Übungsblatt 2 - Informationsvisualisierung und Visual Analytics

Nick Martin, Tim Lukas

Aufgabe 1: Daten- und Task-Analyse

What? (Datenanalyse)

Bei den gegebenen Rohdaten handelt es sich um eine zweidimensionale Tabelle. Die Items (Zeilen dieser Tabelle) sind sechs individuelle Personen identifiziert von A bis F, wobei jede Person eine eindeutige Entität darstellt. Diese Personen werden durch fünf Attribute (Spalten) beschrieben: dabei handelt es sich um die Programmiersprachen Java, Perl, JavaScript, C#, Python. In den Zellen der Tabelle werden die jeweiligen Werte angegeben, welche die Selbsteinschätzung der jeweiligen Person für eine spezifische Programmiersprache repräsentieren. Diese Werte liegen auf einer Skala von 1 bis 5, bei diesen Werten gilt: Je höher der Wert, desto besser die Selbsteinschätzung in der jeweiligen Programmiersprache.

Attirbut Typen die Attribute der Tabelle entsprechen folgenden Klassifikationen:

- **Personen:** Personen sind kategorische Attribute. Sie haben keine implizite Ordnung und dienen in dieser Tabelle als eindeutiger Schlüssel, um auf die Werte eines Items zuzugreifen.
- **Programmiersprachen:** Auch die Programmiersprachen sind sind kategorische Attribute.
- **Werte der Selbsteinschätzung:** Die Selbsteinschätzungen sind ordinale Attribute. Es existiert eine klare Rangfolge (5 ist besser als 4), die Abstände zwischen den Werten sind jedoch weder quantifizierbar noch notwendigerweise gleichmäßig.

Why? (Task)

Bei der gegeben Fragestellung *"Wer ist der beste und wer der schlechteste Programmierer"* handelt es sich um eine Domänenfrage.
Die Frage ist:

- **Mehrdeutig:** Die Definition von "bester" ist nicht klar. Ist es die Person mit konstant hohen Werten oder eher die Person mit Spitzenwerten in wenigen Kategorien?
- **Unpräzise:** Es wird keine klare Metrik für den Vergleich vorgegeben.
- **Nicht direkt ausführbar:** Die Frage kann nicht ohne weitere Annahmen direkt an die Daten gestellt werden.

Die Frage könnte beispielsweise in folgende Datenfragen übersetzt werden:

"Welche Person hat die höchste Summe an Fähigkeitspunkten über alle fünf Programmiersprachen, und welche die niedrigste?"

Abstraktion der Aufgabe: Das übergeordnete Ziel ist der Vergleich der Programmierfähigkeiten aller Personen, um die Extremwerte zu erkennen. Um dieses Ziel zu erreichen sind folgende Aktionen notwendig:

1. **Zusammenfassen / Agregieren:** Da jede Person durch mehrere Attribute (fünf Programmiersprachen) beschrieben wird, kann kein direkter Verlgeich stattfinden. Zuerst müssen diese verschiedenen Werte für jede Person zu einem einzigen, repräsentativen Gesamtwert zusammengefasst werden.
    - **Beispiel:** Berechnung der Summe oder des Medians der Fähigkeitswerte pro Person.
2. **Ordnen:** Die aus der Aggregation resultierenden Gesamtwerte können dann verwendet werden, um eine Rangliste aller Personen zu erstellen. Diese Aktion bringt die Items(Personen) in eine explizite Reihenfolge.
3. **Suchen/Identifizieren:** Im finalen Schritt müssen die beiden gesuchten Items aus der im vorherigen Schritt erstellten, geordneten Liste extrahiert werden: Die erste Person (der "beste") und die letzte Person (der "schlechteste").

Ergebnis der Task-Abstraktion: Die Aufgabe besteht darin, die Attribute einer Menge von Items (Personen) zu aggregieren, um eine Rangordnung zu erstellen, mit der die Extremwerte identifiziert werden können.

Aufgabe 2 - Datenattribute

Diskussion zwischen den in der Vorlesung vorgestellten und in der Publikation vorgeschlagenen Attribut-Typen

Der primäre Unterschied zwischen der Klassifikation von S. S. Stevens und dem in der Vorlesung behandelten System liegt in ihrer jeweiligen Ausrichtung. Stevens' Gliederung in Nominal-, Ordinal-, Intervall- und Verhältnisskalen ist auf die statistische Analyse ausgelegt. Sie definiert, welche mathematischen Operationen für einen Datentyp zulässig sind, um die statistische Validität gewährleisten zu können. Die in der Vorlesung behandelten Kategorien Kategorial, Geordnet und Quantitativ sind hingegen auf die grafische Umsetzung ausgelegt. Hierbei ist die zentrale Frage, wie die Struktur der Daten am effektivsten und sinngemäß durch visuelle Kanäle (z.B. Farbe, Form und Position) dargestellt werden kann.

Betrachtet man die jeweils vorgeschlagenen Attributstypen fällt auf, dass Stevens' Nominalskala direkt dem in der Vorlesung behandelten kategorialen Attributstyp entspricht. Auch die Ordinalskala entspricht dem geordneten (ordered) Typ.

Die größte Abweichung liegt bei den numerischen Daten. Stevens unterscheidet streng zwischen Intervallskalen, die keinen echten Nullpunkt haben (bspw. Celsius), und Verhältnisskalen mit einem absoluten Nullpunkt (z.B. Größe). Diese Trennung ist statistisch

entscheidend, da Verhältnisse nur bei einem absoluten Nullpunkt sinnvoll sind. Das in der Vorlesung behandelte Modell fasst beide Typen zu einem einzigen quantitativen Typ zusammen, dies hat den Grund, dass für die grafische Umsetzung die Art des Nullpunkts meist keine Rolle spielt. Sowohl Intervall- als auch Verhältnisskalen werden typischerweise durch dieselben visuellen Kanäle, wie die Position auf einer Achse oder die Balkenlänge, abgebildet.

Abschließend lässt sich sagen, dass das in der Vorlesung behandelte Modell eine für den Anwendungsfall der Informationsvisualisierung optimierte Abstraktion ist. Es vereinfacht Stevens' strenge statistische Regeln zu einem Modell, das Design-Entscheidungen für die effektive visuelle Repräsentation von Daten anleitet.