Analyse von Verfahren zur Effizienzsteigerung von multidimensionaler Skalierung

Anita Schilling

Web Technology and Information Systems
Bauhaus-Universität Weimar

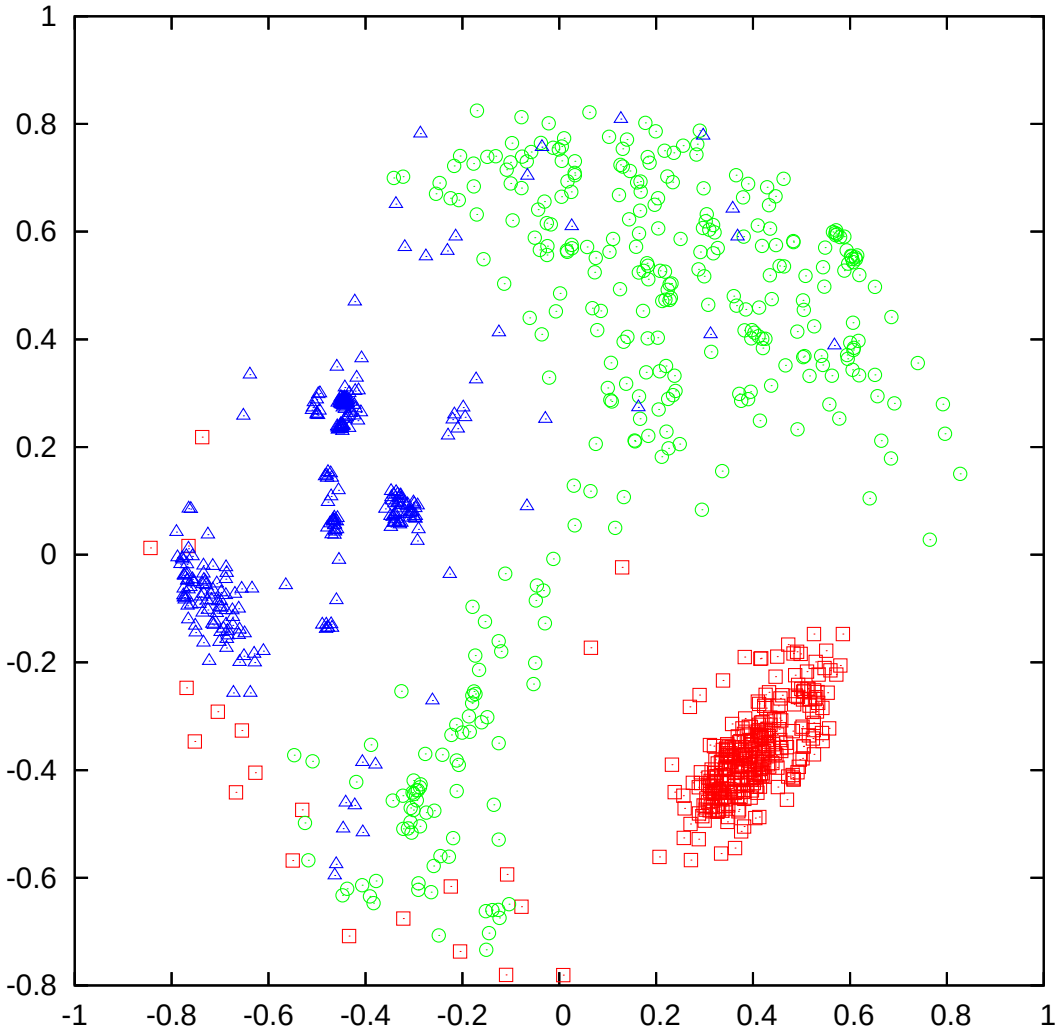
1. März 2007

Outline

Motivation

- Visualisierung von hochdimensionalen Daten, die keine unmittelbare Entsprechung im 2- oder 3-dimensionalen Raum haben
- Zusammenfassung von Methoden zur Dimensionsreduktion unter *multidimensionaler Skalierung*

Visualisierung

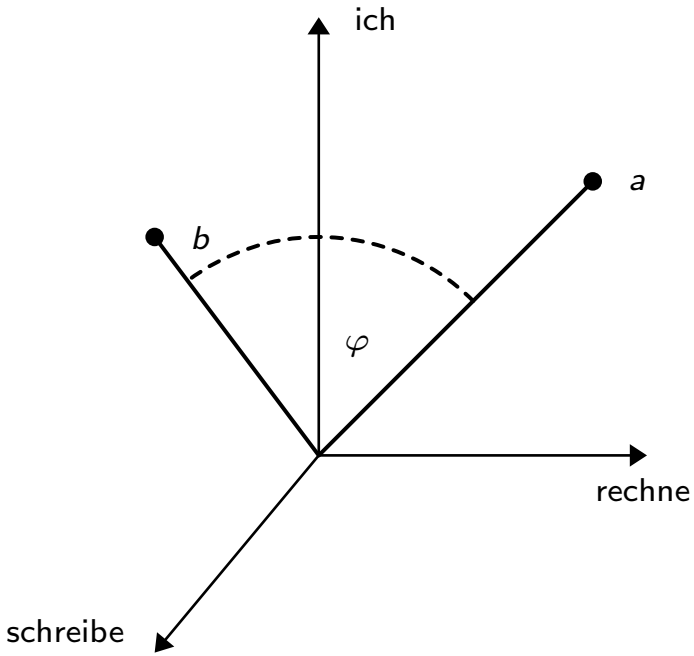


Definition

- Repräsentation der Distanzen δ zwischen Objektpaaren (a, b) als Distanzen Δ zwischen Punkten (p_a, p_b) in einer niedrigeren Dimension

$$\delta(a, b) - \Delta(p_a, p_b) \rightarrow 0 \quad \forall a, b \in D,\ |D| = N$$

Vektorraummodell



$a =$ "ich rechne"
$b =$ "ich schreibe"

$w_0 =$ ich
$w_1 =$ rechne
$w_2 =$ schreibe

$a \longrightarrow \mathbf{a} = (1, 1, 0)$
$b \longrightarrow \mathbf{b} = (1, 0, 1)$

$1 - cos(\varphi) = 0,5$

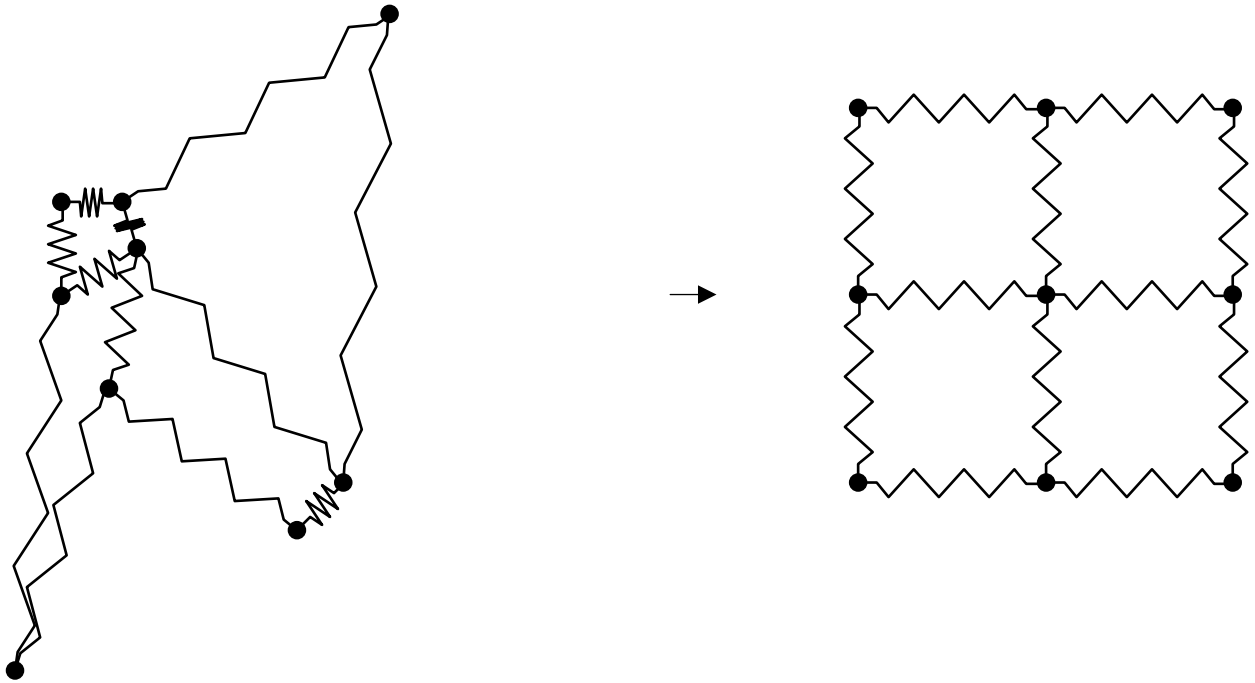
- Repräsentation eines Dokuments als Vektor
- Dimension des Vektors ist die Anzahl von Indextermen in der Dokumentekollektion
- Elemente des Vektors sind die Wichtigkeiten der Indexterme in dem Dokument
- Kosinus-Unähnlichkeit als Distanz zwischen zwei Vektoren

Bewertungsmaß der Punktkonfiguration

- *Stress* bezeichnet den Fehler, den die generierte Punktkonfiguration enthält
- Verfahren sollten Punktkonfigurationen mit möglichst geringem Stress für alle Ausgangdaten generieren

$$\sigma = \sqrt{\frac{\sum_{a,b \in D} \left(\delta(a,b) - \Delta(p_a, p_b)\right)^2}{\sum_{a,b \in D} \left(\Delta(p_a, p_b)\right)^2}}$$

Spring-Modell



Feder zwischen zwei Punkten im Ruhezustand

Spring-Modell

Spring-Modell

Kräfteberechnung zwischen allen Nachbarn
$O(N^3)$

Verfahren von Chalmers (1996)

Kräfteberechnung zwischen einer Stichprobe von Nachbarn
$O(N^2)$

Hybride Verfahren

1. Generierung einer Punktkonfiguration für eine Teilmenge S der Größe $\sqrt{N}$ mit dem Verfahren von Chalmers (1996) $O(N)$
2. Für alle übrigen Objekte i der Datenmenge
   1. Suche des nächsten Nachbarn zu i in der Teilmenge S
   2. Interpolation von Punktkoordinaten für i auf Basis des nächsten Nachbarn und der Teilmenge S $O(N\sqrt{N})$
3. Verfeinerung der Punktkonfiguration mit dem Verfahren von Chalmers (1996) für die gesamte Datenmenge $O(N)$

Betrachtete Verfahren

Verfahren von Chalmers u.a. (2003)

lineare Nächste-Nachbar-Suche

$O(N\sqrt{N})$

Verfahren von Jourdan und Melançon (2004)

Nächste-Nachbar-Suche mit geordneten Listen

$O(N^{5/4}log(N))$

Multiscale-Verfahren von Jourdan und Melançon (2004)

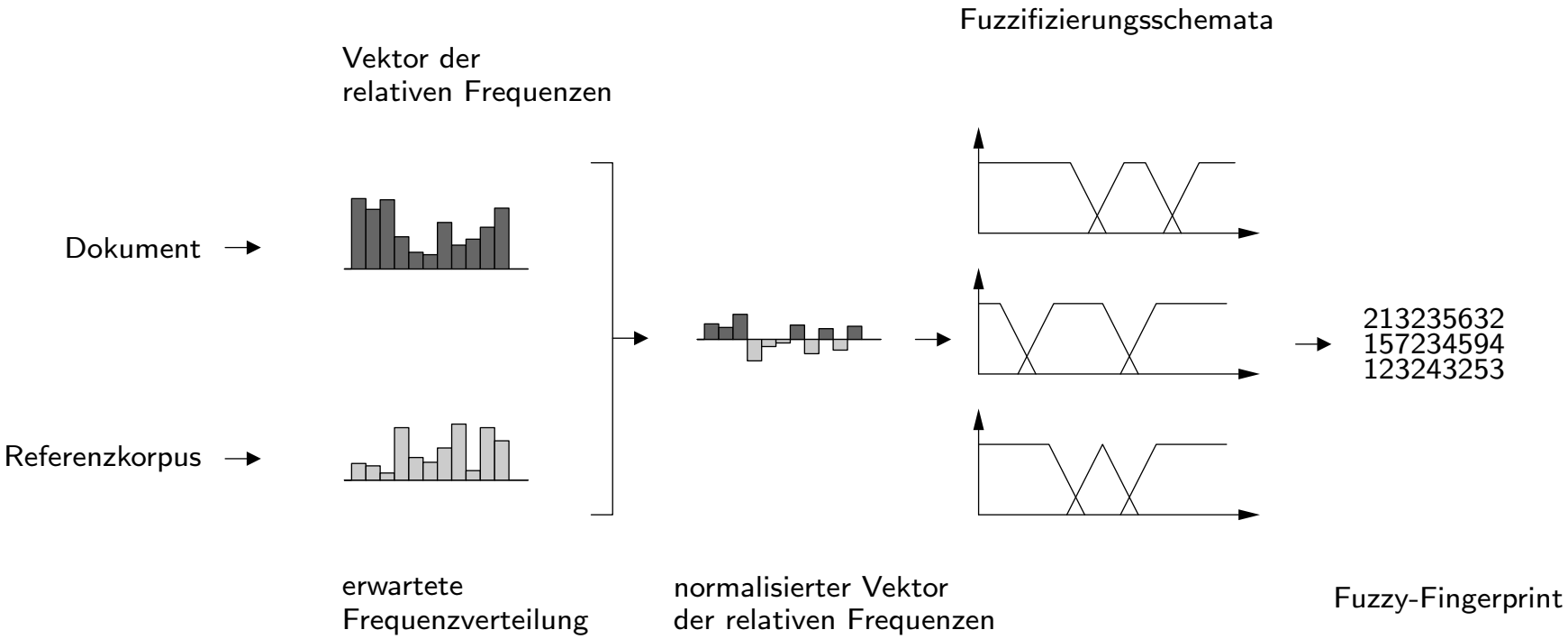
Generierung der Teilmengenkonfiguration durch rekursive Anwendung des Verfahrens

$O(N^{5/4}log(N))$

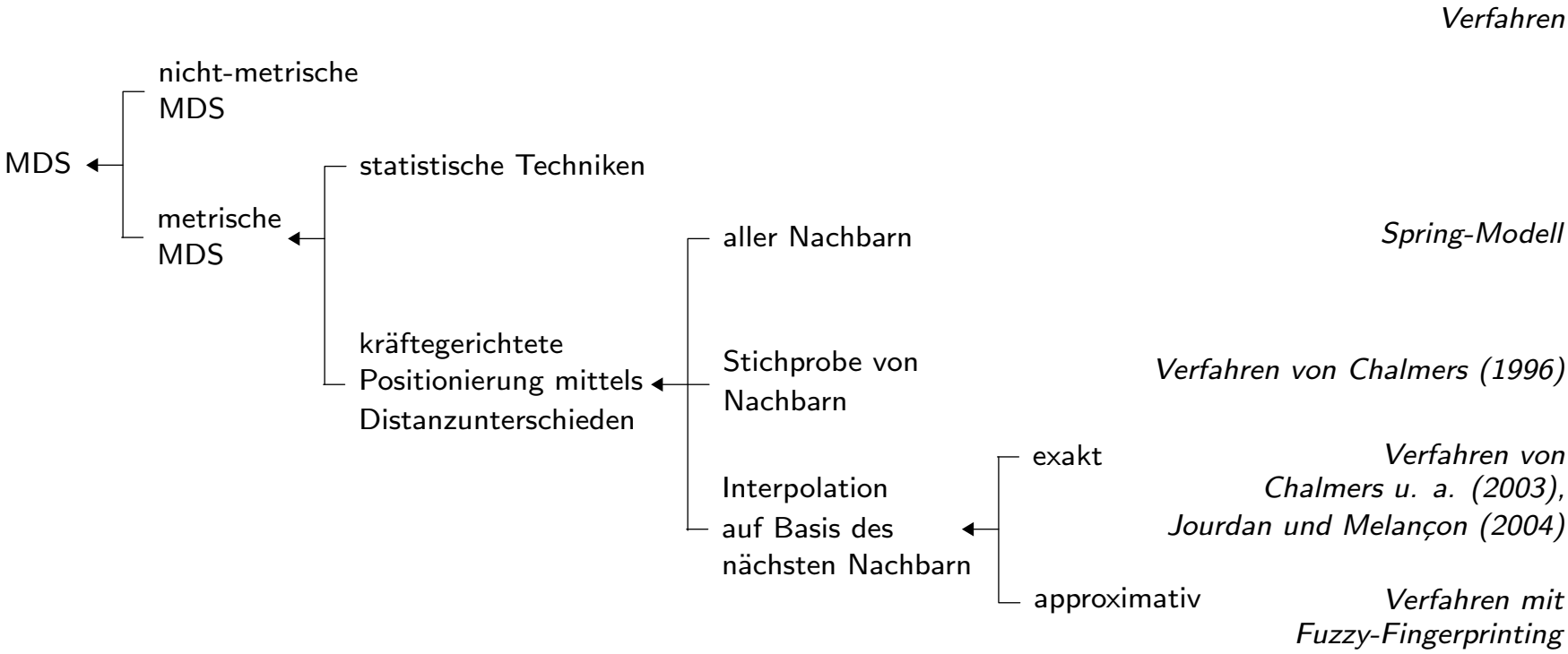
Verfahren mit Fuzzy-Fingerprinting

- Variante des Similarity-Hashings
- Hashkollisionen als Indikator für Ähnlichkeit zwischen Objekten
- Nächste-Nachbar-Suche mit Hashing in $O(N)$

Fuzzy-Fingerprinting



Taxonomie der Verfahren

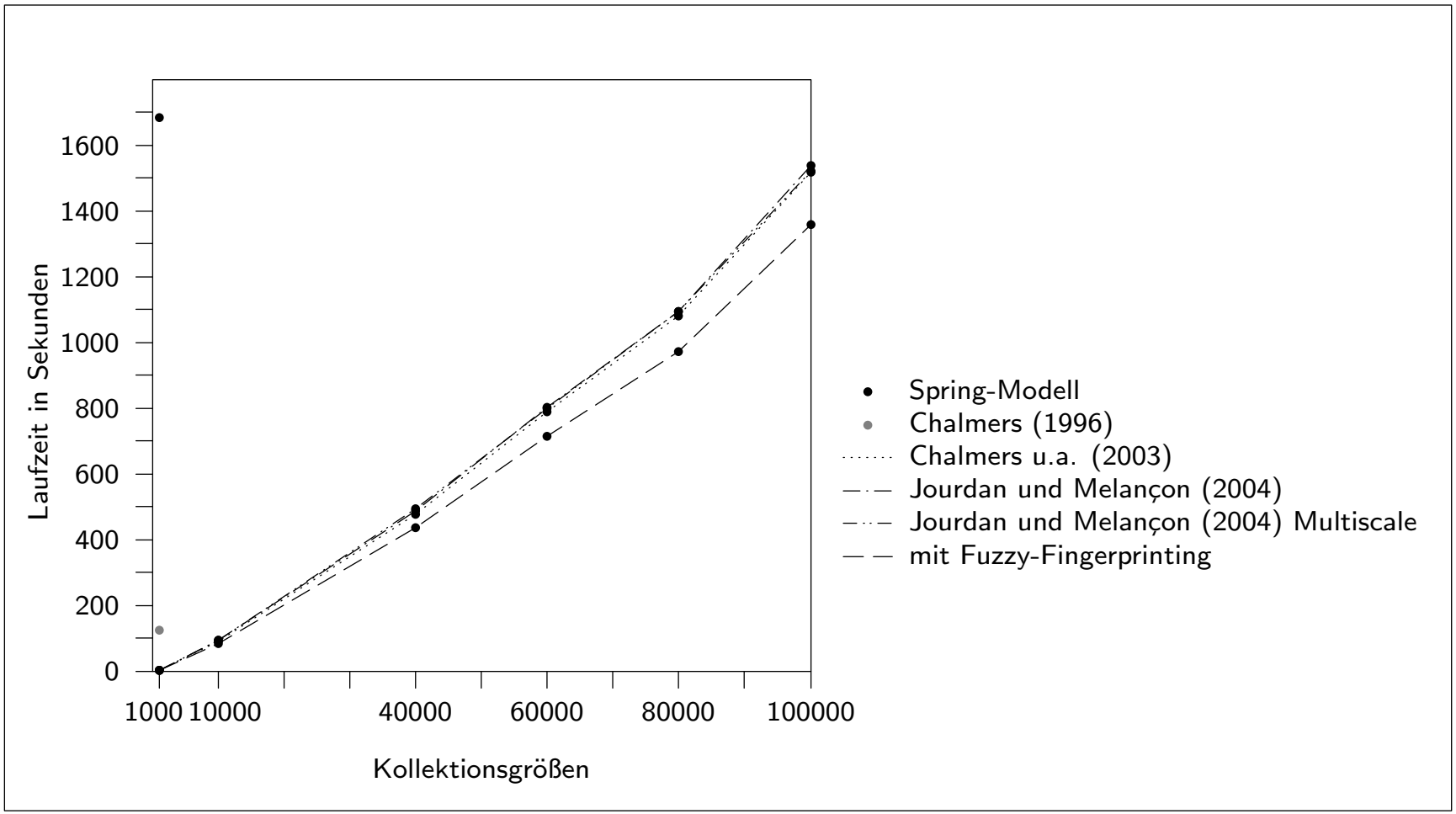


Testbedingungen

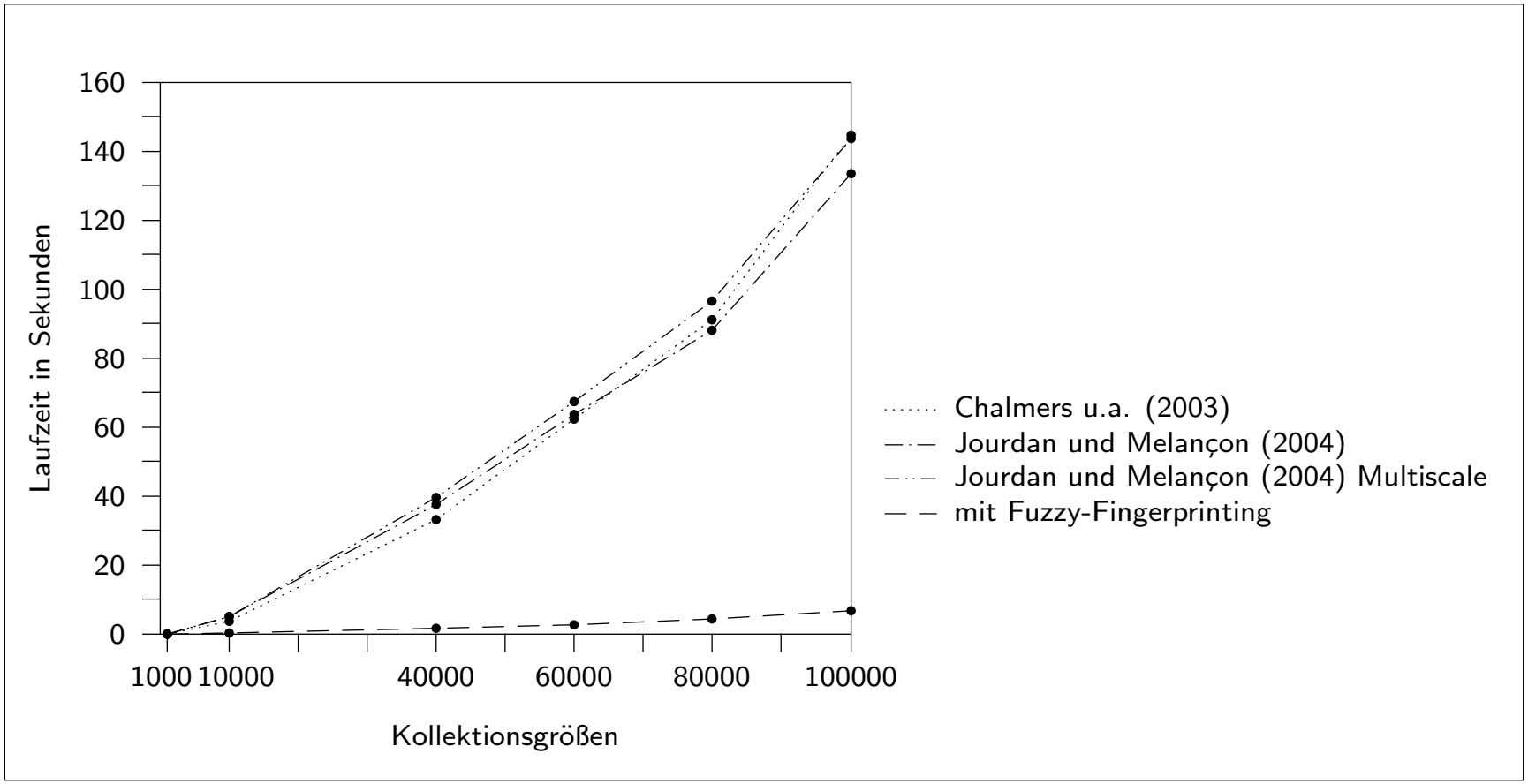
Dokumentenkollektionen	
Größe:	**Dimension:**
1 000	6 992
10 000	27 612
40 000	52 420
60 000	105 976
80 000	94 798
100 000	124 093

- 6 Dokumentenkollektionen aus dem Reuters-Korpus
- 6 Durchläufe je Verfahren

Gesamtlaufzeiten der Verfahren



Laufzeiten der Nächsten-Nachbar-Suche



Plots

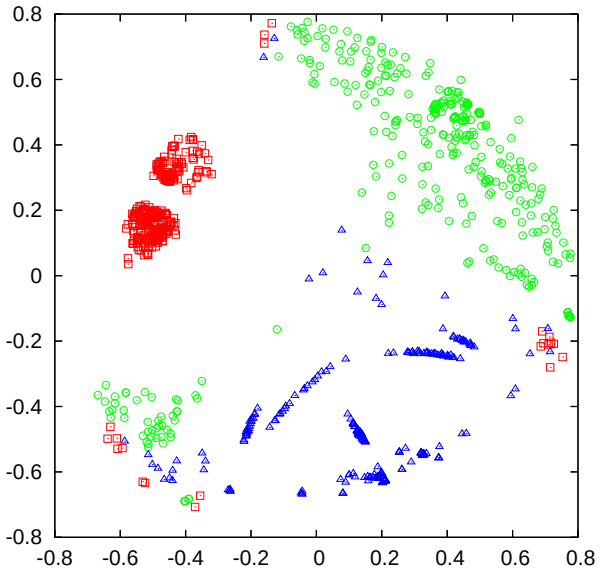


Abb. Spring-Modell

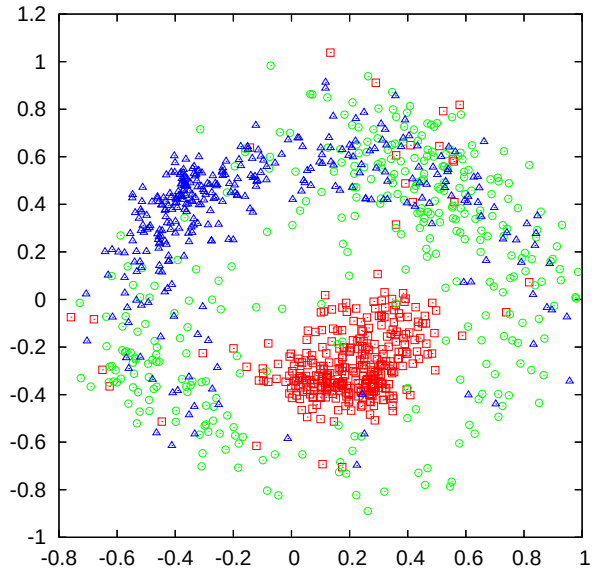
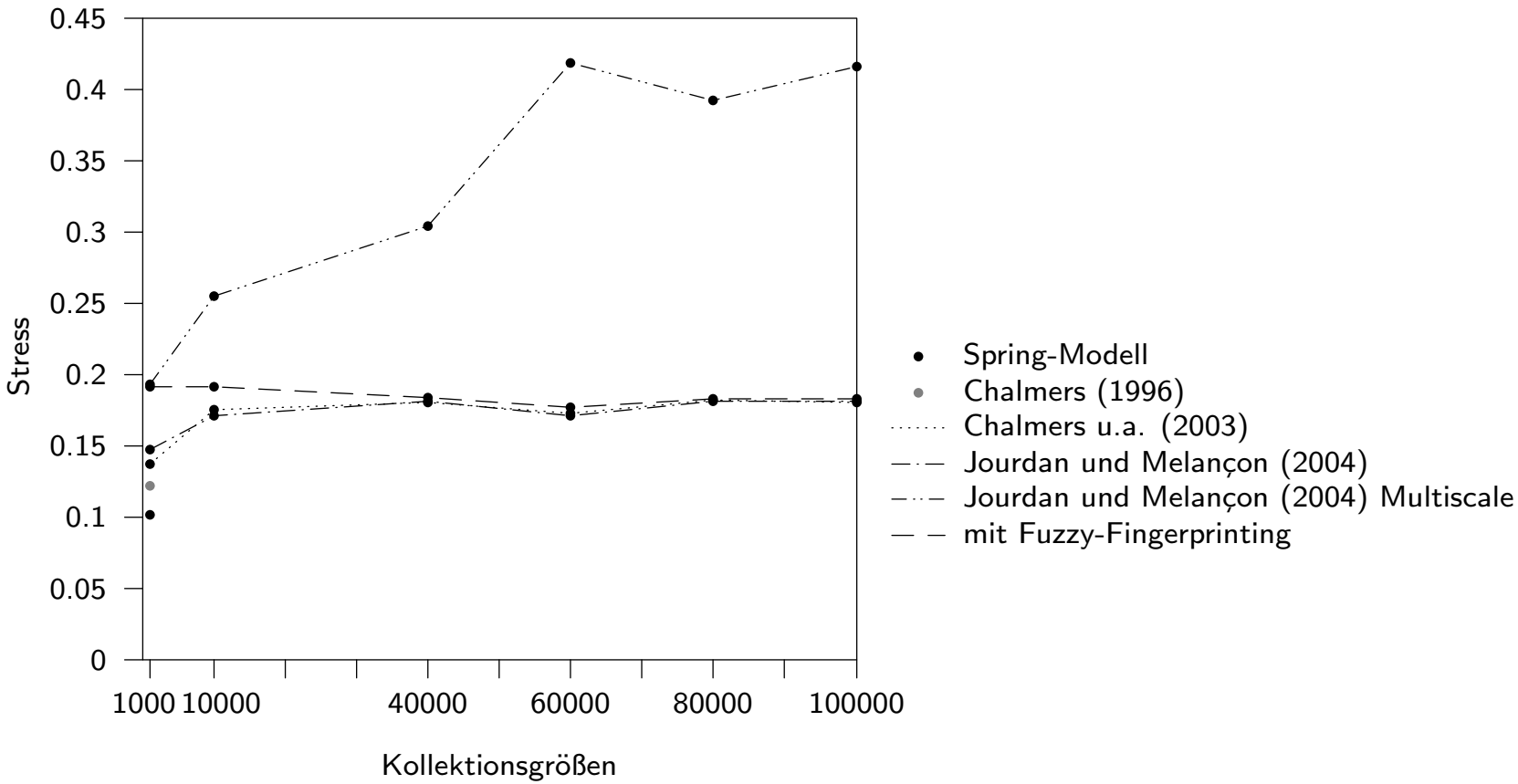


Abb. Verfahren mit Fuzzy-Fingerprinting

Kollektion mit 1 000 Dokumenten

Stresswerte der Punktkonfigurationen



Zusammenfassung

- Verfahren mit Fuzzy-Fingerprinting ist effizientestes Verfahren hinsichtlich Laufzeit und Stresswert
- Interpolation ist Ansatzpunkt für weitere Optimierung