Visualisierung

Assignment 1

Aufgabenstellung

Finden von je einem Beispiel von guter beziehungsweise schlechter Visualisierung von Daten im wissenschaftlichen Bereich und anschließender Besprechung der gegebenen Beispiele.

Beispiele

Folgende Beispiele wurden ausgewählt:

Gute Visualisierung von Daten

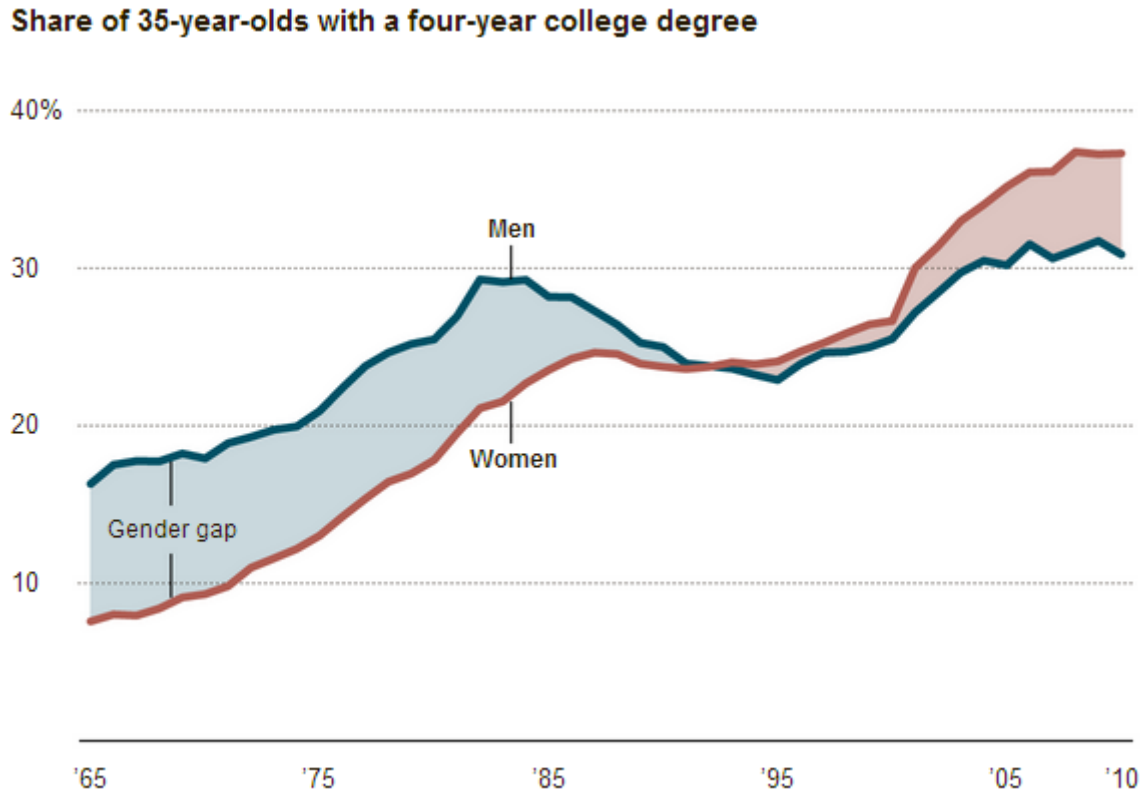

Abbildung 1: Prozentsatz an Absolventen eines vierjährigen Colleges in den USA

Beschreibung des Graphen

Der Graph zeigt den Prozentsatz an Frauen und Männern in den USA, welche mit 35 Jahren einen vierjährigen Collegeabschluss absolviert haben. Wie die Statistik zeigt, wurden die Männer, welche in früheren Jahren deutlich in "Führung" lagen, dabei in den letzten Jahren - speziell zwischen 2005 und 2010 - stark von den Frauen "überholt". Während es in den 1970er Jahren für Frauen wenig üblich war, einen Collegeabschluss zu erreichen (unter 10 Prozent), war dieser "gender gap" zwischen 1990 und 2000 relativ ausgeglichen (die Prozentzahl der Frauen

mit Abschluss war aber auch da schon etwas höher), im letzten Jahrzehnt stieg die Anzahl der weiblichen Absolventinnen mit 35 Jahren aber deutlich stärker an. Außerdem fällt auf, dass der Prozentsatz an Männern mit Abschluss zwischen 1985 und 1995 um knapp 5% zurück ging, bei den Frauen aber ein stetiger Anstieg zu vermerken war.

Bewertung des Graphen

Das gegebene Diagramm aus der *New York Times* wurde ausgewählt, da es die zu präsentierenden Daten sehr schlicht und übersichtlich darstellt, was dazu führt, dass sie schnell und einfach vom Leser interpretiert werden können. Bei der Erstellung des Graphen wurde dabei auf wichtige Design-Prinzipien wie beispielsweise einer maximale Data-Ink Ration, der Vermeidung von Chartjunk und der Einhaltung einer nicht zu großen Datendichte geachtet. Auch die Beschriftung der Achsen sowie der beiden Linien des Graphen tragen zur Übersichtlichkeit bei, die gepunkteten "Hilfslinien" helfen dem Betrachter, die Prozentzahlen besser einschätzen zu können, ohne dabei den eigentlich Graphen zu stören. Auch die Färbung der Fläche zwischen den beiden Linien verdeutlicht die repräsentierten Daten beziehungsweise deren Unterschiede. Insgesamt bietet der Graph eine sehr gelungene Präsentation der zugrunde liegenden Daten.

Schlechte Visualisierung von Daten

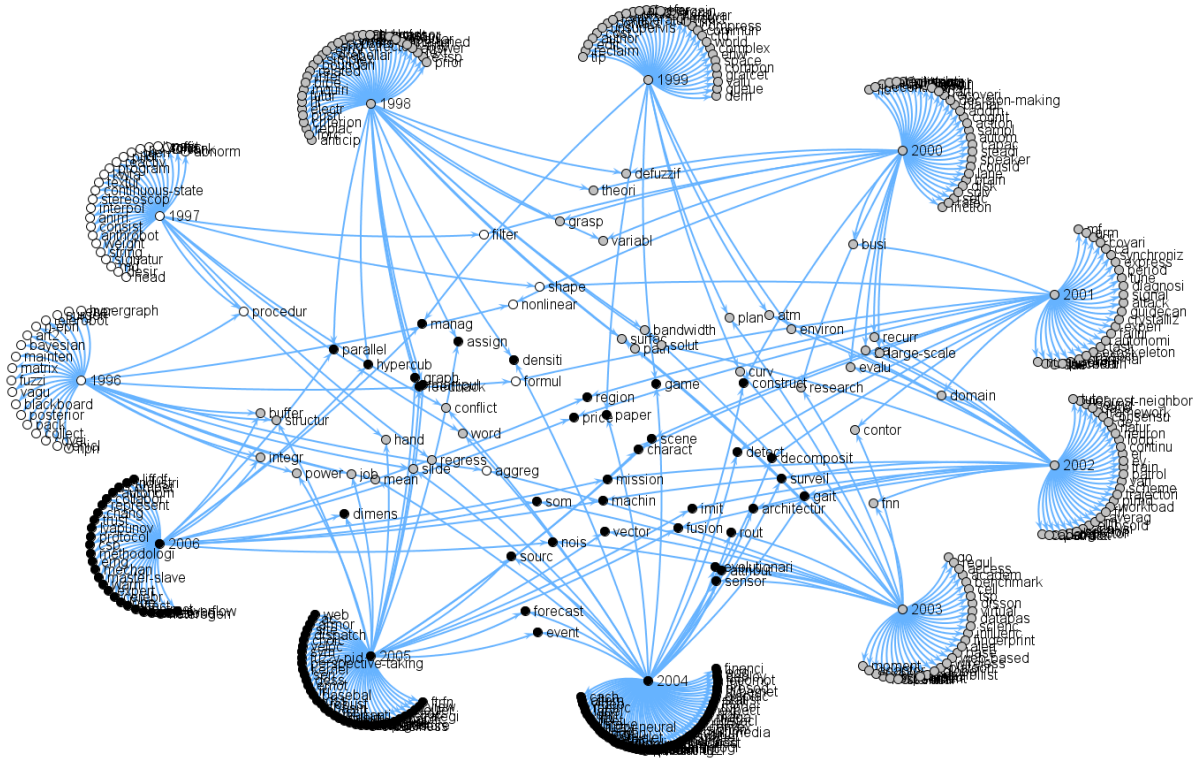

Abbildung 2: Veränderung der Themen in Abstracts der IEEE SMC von 1996 bis 2006

Beschreibung des Graphen

Der zweite gewählte Graph repräsentiert die Veränderung der Themengebiete in den Abstracts der IEEE Systems, Man & Cybernectics Society zwischen den Jahren 1996 und 2006. Dafür wurden die Abstracts der Publikationen ("IEEE SMC A, B and C transactions") eingelesen und tokenisiert. Mit Hilfe von "part-of-

speech" Tags und dem Porter Stemming Algorithmus sowie der "term frequency-inverse document frequency" wurden die wichtigsten Begriffe (meist 2 pro Publikation) ausgelesen und gesammelt. Diese Begriffe wurden dann in einem Graphen dargestellt, wobei eine Untergliederung in die verschiedenen Jahre vorgenommen wurde: alle Begriffe, die nur in einem bestimmten Jahr aufzufinden waren, sind nur zu dem Knoten des jeweiligen Jahres verbunden. Wurde ein Ausdruck aber in Publications aus mehreren Jahren gefunden, so wird er in der Mitte mit Verknüpfung zu allen betroffenen Jahren dargestellt.

Bewertung des Graphen

Da die Anzahl der ausgewerteten Begriffe riesig war, wurde eine Reduzierung auf rund 5000 Stichworte durchgeführt. Trotz dieser Reduzierung ist die Menge der dargestellten Knoten und Kanten noch immer immens, was es sehr schwer bis unmöglich macht, Informationen aus diesem Graphen zu gewinnen. Zwar bietet der Autor viele Informationen auf einmal an und stellt auch die Verknüpfungen zwischen diesen her, jedoch geschieht dies auf Kosten der Übersichtlichkeit und Lesbarkeit, was im Endeffekt auch zu der Bewertung als schlechte Repräsentation von Daten führte. Zwar bleibt die grafische Integrität der Daten bewahrt (es erfolgt keine "Verfälschung" durch z.B. unterschiedliche Größen oder Skalen), jedoch ist die "data density" sehr hoch, da sehr viele Informationen auf kleinem Raum dargestellt werden müssen. Da es schwer möglich sein wird, eine solch große Datenmenge effektiv in einem geeigneten Graphen darzustellen, wäre es wahrscheinlich angebracht, die Daten entweder noch weiter zu reduzieren (was im Gegenzug aber eine ungenauere Darstellung bedeutet, da Informationen ausgelassen werden) oder den Graphen in mehrere "Subgraphen" (beispielsweise pro Jahr) aufzuteilen und nur die gemeinsamen Informationen möglich übersichtlich darzustellen.

Fazit

Zwar sind die Datenmengen der beiden ausgewählten Graphen wahrscheinlich sehr unterschiedlich, die Art der Visualisierung ist aber in der ersten Grafik deutlich besser gewählt und ermöglicht es dem Betrachter sehr viel einfacher, die präsentierten Informationen zu erfassen und zu verarbeiten. Wird die Anzahl der zu visualisierenden Daten zu groß, muss deswegen eine geeignete Möglichkeit zur Darstellung gefunden werden (dies kann auch durch Reduzierung beziehungsweise Aufteilung der Ausgangsdaten erfolgen). Wird trotzdem versucht, möglichst viel Information auf kleinem Raum darzustellen, so kann dies schnell "ausarten" und unübersichtlich werden.

Quellenangaben

Abbildung 1: The New York Times, http://www.nytimes.com/interactive/2013/03/20/business/diverging-fortunes-for-men-and-women.html?ref=multimedia (22.03.2012 - 16:44)

Abbildung 2: Towards Visual Exploration of Topic Shifts, http://www.inf.uni-konstanz.de/bioml2/publications/Papers2007/ThDKB07_TowVisExplOfTopicShifts.pdf (22.03.2012 - 16:44)