Master-Entwicklungsprojekt „Visualisieren und Auswerten großer Datenmengen"

UNIVERSITY OF APPLIED SCIENCES
FACHHOCHSCHULE GIESSEN FRIEDBERG

Anwendungsdomäne: Media Systems

Erzeugung eines Dotplots

Das Eclipse-Plugin „Dotplot" stellt mit einer aus der Genetik stammenden Methode Gemeinsamkeiten einer Menge von Zeichenketten, Wörtern, Wortsequenzen oder Sätzen - allgemein „Tokens" genannt - grafisch dar. Bild 1 zeigt anhand diverser Beispiele, wie ein Quelltext auf zwei Achsen einer Matrix verteilt wird und Matches (gleiche Tokens) als Punkt markiert werden. Spezifische Muster im Dotplot lassen den erfahrenen Betrachter schnell auf die Art und den Grad von Ähnlichkeiten schließen. Bild 2 zeigt einen Dotplot zum Text aus Bild 3. Häufig auftretende Muster sind dort synthetisch erzeugt worden. Bild 4 ist eine Selbstanwendung des Dotplot-Plugins auf seine Sourcen. Die Farbskala unter den Bildern 2 und 4 bildet die Gewichte von Matches auf Farben im Dotplot ab. Man kann so die seltenen Treffer von häufigen besser unterscheiden.

Projektziel

- ☐ Open-Source-Produkt für die Dotplot-Erzeugung aus beliebigen Textsorten
- ☐ Automatische Plagiaterkennung
- ☐ Refactoring von Programmen
- ☐ Autorenstile identifizieren: „Ein echter Shakespeare?"

Methode

- ☐ Grafisches Dotplot-Verfahren aus der Genforschung
- ☐ Interaktive Muster-Erkennung
- ☐ Sequenz-Alignment

Softwaretechnik

- ☐ Open-Source-Projekt (GNU GPL)
- ☐ Java-Implementierung als Eclipse-Plugin
- ☐ Kooperationsplattform SourceForge
- ☐ Vorgehensweise: Extreme Programming (XP)

Features

- ✓ Berechnung im Grid
- ✓ Information-Mural-Algorithmus (verlustarme Interpolation)
- ✓ Dotplot-Perspektive in Eclipse
- ✓ Export in diversen Dateiformaten (JPEG, PNG, PDF)
- ✓ PDF-Konverter und Inputfilter für Java, C++, PHP, ...

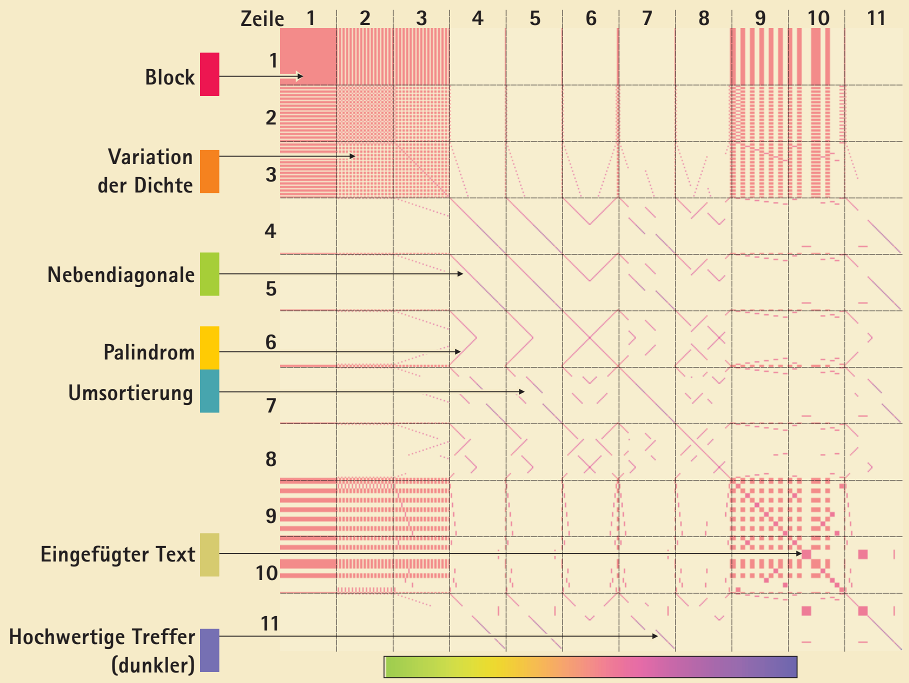

Bild 2: Dotplot (per JAI) der Tokens aus Bild 3. Farben und Zeilenangaben entsprechen den Markierungen im Quelltext. Die synthetisch erzeugten Muster sind mit den entsprechenden Bezeichnungen gekennzeichnet.

```
1  a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a
2  a a b a a b a a b a a b a a b a a b a a b a a b a a b a a b a a b a a b a a b a a b a a b a a b
3  a a b a a c a a d a a e a a f a a g a a h a a i a a j a a k a a l a a m a a n a a o a a p a a q
4  a b c d e f g h i j k l m n o p q r s t u v w x y z 0 1 2 3 4 5 6 7 8 9 ä ö ü ß + - * / = ; : #
5  a b c d e f g h i j k l m n o p q r s t u v w x y z 0 1 2 3 4 5 6 7 8 9 ä ö ü ß + - * / = ; : #
6  a b c d e f g h i j k l m n o p q r s t u v w x x w v u t s r q p o n m l k j i h g f e d c b a
7  a b c d e f g h i s t u v w x y z 0 1 2 3 4 5 j k l m n o p q r 6 7 8 9 ä ö ü ß + - * / = ; : #
8  a b c d e f g h i j k l m n s r q p o n m l k j i h g f o p q r s t u v w x x w v u t e d c b a
9  a a a a b b b b a a a a c c c c a a a a d d d d a a a a e e e e a a a a f f f f a a a a g g g g
10 a a a a c c c c a a a a * * * * * * * * a a a a a a a a f f f f a a a a e e e e g g g g b b b b
11 a b c d e f g h i j k l * * * * * * * * u v w x y z 0 1 2 3 4 5 6 7 8 9 ä ö ü ß + - * / = ; : #
```

Bild 3: Datenquelle für den in Bild 2 abgebildeten Plot

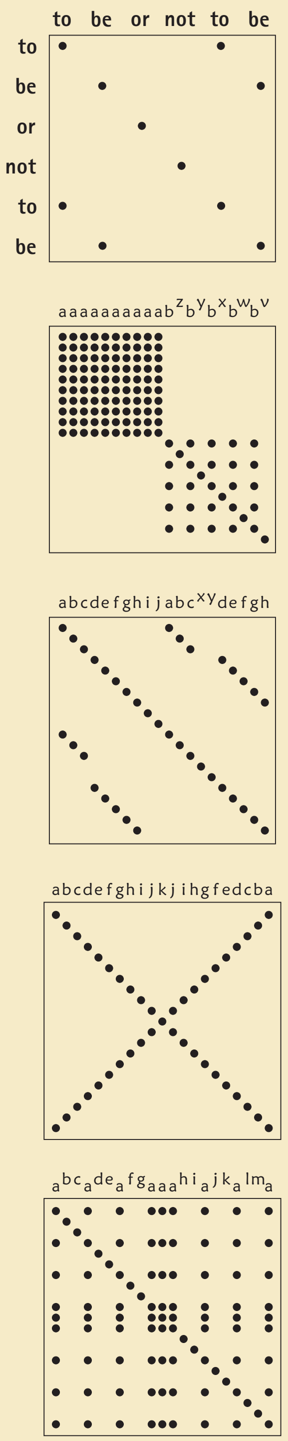

Bild 1: Beispiele zur Entstehung prinzipieller Muster im Dotplot. Von oben nach unten:
- Wörter als Tokens
- Block mit variierender Dichte
- Unterbrochene Nebendiagonalen
- Palindrom
- Dunkles Kreuz

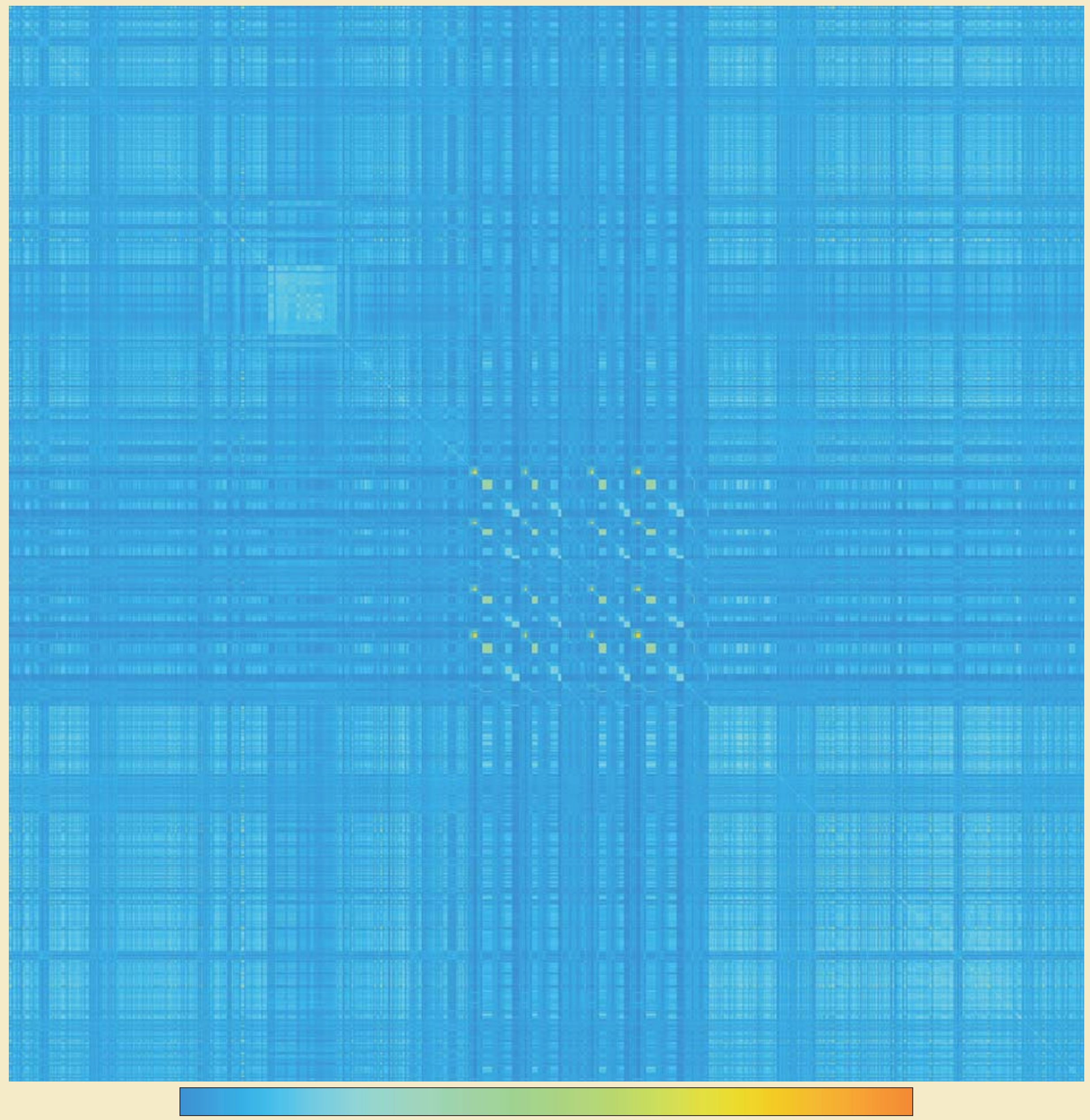

Bild 4: Dotplot (Information Mural) der Java-Sourcen des Eclipse-Plugins: 175 Klassen, 30.000 Zeilen Code

ToDos

- ☐ Automatische Wortstammreduktion und Satzende-Erkennung bei natürlichen Sprachen
- ☐ Implementierung als Rich-Client-Plattform (RCP)
- ☐ Information Mural als Navigationshilfe

Herausforderungen

- ☐ Performancesteigerung
- ☐ Webinterface für Online-Service
- ☐ Visualisierung multimedialer Daten

Projektleiter: Prof. Dr. Klaus Quibeldey-Cirkel
Tutor: cand. ing. Sascha Hemminger
SourceForge-Admin: Dipl.-Inform. (FH) Tobias Gesellchen

Fachbereich MNI
Mathematik, Naturwissenschaften und Informatik

- Projekt-Homepage: http://www.dotplot.org/
- eStudy-Masterkurs „DotPlot": http://estudy.mni.fh-giessen.de/
- Sourceforge-Projekt: http://sourceforge.net/projects/dotplot/