Bachelorarbeit/Masterarbeit

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Bachelorarbeit/Masterarbeit

Untersuchung der Möglichkeiten der Platzierung von Ausreißern in Datenbeständen zur Vermeidung von Online-Betrug und Sabotage

Gegenstand der Betrachtung sind hochdimensionale Datenbestände. Beispielsweise lässt sich ein Einkauf in einem Online-Shop als Objekt in einem Raum repräsentieren, den Eigenschaften/Dimensionen wie Datum, Produkt, Betrag usw. aufspannen. Ausreißer in solchen Daten repräsentieren oft betrügerische Transaktionen; es ist deshalb gut, sie zu kennen. Ein Ausreißer manifestiert sich in hochdimensionalen Merkmalsräumen jedoch meist nur in einer Teilmenge der Dimensionen, d. h. in einem Unterraum. Ein betrügerischer Einkauf ist z. B. i. d. R. nicht in allen Eigenschaften auffällig, sondern nur in manchen. Da die Zahl der möglichen Unterräume exponentiell mit der Anzahl der Dimensionen wächst, kommen für die Suche nach Ausreißern üblicherweise Heuristiken zur Anwendung, die nur einen Teil der Unterräume danach absuchen. Diese Unterräume sind i. d. R. die mit ausgeprägtem Kontrast, d. h. die Datenverteilung weicht stark von der ab, die man aufgrund der Verteilungen in Unterräumen erwarten würde. Man findet so also keine Ausreißer, deren Projektionen in kontrastreiche Unterräume unauffällig sind, d. h. in Teilen der Unterräume mit recht hoher Dichte liegen. Wir fragen uns jetzt zum einen, ob derartige Ausreißer in der Wirklichkeit vorkommen. Zum anderen – und das ist der Kern der Aufgabe – fragen wir uns, ob sich künstliche Datenobjekte so in den Daten platzieren/verstecken lassen, dass sie diese Eigenschaften haben. Ein Angreifer/Betrüger könnte dies ausnutzen. Umgekehrt kann man sich wohl vor Betrug/Sabotage schützen, wenn man dieses Phänomen verstanden hat. Daraus ergibt sich die folgende Aufgabenstellung, für die Daten verfügbar sind und Anwender zur Verfügung stehen:

- Gegeben ein Datenbestand und eine einfache dichtebasierte Definition von 'Ausreißer', z. B. die aus der Vorlesung 'Datenbanksysteme', Entwicklung eines effizienten Verfahrens, das die Teile eines Raums identifiziert, in denen sich kein Ausreißer platzieren lässt (weil die Dichte um ihn herum zu hoch wäre).
- Darauf aufbauend Entwicklung eines effizienten Verfahrens, das die Teile eines Raums mit folgenden Eigenschaften identifiziert:
  - Die Dichte in einem solchen Teil des Raums ist niedrig (d. h. hier ließe sich ein Ausreißer platzieren).
  - Für alle Projektionen dieses Teils des Raums in kontrastreiche Unterräume gilt, dass ihre Dichte eher hoch ist (d. h. die Projektion jenes Ausreißers wäre unauffällig).
- Verallgemeinerung der Verfahren für ausgefeiltere Ausreißer-Definitionen, z. B. LOF.
- Interaktion mit Anwendern (z. B. Experten für Online-Shopping), inwieweit Objekte in den gefundenen Teilen des Raums realistisch wären, und ggf. Anpassung der Verfahren.

Sie erwerben mit der Bearbeitung eine ausgeprägte Kompetenz (sowohl theoretisch als auch in sehr praktischer Hinsicht) im Bereich Big Data Analytics und beim Entwurf und der Bewertung von Algorithmen. Kenntnisse aus einer einschlägigen Vorlesung zu Datenanalyse sind keine Bedingung für die Bearbeitung. – Der Umfang dieser Arbeit lässt sich variieren, dadurch ist die Bearbeitung als Bachelor- oder als Masterarbeit möglich.

Ansprechpartner

Prof. Dr. Klemens Böhm	klemens.boehm@kit.edu	+49 721 608-43968	Raum: 366
Am Fasanengarten 5	76131 Karlsruhe	Gebäude: 50.34	