Dr. Ralph Guderlei *31. Januar 2017*

Predictive Analytics – Von der Idee zur Umsetzung

© Sergey Nivens / Fotolia.com

Die meisten Unternehmen bilden ihre Kernprozesse über eine Vielzahl von betrieblichen Informationssystemen ab. Diese werden über einen langen Zeitraum betrieben und enthalten daher oft große Mengen an Daten. Bei diesen Informationssystemen handelt es sich teilweise um Standard-Lösungen, aber oft auch um speziell für einen Zweck entwickelte Individuallösungen.

Der von den Informationssystemen verwaltete Datenbestand wird jedoch oft nur innerhalb der abgebildeten Prozesse verwendet. Die Datenmenge und der lange Zeitraum, über den die Daten gesammelt wurden, bietet sich jedoch dazu an, die Daten auch für tiefergehende Analysen zu nutzen. Diese können dann dazu verwendet werden, um die Prozesse zu verbessern. Die Methoden und Vorgehensweisen um diese Ziele zu erreichen, werden als "Predictive Analytics" bezeichnet.

Der Artikel soll anhand einer nicht-öffentlichen Online-Community zeigen, wie mit einfachen Mitteln das Prozessverständis verbessert und damit eine höhere Kundenzufriedenheit erreicht werden kann.

Definitionen und Beispiele

Predictive Analytics beschäftigt sich mit Verfahren für die Extraktion von Informationen aus Daten mit dem Zweck, aus den gewonnenen Informationen Trends oder Verhaltensmuster abzuleiten. In der Regel wird der Begriff für Prognosen von zukünftigen Ereignissen verwendet, die Analysen können sich aber auch auf gegenwärtige oder vergangene Ereignisse beziehen. Für Predictive Analytics werden Verfahren aus dem Data Mining und induktive statistische Methoden verwendet. Typische Verfahren sind Regressionen oder Klassifikationsmethoden.

Die eingesetzten Verfahren und das Ziel, zukünftige Entwicklungen einzuschätzen, unterscheiden Predictive Analytics von der klassischen Business Intelligence (BI). Bei dieser steht die Aggregation und Beschreibung von Daten im Vordergrund.

Grobes Vorgehensmodell

Die Umsetzung einer Predictive-Analytics-Aufgabe erfolgt in vier Schritten:

1. Rohdaten sammeln
2. Datenaufbereitung und Informationsextraktion
3. Explorative Datenanalyse und Modellbildung
4. Anwendung des Modells auf zukünftige Daten

Im ersten Schritt werden die Rohdaten gesammelt. Im einfachsten Fall liegen die Daten bereits in Dateien vor (beispielsweise bei Log-Dateien) oder werden aus einer Datenbank exportiert. Die Rohdaten dienen dazu, einen Trainings-Datenbestand aufzubauen, der dann in den nächsten Schritten verwendet wird, um das gewünschte Verfahren zu kalibrieren.

Die Rohdaten enthalten normalerweise entweder zu viele Daten, oder die gewünschten Informationen müssen zuerst aus den Daten extrahiert werden. Gegebenenfalls können Informationen für Trainingsdaten auch manuell hinzugefügt werden. Das ist oft bei Klassifikationsproblemen der Fall, wenn die Trainings-Datenbestände manuell klassifiziert werden.

Bei der folgenden explorativen Datenanalyse versucht man, Annahmen über Strukturen und Zusammenhänge der Daten zu entwickeln. Dazu werden Mittel der deskriptiven Statistik und unterschiedliche graphische Aufbereitungen der Daten verwendet. Das Ziel der explorativen Datenanalyse ist es, mit Hilfe der getroffenen Annahmen Vorgehensweisen und Verfahren für die eigentliche Predictive Analytics-Lösung auszuwählen. Als Werkzeuge für die explorative Datenanalyse wird häufig das Jupyter Notebook → [1] oder R → [2] genutzt. Diese

Autor

Dr. Ralph Guderlei

Dr. Ralph Guderlei ist Technology Advisor bei der eXXcellent solutions GmbH in Ulm. Neben der Arbeit als Architekt/Projektleiter in...
>> Weiterlesen

Nachrichten

24.01.2017
Sicherheitsupdate: Apple bringt macOS 10.12.3, iOS 10.2.1, watchOS 3.1.3 und tvOS 10.1.1

Apple hat iOS 10.2.1, macOS 10.12.3, watchOS 3.1.3 und tvOS 10.1.1 veröffentlicht. Die Updates der Apple-Betriebssysteme dienen der...
>> Weiterlesen

24.01.2017
3. Industrie 4.0-Konferenz des Hasso-Plattner-Instituts

Am 31.01.2017 findet zum dritten Mal die Industrie 4.0-Konferenz des Hasso-Plattner-Instituts statt. Führende Experten aus Wirtschaft,...
>> Weiterlesen

20.01.2017
SQLGrillen3 im Juni 2017: Agenda online!

Die Agenda für das dritte SQLGrillen in Lingen, am 2. Juni 2017, ist online. Das kostenlose Programm enthält 28 Sessions in vier parallelen...
>> Weiterlesen

ermöglichen ein schnelles, quasi interaktives Arbeiten.

Aufgabenstellung

Die Basis für unser Beispiel ist die nicht-öffentliche Online-Community einer privaten Fern-Hochschule. Da Studenten und Lehrkräfte sich nur selten bei Präsenzveranstaltungen treffen, ist diese Community eine einfache und direkte Art miteinander zu kommunizieren. Auch die Kommunikation von Studenten untereinander (z. B. in Arbeitsgruppen) findet über diese Plattform statt. Die Community ist in Diskussions-Threads organisiert, die wiederum Bereichen zugeordnet sind.

Die Zufriedenheit der Studenten mit der Community ist im Wesentlichen dadurch bestimmt, ob und wie schnell sie Antworten auf ihre Fragen bekommen. Die Aufgabe hier ist es, Informationen zur Verfügung zu stellen, um Reaktionszeiten zu verbessern und Bereiche mit sich verschlechternden Reaktionszeiten rechtzeitig zu identifizieren, um gezielt gegensteuern zu können. Es ist nicht das Ziel, ein vollkommen autonomes System zu entwickeln, daher sind gewisse Ungenauigkeiten in den Ergebnissen zulässig.

Klassifikationsregeln

Um diese Aufgabe zu lösen, muss zunächst ein Verfahren entwickelt werden, um Fragen und Antworten hinreichend genau identifizieren zu können. Die einfachste Möglichkeit sind simple Regeln zu Klassifikation:

- Jede Nachricht, die ein "?" enthält und von einem Studierenden verfasst wurde, ist eine Frage
- Jede auf eine Frage folgende Nachricht, die selbst keine Frage ist, ist eine Antwort

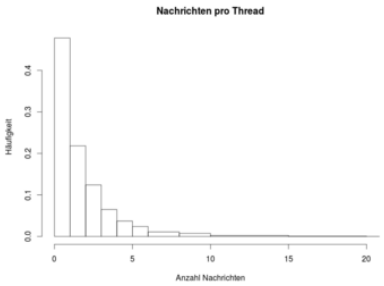

Abb.1: Nachrichten pro Threat. © Dr. Ralph Guderlei

Diese Regeln scheinen zunächst viel zu einfach zu sein. Betrachtet man jedoch die Anzahl der Nachrichten pro Thread, sieht man, dass ca. 80 Prozent der Threads maximal drei Nachrichten enthalten. Eine stichprobenartige Überprüfung der Daten hat gezeigt, dass in diesen kurzen Threads tatsächlich gezielt geantwortet wurde. Nur selten gibt es Nachrichten in der Art "*ja, will ich auch wissen*", die die Klassifikationsregel zwar als Antwort interpretieren würde, die aber de facto keine ist.

Generell lässt sich sagen, dass einfache Methoden oft ausreichen, um akzeptable Ergebnisse zu erzielen. Insbesondere bei Textklassifikationen erzielen deutlich komplexere Methoden oft nur geringfügig bessere Resultate.

Sinnhaftigkeit der Aufgabe

Als nächstes stellt sich die Frage, wie relevant die Community tatsächlich für die Hilfe bei Problemen ist. Abb.2 zeigt, dass ca. 30 Prozent der Nachrichten als Fragen klassifiziert wurden und damit einen erheblichen Anteil am Nachrichtenaufkommen darstellen.

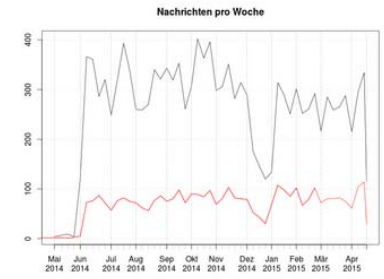

Abb.2: Fragen pro Woche. © Dr. Ralph Guderlei

Analyse der Antwortzeiten

Der letzte Baustein für die Aufgabenstellung ist die Analyse der gemessenen Antwortzeiten. In Abb.3 ist erkennbar, dass über 80 Prozent der Fragen innerhalb von 5 Tagen beantwortet werden. 40 Prozent der Fragen werden sogar innerhalb eines Tages beantwortet. Die durchschnittliche Antwortzeit beträgt 3,5 Tage. Allerdings ist die Standardabweichung mit fast 11 Tagen ziemlich hoch, da es einige Fragen mit sehr hohen Antwortzeiten gibt. Aber man kann sagen, dass in der Regel Fragen schnell beantwortet werden.

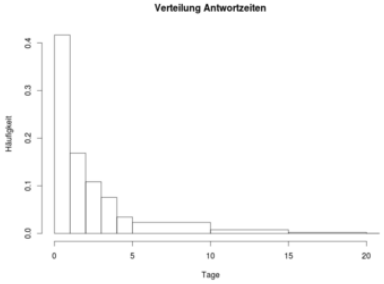

Abb.3: Analyse der gemessenen Antwortzeiten. © Dr. Ralph Guderlei

Prognose

Mit Hilfe der gewonnenen Informationen kann nun ein Verfahren für die Verbesserung des Antwortverhaltens entwickelt werden. Zum einen können die Bereiche identifiziert werden, deren durchschnittliche Antwortzeiten deutlich über dem Mittel liegen. Zum anderen kann eine Trendanalyse benutzt werden, um Bereiche zu identifizieren, die steigende Antwortzeiten aufweisen. Dazu werden alle Fragen & Antworten aus einem Zeitraum von beispielsweise vier Wochen betrachtet und der Trend mittels linearer Regression bestimmt. Die Ergebnisse sind allerdings mit Vorsicht zu genießen, da die Daten mit Sicherheit zyklische Effekte beinhalten: Es ist davon auszugehen, dass Fragen, die am Wochenende gestellt werden, erst nach dem Wochenende beantwortet werden und damit eine höhere Antwortzeit aufweisen, als Fragen, die am Anfang der Woche gestellt werden.

Betrachtet man einen ausreichend langen Zeitraum (vier Wochen), dann sollten diese Effekte gedämpft werden.

Über die Trendanalyse können dann die Antwortzeiten für künftige Fragen prognostiziert werden. Liegen die prognostizierten Antwortzeiten in einem Bereich über einem Grenzwert (bspw. 5 Tage), können gezielt Maßnahmen ergriffen werden.

Umsetzung

Die tatsächliche Umsetzung der Aufgabenstellung hat eine Reihe von Anforderungen, die die Werkzeuge für explorative Datenanalyse nicht abdecken, z. B. den Umgang mit großen Datenmengen und die Anbindung an das bestehende Community-System. Für die Umsetzung wurde deswegen Apache Spark (Streaming) gewählt: Spark lässt sich über Message Queues oder Datenbank einfach mit Daten aus dem Community-System versorgen und bringt über die MLLib die benötigten Algorithmen mit. Darüber hinaus erlaubt Spark unterschiedliche Betriebsmodi von Stand-alone bis Cluster-Betrieb und kann flexibel an den erforderlichen Durchsatz angepasst werden.

Die Umsetzung beinhaltet die Schritte

1. Gruppieren der Nachrichten nach Threads
2. Klassifizieren von Fragen und Antworten
3. Berechnung der Antwortzeiten
4. Zusammenfassen der Antwortzeiten pro Bereich
5. Trendanalyse der Antwortzeiten pro Bereich
6. Prognose der künftigen Antwortzeiten pro Bereich

Fazit

Mittlerweile gibt es eine große Zahl guter Werkzeuge für die Datenanalyse und die performante Umsetzung von Analysen. Aus vorhandenen Daten (in diesem Fall die Nachrichten einer Online-Community) lassen sich dadurch zusätzliche Informationen gewinnen, die helfen, ein Produkt oder einen Service zu verbessern oder um neue Funktionen zu erweitern.

Quellen

[1] Jupyter Notebook
[2] The R Project for Statistical Computing

Autor

Dr. Ralph Guderlei

Dr. Ralph Guderlei ist Technology Advisor bei der eXXcellent solutions GmbH in Ulm. Neben der Arbeit als Architekt/Projektleiter in unterschiedlichen Kundenprojekten berät er Teams in technologischen und methodischen...
>> Weiterlesen