Daten verstehen mit Topologie
Understanding Data using Topology

Michael Kerber

Die wachsenden Größen heutiger Datensätze werfen die Frage nach neuartigen Methoden auf, um in diesen Daten relevante Informationen zu finden. Die mathematische Disziplin der algebraischen Topologie hat sich als attraktive Methode herausgestellt, um High-Level-Information für reale Daten zu extrahieren. Eine der Hauptfragen in diesem Bereich dreht sich um die algorithmischen Aspekte: Wie können die topologischen Eigenschaften effizient berechnet werden?

Wir leben im Informationszeitalter: Riesige Datenmengen werden jede Minute produziert. Beispielsweise werden täglich etwa 350 Millionen Bilder bei Facebook hochgeladen. Als „Datenanalyse" bezeichnet man den Prozess, bei dem relevante Informationen aus einer Datensammlung extrahiert und daraus Schlussfolgerungen gezogen werden. Ein Beispiel ist die Analyse des Nutzerverhaltens auf einer Internetseite wie Youtube oder Netflix, um Videos personalisiert zu empfehlen. Für reale Datensammlungen wird die Datenanalyse durch unvermeidbare Ungenauigkeiten, sogenanntes „Rauschen", erschwert.

In vielen Fällen sind qualitative Zusammenfassungen für die Datenanalyse notwendig. Beispielsweise besteht ein erster Schritt der Analyse einer Menge von Bildern darin, eine Klassifikation in ein paar wenige Kategorien vorzunehmen, zum Beispiel Bilder von Personen, von Gebäuden, von Landschaften etc. Durch das wachsende Interesse an der Datenanalyse werden neue Ansätze benötigt, um solche High-Level-Informationen aus Daten zu extrahieren.

Topologische Datenanalyse

Eine vielleicht überraschende Verbindung wurde zwischen der Datenanalyse und der Topologie von geometrischen Formen beobachtet. Die Topologie ist die mathematische Sprache zur Klassifizierung von Formen anhand der Art, wie sie zusammen- >

The growing sizes of contemporary datasets require novel ways to find the relevant information within that data. The mathematical discipline of algebraic topology has been identified as an attractive candidate to obtain high-level information on real data. One of the major questions in this field lies in the algorithmic aspects: how can such topological information be computed efficiently?

We are living in the era of information: huge amounts of data are generated every minute. For instance around 350 million photos are uploaded to Facebook every single day. "Data analysis" is the task of extracting meaningful information content and drawing conclusions from a data collection. One example are recommendation systems for user navigation on a website like Youtube or Netflix based on the choices made. In general, data analysis becomes more challenging through the inevitable presence of noise in almost any form of real data.

In many cases, qualitative, high-level summaries are required for data analysis. For instance, a first step for analysing a collection of images is a clustering into few categories, like images of people, buildings, landscapes, etc. The growing interest of data analysis asks for novel ways of extracting such high-level information.

Topological data analysis

A perhaps surprising connection has been established between data analysis and the "topology" of geometric shapes. Topology is a mathematical language for classifying shapes according to how they are connected. To illustrate the idea, a bagel and a Kaiser roll are topologically different because the former contains a hole that is missing in the roll. Moreover, a pretzel is different from the former, too, again because it has more than one hole. On the other hand, a coffee mug and a bagel are, topologically speaking, the same because both have one hole, and we can transform one into the other without ever changing the connectivity of the shape. >

Michael Kerber beschäftigt sich als Professor des FoE „Information, Communication & Computing" mit algorithmischen Fragestellungen der Topologie und Geometrie.

Michael Kerber is professor of the FoE Information, Communication & Computing and focuses on algorithmic problems in topology and geometry.

hängen. Um diese Idee zu illustrieren, betrachten wir eine Semmel und einen Bagel. Diese sind topologisch verschieden, weil Letzterer einen „Tunnel" enthält, der in der Semmel nicht existiert. Eine Brezel hingegen ist topologisch wiederum verschieden, weil sie mehr als einen Tunnel formt. Andererseits sind ein Bagel und eine Kaffeetasse unter topologischen Gesichtspunkten nicht unterscheidbar, da beide nur einen Tunnel besitzen und wir eine Form in die andere überführen können, ohne zu irgendeinem Zeitpunkt den Zusammenhang der Form zu verändern.

Die Verbindung zur Datenanalyse ergibt sich daraus, dass sich viele Datensammlungen geometrisch interpretieren lassen (unter Umständen in einem hochdimensionalen Raum) und die Topologie eine High-Level-Zusammenfassung dieser Daten liefert, die Details wie den genauen Abstand von zwei Punkten ignoriert. Jedoch reichen die klassischen Methoden der Topologie nicht aus, da sie für den idealisierten Fall von „sauberen" Formen entwickelt wurden und damit sehr anfällig für Rauschen sind. Dieses Problem wird durch die „persistente Topologie" entschärft. Die grobe Idee besteht darin, nicht nur die Anzahl der topologischen Features (zum Beispiel die Tunnel im oben beschriebenen Beispiel) zu zählen, sondern auch jedem dieser Features einen „Bedeutungswert" zuzuordnen. Das erlaubt eine feinere Analyse der topologischen Eigenschaften, insbesondere eine Unterscheidung zwischen Rauschen und relevanten Eigenschaften der geometrischen Form.

Persistente Topologie wurde auf viele verschiedene Probleme in der Datenanalyse angewandt. Zum Beispiel wurde gezeigt, dass der Raum der „natürlichen" Bilder geometrisch eine „Klein'sche Flasche" bildet. Dies ist eine verdrehte Version eines (ausgehöhlten) Bagels.

The relation to data analysis is that many data sets can be easily interpreted as geometric data (in some high-dimensional space), and topology provides a high-level summary of that data, ignoring details like the distance between points. However, the classical notions of topology are insufficient because they are developed for the idealized situation of "clean" shapes and are therefore sensitive to noise. This problem has been overcome with the development of "persistent topology". The rough idea is to not just count the number of topological features (like the number of holes in the example), but also to provide an "importance value" to each feature. This allows a more fine-grained analysis of the topological features, in particular a distinction between noise and relevant features of the shape.

Persistent topology has been applied to various questions in data analysis. As an example, the space of "natural images" has been shown to fit the geometric shape of a "Klein Bottle", which is a twisted version of a (hollow) bagel.

Computational challenges

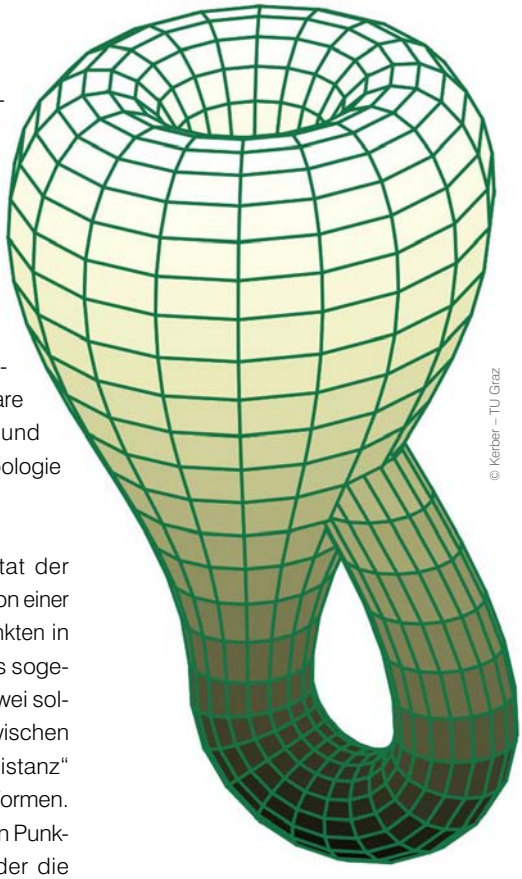
In light of the increasing size of data sets, efficient ways of computing and analysing the persistence information of data are required. Michael Kerber's research is devoted to this goal. The computational problems connect with various classical areas in algorithmics, for instance, approximation algorithms to create a shape from data, linear algebra to compute the persistence of a shape, and combinatorial optimization to compare the topologies of different shapes.

We highlight one recent result from the group: the topological information of a shape can be summarized by a multi-set of dots in the plane, called the "persistence diagram". Having two such diagrams, a distance measure between the dots yields

Algorithmische Herausforderungen

Wegen der wachsenden Größe von Datenmengen sind effiziente Ansätze zur Berechnung und Analyse von Persistenzinformationen notwendig. Die Forschungsarbeit von Michael Kerber ist auf dieses Ziel ausgerichtet. Die praktischen Probleme in diesem Themenfeld haben Verbindungen zu klassischen Feldern der Algorithmik, zum Beispiel Approximationsalgorithmen zur Generierung von geometrischen Formen, lineare Algebra zum Berechnen der Persistenz und kombinatorische Optimierung, um die Topologie von zwei Formen effizient zu vergleichen.

Wir stellen ein kürzlich erzieltes Resultat der Gruppe heraus: Die topologische Information einer Form kann durch eine Multimenge von Punkten in der Ebene zusammengefasst werden, das sogenannte „Persistenzdiagramm". Hat man zwei solche Diagramme, liefert ein Distanzmaß zwischen den Punktmengen eine „topologische Distanz" zwischen den beiden zugrundeliegenden Formen. Eine verbreitete Wahl einer Distanz zwischen Punkten ist die „Flaschenhalsdistanz", bei der die Punkte eines Diagramms auf die Punkte des anderen Diagramms abgebildet werden, sodass keine Verbindung zu lange wird. Die Berechnung einer solchen Distanz kann auf ein graphentheoretisches Matchingproblem zurückgeführt werden und zum Beispiel mithilfe des Hopcroft-Karp-Algorithmus gelöst werden. Es ist wohlbekannt, dass unter dem gängigen Berechnungsmodell die geometrische Natur des zugrunde liegenden Problems ausgenutzt werden kann, um die theoretischen Garantien des Algorithmus zu verbessern. Die aktuelle Arbeit der Gruppe zeigt, dass diese Techniken ebenso zu einem sehr viel schnelleren Algorithmus zur Berechnung von Flaschenhalsdistanzen in der Praxis führen. ■



© Kerber – TU Graz

Abbildung 1:
Eine Klein'sche Flasche eingebettet in drei Dimensionen.
Figure 1:
A Klein bottle embedded in three dimensions.

a "topological distance" between the two underlying shapes. A common choice of distance of dots is the "bottleneck distance", where the dots of one shape are matched to the dots of the other, such that no dot-to-dot connection is too large. The problem of computing this distance can be reduced to a graph-theoretic matching problem, and can be solved for instance using the Hopcroft-Karp algorithm. It is well-known that under the common model of computation, the geometric nature of the underlying problem can be exploited to improve the theoretic guarantees of the algorithm. The recent work of the group demonstrates that these techniques also lead to a much faster algorithm to compute bottleneck distances in practice. ■

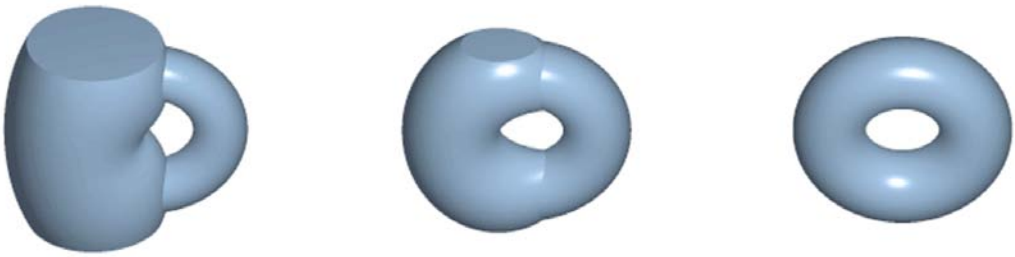


Abbildung 2:
Ein Bagel und eine Kaffeetasse sind unter topologischen Gesichtspunkten nicht unterscheidbar: Beide besitzen einen Tunnel und eine Form kann in die andere überführt werden, ohne den Zusammenhang der Form zu verändern.
Figure 2:
A coffee mug and a bagel are, topologically speaking, the same. Both have one hole, and can be transformed one into the other without changing the connectivity of the shape.