Data Mining oder Wissensentdeckung in Datenbanken

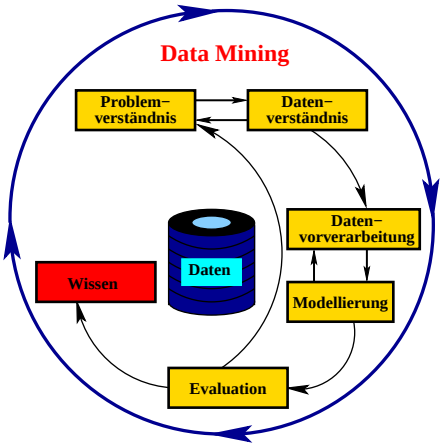
In den letzten beiden Dekaden haben wir ein enormes Wachstum der elektronisch verfügbaren Daten zu verzeichnen. Etwa alle 20 Monate verdoppelt sich die Menge der bereitgestellten Daten. Erinnert sei hier nur an das Human Genome Project, die weltweite Erfassung von Klimadaten, das ständig wachsende World Wide Web (WWW) oder die im Mobilfunk anfallenden Daten.

Die wichtigste Aufgabe besteht nun darin, das in den Daten enthaltene Wissen zu extrahieren. Diesen Prozess nennt man *Data Mining* oder *Wissensentdeckung in Datenbanken*. So möchte man z.B. wissen:

1. Welche Gene bewirken was im Organismus, oder welche Veränderungen an Genen sind krankhaft?
2. Welche Veränderungen der Sonnenaktivität beeinflussen das Klima auf der Erde nachhaltig?
3. Welche neuen Publikationen sind im WWW verfügbar, die für meine Forschungen/Entwicklungen bedeutsam sind?
4. Welche Abweichungen vom typischen Kundenverhalten deuten auf einen Betrugsversuch hin?

Wie die Beispiele zeigen, ist es häufig wichtig, das in den Daten enthaltene Wissen nicht nur zu entdecken, sondern es auch für Menschen verständlich darzustellen.

Die Lösung dieser Aufgabe ist schwierig, da die Menge der zu analysierenden Daten oft riesig ist (mehrere Gigabyte), die Daten oft in sehr inhomogener Form vorliegen, unvollständig und mit Fehlern behaftet sind und häufig auch noch relevantes Vorwissen zu integrieren ist. Die vorliegenden Daten müssen daher vorverarbeitet, transformiert, angereichert und vereinheitlicht werden, bevor mit der eigentlichen Wissensentdeckung begonnen werden kann. Dann sind geeignete maschinelle Lernverfahren oder statistische Methoden auf die so erhaltenen Daten anzuwenden. Abschließend ist das gefundene Wissen zu auszuwerten und ggf. der gesamte Prozess solange zu wiederholen, bis wirklich relevantes Wissen entdeckt wurde (siehe Abbildung).



Data Mining wird von uns nicht nur intensiv untersucht, sondern auch im Rahmen des Zukunftsinvestitionsprogramms der Bundesregierung im Verbundprojekt *DaMiT* für die Lehre an Universitäten und für potentielle Anwender aufbereitet.

Vortrag: 15:00–15:45 im Seminarraum 3/4

Ansprechpartner: Thomas Zeugmann
`http://www.tcs.uni-luebeck.de/pages/thomas/`