2.6 Bewertung von Methoden zur Ähnlichkeitssuche

– Fragestellung
  - Anfragebearbeitung in metrischen Räumen oder Vektorräumen
  - Gesucht: Feature-Transformation zur Umwandlung komplexer STMM-Objekten in metrische Objekte/Featurevektoren
  - Wie gut drückt die Feature-Transformation die Ähnlichkeit der realen Objekte aus, d.h. wie gut approximiert die Distanz im Feature-Raum die Distanz im Objektraum?
  - Bewertung von Methoden zur Ähnlichkeitssuche („Ähnlichkeitsmodelle“)
    – Testset von Objekten
    – Stelle für alle Objekte des Testsets Ähnlichkeitsanfragen (typischerweise *k*-NN-Queries)
    – Evaluiere das Ergebnis dieser Anfragen

– Objekte mit bekannten Kategorien
  - Objekte sind in Kategorien eingeteilt und entsprechend markiert (z.B. „Schrauben“, „Nägel“, „Bolzen“, …), d.h. Ergebnis der Anfragen ist bekannt
  - Übersicht

	erwünscht	**unerwünscht**
gefunden	richtig positive (rp)	falsch positive (fp)
nicht gefunden	falsch negative (fn)	richtig negative (rn)

  - Recall (Sensitivität): Wie viele der erwünschten Objekte wurden gefunden?

$$\frac{rp}{rp+fn} = \frac{\text{gefundene erwünschte Objekte}}{\text{alle erwünschten Objekte}}$$

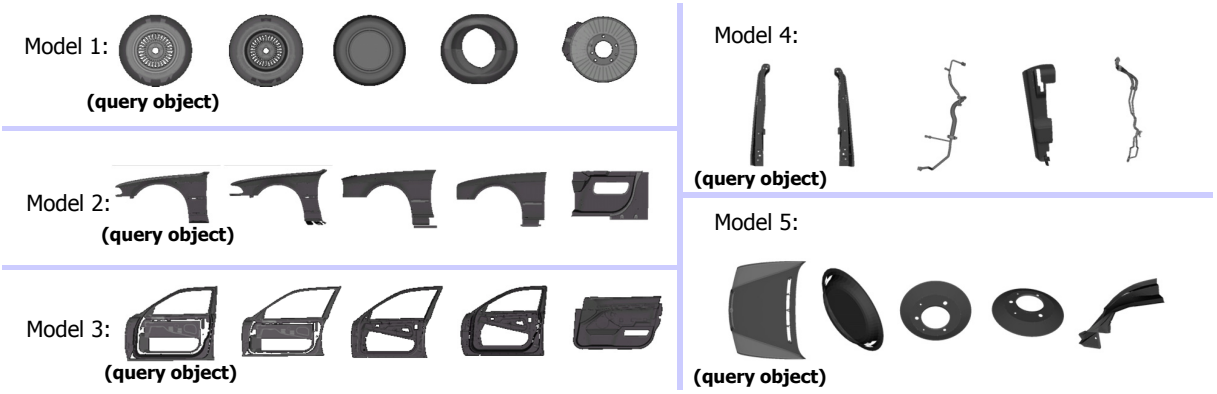
  - Precision: Wie viele der gefundenen Objekte sind erwünscht?

$$\frac{rp}{rp+fp} = \frac{\text{gefundene erwünschte Objekte}}{\text{alle gefundenen Objekte}}$$

  - Spezifität: WS, dass Test für unerwünschtes Obj. negativ verläuft

$$\frac{rn}{rn+fp} = \frac{\text{richtig negativ}}{\text{alle unerwünschten Objekte}}$$

– Objekte mit unbekannten Kategorien
  - Ergebnis der Anfragen ist unbekannt
  - Manuelle Evaluation weniger zufälligen *k*-NN-Queries
  - Problem: Qualität des Modells hängt ab von
    – einer geringen Anzahl von Query-Objekten
      » Besser: möglichst alle Objekte der DB spielen eine Rolle bei der Evaluation
    – der Wahl dieser Query-Objekte
      » Schlechtes Anfrageergebnis für gegebenes *q* bedingt nicht schlechtes Modell
      » Gutes Anfrageergebnis für gegebenes *q* bedingt nicht gutes Modell



- BOSS (Browsing OPTICS-plots for Similarity Search)
  [Brecheisen, Kriegel, Kröger, Pfeifle. Proc. SIAM Int. Conf. Data Mining (SDM), 2004]
  – Idee: benutze Data Mining Methoden
  – Clustering
    » Fasse Objekte in Gruppen zusammen, sodass die Objekte in einer Gruppe (Cluster) ähnlich, Objekte aus verschiedenen Clustern unähnlich sind
    » Hierarchisches Clustering: erstelle eine Hierarchie von ähnlichen Objekten
  – Clustererkennung und Clusterrepräsentation
    » Erkenne automatisch geeignete Cluster in der Hierarchie
    » Stelle jeden Cluster durch einen geeigneten Repräsentanten dar
  – Evaluation/Retrieval
    » Hierarchie von Clusterrepräsentanten ist navigierbar
    » Evaluation der Cluster um Ähnlichkeitsmodell zu evaluieren
    » Ähnlichkeits-basierte Suche nach Objekten ohne konkretes Anfrageobjekt angeben zu müssen