteranalyse und maschinellem Lernen befasst. Ebenfalls von ,,Datasolut" ist der Artikel [Datasolut2] in dem erfolgreiche Anwendungen und Beispiele zum Thema künstlicher Intelligenz im E-Commerce aufgezeigt werden.

2.3 Big Data

Big Data beschäftigt sich nach [DataSolut3] und [Oracle] mit dem Sammeln, Verarbeiten und Zusammenführen von großen Datenmengen. Um diese Daten für die Entscheidungsfindung und Prozessautomatisierung zu verwenden. Dabei stammen die Daten aus den unterschiedlichsten Quellen, aus verschiedenen Datenbanken oder auch direkt aus Programmen. Als Datenquellen können folgende infrage kommen:

- Internetnutzung
- Social Media

- Geo-Tracking
- Cloud Computing
- Vitaldaten-Messung
- Media-Streaming

Diese Daten können strukturiert, aber auch unstrukturiert vorliegen. [Gratner] beschrieb Big Data anhand von den ,,4 V's". Mit der Zeit wurde es um ein ,,V" erweitert. Diese Beschreibung wird in unterschiedlichen Publikationen aufgegriffen, unter anderen auch in [Oracle].

Volume (Volumen)

Immer größere Datenmengen müssen Verarbeitet werden. Durch die stetig zunehmende Digitalisierung in immer mehr Lebensbereichen wächst die erzeugte Datenmenge pro Zeiteinheit immer mehr an. So werden großen Datenmengen nicht nur durch die oben genannten Quellen erzeugt, sondern auch z. B. durch Gerätesensoren. Hierbei können etliche Terabytes oder hunderte Petabytes an Daten anfallen. Wie die Abbildung 1 [Statistika] zeigt, wird das Datenvolumen im Jahr 2025 auf 181 Zettabyte vorhergesagt.

Hier kommt Abbildung 1 hin.

Variety (Vielfalt)

Durch die unterschiedlichen Bereiche, in denen die Datenmengen entstehen sind, diese sehr unterschiedlich und zu meist unstrukturiert. Oft liegen diese in relationalen Datenbanken und können dort nicht ausgewertet werden. Neben Texten liegen die Daten in Bildern und Videos vor die Analyse erfolgt durch Machine Learning Algorithmen .

Velocity (Geschwindigkeit)

Mit der Entwicklung der Technik produzieren Softwaresysteme mit einer höheren Geschwindigkeit mehr Daten. Bei vielen Produkten fließen die Daten nicht auf eine Festplatte, sondern werden direkt im Speicher verarbeitet. Solche Produkte arbeiten in Echzeit oder beinahe in Echtzeit. Deren Verarbeitung in immer kürzerer Zeit erfolgt. Für Unternehmen und verschiedenen Use Cases kann die Verarbeitung in Echtzeit einen erheblichen Wettbewerbsvorteil bedeuten.

Veracity (Wahrhaftigkeit)

Da die Daten oft aus Quellen kommen, deren Wahrheitsgehalt nicht sicher ist und die Daten in nicht geeigneter Qualität vorliegen, können diese nicht ohne eine aufwendige Nachbearbeitung eingesetzt werden.

Value (Mehrwert)

Durch die Verknüpfung der Daten, die beim Einsatz der Techniken des Machine Learning entstehen, ist dieser Mehrwert eines der wichtigsten ,,V" bei Big Data. Ohne diesen Mehrwert würde Big Data keinen Sinn ergeben.

2.3.1 Technologien

Rund um das Thema Big Data, haben sich verschiedene Technologien entwickelt, die Ansätze für die Verarbeitung von großen Datenmengen liefern. Nachfolgende sind Open Source Produkte stellvertretend einige genannt.

Apache Spark

Es ist nach eigenen Angaben [Spark] eine mehrsprachige Engine zur Ausführung von Data Engineering, Data Science und maschinelles Lernen auf Single-Node-Maschinen oder Clustern.

Apache Hadoop

Hadoop ist nach [Hadoop] ein Framework das mit einfachen Programmiermodellen, welches eine verteilte Verarbeitung von großen Datenmengen anbietet, das von einzelnen Server auf mehrere tausend skaliert werden kann.

Apache Cassandra

Nach eigener Beschreibung [Cassandra], ist Cassandra ist ein skalierbares und hochverfügbares verteiltes Datenbanksystem. Es basiert auf NoSQL, ist Open-Source und kann ebenfalls auf einzelnen Servern oder in der Cloud eingesetzt werden.

2.4 Data-Mining

2.5 Clustering

Abkürzungen
ML Mashine Learning
KI Künstliche Intelligenz

Kapitel 3

Kern der Arbeit

3.1 Probleme und Lösungsansätze

3.2 Methodiken und Vorgehen

3.3 Architektur

3.4 Algorithmen

Hier wird der k-Means-Algorithmus erläuter.

Kapitel 4

Implementierung

4.1 Umsetzung der Datenverarbeitung

So was wie Daten bereinigen und zusammenführen.

4.2 Umsetzung der Clusterung

Kapitel 5

Evaluation

5.1 Ausbau der Umgebung

Hier könnte das CMS erwähnt werden.

5.2 Ergebnisse

5.3 Bewertung und Diskussion

Kapitel 6

Zusammenfassung

Kapitel 7

Ausblick

Anhang

Literaturverzeichnis

Glossar

CMS Nutzerfreudliche Bedienungsoberfläche einer Software.. 11