Hochdimensionale Indexierung: Formale Grundlagen und neue Ansätze

Sören Balko

UMIT Innsbruck
Soeren.Balko@umit.at

Abstract: Multimediale Dokumentenkollektionen finden in vielen IT-Bereichen zunehmende Verbreitung. Die Handhabung großer Datenbestände erfordert effiziente Suchoperationen, die es erlauben, Medienobjekte in akzeptablen Zeiten aufzufinden. Darunter fallen auch inhaltsbasierte Anfragen. Häufig werden charakteristische Eigenschaften auf mehrdimensionale Merkmalsvektoren abgebildet, deren Distanz als ein Maß für die (Un-)Ähnlichkeit der repräsentierten Medienobjekte aufgefasst wird. Im Datenbankenkontext bilden geeignete Indexstrukturen und Suchalgorithmen die elementare Voraussetzung für eine effiziente Anfragebearbeitung.

In diesem Beitrag stellen wir Ergebnisse aus [Bal04] dar und beschäftigen uns mit (1) den formalen Grundlagen hochdimensionaler Indexierung, (2) der Einführung eines adaptiven Indexierungsschemas, (3) Fragen des Index-Tunings auf der Grundlage eines analytischen Kostenmodells und (4) dem experimentellen Vergleich konkurrierender Indexierungsvorschläge. Im Vorgriff auf die folgende Darstellung dieser Beiträge ist es gelungen, eine neue Indexierungsmethode zu entwickeln, die bei der Anfragebearbeitung deutliche Kostenvorteile gegenüber bestehenden Ansätzen erzielt.

1 Einleitung

Der Redakteur eines Reisemagazins möchte kurz vor Redaktionsschluss noch einige typische Landschaftsfotos in seinen Artikel über die Kanaren einfügen. Dem Journalisten steht ein großes Bildarchiv seines Verlagshauses zur Verfügung, aus dessen Bestand er sich frei bedienen kann. Aus Kostengründen ist keines der Fotos mit Stichworten annotiert, so dass inhaltsbasierte Ähnlichkeitsanfragen als einzige Suchfunktionalität verfügbar sind. In diesem Anwendungsbeispiel kann der Redakteur auf eigene Urlaubsfotos zurückgreifen, die er als Anfrageobjekte für die Suche verwendet. Mit großen Erwartungen startet er die Anfrage, welche die 30 Fotos mit der ähnlichsten Farbverteilung aus einem Datenbestand von 10 Millionen Bildern ermitteln soll. Zum Redaktionsschluss hat das System noch keine Ergebnisse geliefert, so dass der Autor seinen Artikel frustriert ohne Fotos in den Druck geben muss.

Obwohl ein großer Datenbestand verfügbar ist, der die gesuchten Dokumente mit hoher Wahrscheinlichkeit enthält und eine geeignete Anfrageschnittstelle bereitstellt, erweist sich die Suche aufgrund inakzeptabler Antwortzeiten als unpraktikabel. Offenbar verzichtet das verwendete System auf eine Indexunterstützung für die eingesetzten mehrdimensio-

nalen Farbmerkmale. Damit muss die Anfragebearbeitung den kompletten Datenbestand sequenziell durchlaufen, um die 30 ähnlichsten Bilder anhand der geringsten Distanzen ihrer Feature-Vektoren zu einem Anfragevektor zu bestimmen (*SCAN*). Formal handelt es sich um eine Nearest-Neighbor-Anfrage unter den Feature-Vektoren, die als Punkte im hochdimensionalen Raum angesehen werden können.

Die Indexunterstützung für Nearest-Neighbor-Anfragen im hochdimensionalen Raum erfreut sich seit langem beträchtlicher Forschungsbemühungen. Erste Ansätze adaptierten räumliche Zugriffsmethoden, die Punkte in hierarchisch angeordneten Clustern versammeln (siehe [BBK01] für einen Überblick). Die Anfragebearbeitung [HS95] geschieht im günstigsten Fall in logarithmischer Zeit bezüglich der Datenbankgröße N. Für die meisten Datenverteilungen und hohe Dimensionalitäten ($d \gtrapprox 15$) führt der Einsatz hierarchischer Indexstrukturen jedoch zu weitaus höheren Antwortzeiten. Dieses unbefriedigende Ergebnis ist auf eine Reihe formal-stochastischer Effekte im hochdimensionalen Raum zurückzuführen, die im Formalteil [Bal04] einer tief greifenden Analyse unterzogen wurden. So konnte nachgewiesen werden, dass der Erwartungswert und die Standardabweichung Euklidischer Distanzen mit steigender Dimensionalität d konvergieren: $\mu_{\|v-v\|_2} \rightarrow_p \sqrt{d/6}$ und $\sigma_{\|v-v\|_2} \rightarrow_p$ const. Das Wachstum des Erwartungswertes führt bei unveränderter Standardabweichung zur relativen Annäherung der Distanzen. Mit anderen Worten existieren also praktisch keine Punkte, die dicht beieinander liegen. Damit verbleiben für die Indexierung zwei Möglichkeiten: (1) die Cluster-Bildung trotz nahezu gleicher hoher Punktabstände, die mit einem steigendem Approximationsfehler[1] einhergeht oder (2) die Aufnahme sehr weniger Punkte in den Cluster, was zu einem Missverhältnis aus Cluster-Beschreibungskosten und Speicheraufwand der enthaltenen Punkte führt. Beide Entscheidungen lassen die Anfragekosten letztlich derart ansteigen, dass sie sogar über denen des Naivverfahrens *SCAN* liegen.

Auf der Grundlage dieser Erkenntnisse führen Signaturverfahren ein konsequent verändertes Indexierungsprinzip ein. Darin werden den Punkten kompakte, aber möglichst präzise approximative Darstellungen (Signaturen) zugeordnet. Diese Signaturen werden in einer flachen Datei sequenziell abgelegt. Die Anfragebearbeitung setzt auf einem zweistufigen *filter-and-refine*-Verfahren auf. Die Signaturen ermöglichen die Berechnung unterer und oberer Schranken der Distanzen zum Anfragepunkt. Anhand dieser Schranken kann eine Vielzahl von Signaturen bei sequenzieller Durchmusterung sicher vom Ergebnis ausgeschlossen werden, so dass lediglich eine geringe Anzahl an Kandidaten verbleibt. Diese Kandidaten müssen anschließend in wahlfreien Zugriffen angefordert werden, um das endgültige Anfrageergebnis zu bestimmen.

Signaturverfahren profitieren von ausschließlich sequenziellen Plattenzugriffen auf kompakte Signaturen in der ersten Stufe und wenigen (teuren) randomisierten Plattenzugriffen auf die verbleibenden Kandidaten. Darüber hinaus lassen sich Nearest-Neighbor-Anfragen so sehr leicht durch horizontale Partitionierung der Signaturdatei in Grid-Architekturen parallelisieren [WBS00]. Offenbar müssen die Signaturen die Ursprungsdaten kompakt repräsentieren und dabei möglichst präzise approximieren. Mit dem VA-File zeichnet sich der bekannteste Ansatz [WSB98] durch ein besonders einfaches Approximationsprinzip

[1] Unter dem Approximationsfehler verstehen wir die Differenz aus exakter Distanz zwischen Anfrage- und einem betrachteten Datenpunkt sowie dem Abstand zur Cluster-Region.

aus, das die einzelnen Dimensionen vollständig in disjunkte bitkodierte Intervalle partitioniert. Die Signaturen setzen sich aus der Verkettung von Bitkodes der Partitionsintervalle zusammen, in denen die jeweiligen Punktkoordinaten liegen. Andere Ansätze nehmen leichte Modifi kationen dieses Approximationsprinzips vor [FTAA00, CZPC02]. Mit der AV-Methode [BSS04, Bal04] führen wir ein Approximationsprinzip ein, das die Länge der Signaturen dynamisch an die Datenverteilung anpasst.

2 Formal-stochastische Grundlagen hochdimensionaler Indexierung

In diesem Abschnitt wollen wir die formal-stochastischen Effekte hochdimensionaler Indexierung skizzieren und die Grundlagen für ein generisches Index-Tuning-Szenario bereitstellen. Der Kerngedanke unserer Formalisierungsbemühungen liegt in der Herleitung der Distanzverteilungen zwischen Punkten und verschiedenen geometrischen Primitiven im hochdimensionalen Raum. Cluster, aber auch Signaturen, repräsentieren Regionen im Vektorraum von zumeist einfacher geometrischer Gestalt (etwa Hyperrechtecke oder Hyperkugeln). Mit den explizit berechenbaren Verteilungsfunktionen stellen wir ein mächtiges Werkzeug für die Anfragekostenmodellierung bereit.

Analytische Kostenmodelle bilden die Grundlage formaler Vergleiche von Indexierungsansätzen (siehe Abschnitt 4) oder der Umsetzung von Index-Tuning- und Anfrageoptimierungskomponenten, die ohne einen Rückgriff auf Datenstichproben auskommen. Unsere Herleitungen beruhen auf der Annahme in $[0, 1]$ gleichverteilter, stochastisch unabhängiger Feature-Werte. Als pessimistische Festlegung stellt diese Einschränkung die ungünstigste Voraussetzung für eine effi ziente Indexunterstützung dar und ist so besonders gut für formale Kostenabschätzungen und -vergleiche geeignet.

Wir verzichten aus Platzgründen auf die ausführliche Herleitung der folgenden Verteilungsfunktionen. Die Distanzverteilung zwischen zwei Punkten erlaubt eine Beurteilung der Distanzen bei steigender Dimensionalität und dient uns darüber hinaus im Folgenden zur Abschätzung der Nearest-Neighbor-Distanz:

$$F_{\|v-v\|_2}(x) = \Phi((x^2 - d/6)/\sqrt{7d/180})$$

Damit gibt $F_{\|v-v\|_2}(x)$ die Wahrscheinlichkeit dafür wieder, dass eine Distanz zwischen zwei Punkten unterhalb von x liegt. Daneben interessieren uns die Distanzen zwischen einem (Anfrage-)Punkt und verschiedenen geometrischen Primitiven, die uns im Rahmen unserer Kostenmodellierung die Betrachtung verschiedener Regionengeometrien gestattet. Im Einzelnen betrachten wir (Hyper-)Kugeln und (Hyper-)Würfel. Die Distanzverteilung zwischen einem Punkt und einer Hyperkugel mit dem Radius r wird durch

$$F_{\text{sphere}}(x) \approx F_{\|v-v\|_2}(x + r)$$

charakterisiert. Die Distanzverteilung zwischen Punkten und Hyperwürfeln, die an fi xen Intervallgrenzen ausgerichtet sind, kann mittels

$$F_{\text{cube}}(x) = \Phi((x^2 - d(w-1)^2/6)/\sqrt{(-11w^4 + 20w^3 - 16w + 7)d/180})$$

berechnet werden, wobei w die Kantenlänge repräsentiert. Darüber hinaus wurden einige weitere Verteilungen entwickelt, von denen für diesen Beitrag die Charakterisierung der Euklidischen Norm eines Punktes von Bedeutung ist, wobei die Koordinaten in einem Intervall $[0, w]$ gleichverteilt sind:

$$F_{\text{norm}}(x) = \Phi((x^2 - dw^2/3)/\sqrt{4dw^4/45})$$

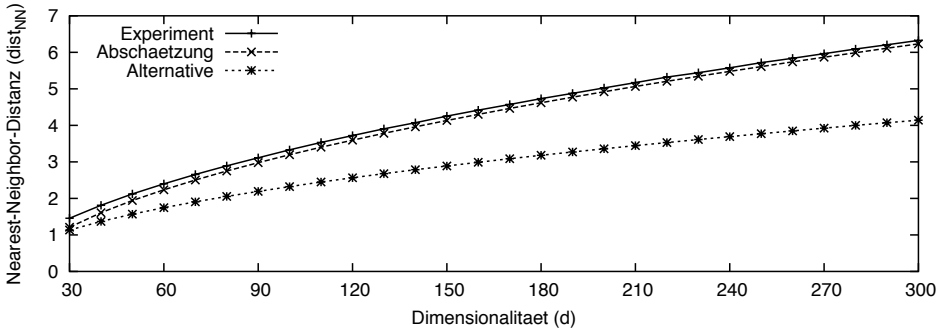

Abbildung 1: Abschätzung der Nearest-Neighbor-Distanz

Als erste Anwendung unserer stochastischen Formalisierungen stellen wir mit der Abschätzung der Nearest-Neighbor-Distanz[2] $\mathsf{dist}_{\text{NN}}$ einen wichtigen Eckpfeiler der Kostenmodellierung bereit. Unsere Abschätzung macht von der bekannten Distanzverteilung zwischen Punkten $F_{\|v-v\|_2}(x)$ derart Gebrauch, dass es keinen Punkt geben kann, der eine geringere Distanz zum Anfragepunkt aufweist als $\mathsf{dist}_{\text{NN}}$[3]:

$$\mathsf{dist}_{\text{NN}} \approx \sqrt{\Phi^{-1}(0{,}5/N) \cdot (d/6) + \sqrt{7d/180}}$$

Nach unserer Kenntnis stellt darüber hinaus lediglich [BBKK97] eine alternative explizite Formel zur Abschätzung der Nearest-Neighbor-Distanz bereit. In Abbildung 1 haben wir beide Ansätze einer experimentellen Ermittlung der Nearest-Neighbor-Distanz bei steigender Dimensionalität gegenübergestellt. Dabei zeigt unsere Abschätzung vor allem bei hohen Dimensionalitäten eine deutlich bessere Näherung und ist so auch für eine analytische Kostenmodellierung praktikabel.

3 Die Active-Vertice-Methode

Mit der *Active-Vertice-* (AV-) Methode führen wir im Folgenden unseren eigenen Indexierungsvorschlag ein. Die AV-Methode lässt die „Bitrate“, also die Länge der Signaturen, flexibel und beschränkt statt dessen den Approximationsfehler mittels eines kontinuierlichen

[2] Die Nearest-Neighbor-Distanz bezeichnet den Abstand zwischen Anfragepunkt und seinem n¨achsten Nachbarn innerhalb einer Datenmenge aus N Punkten in der Dimensionalit¨at d.

[3] $\Phi^{-1}(x)$ bezeichnet die inverse Verteilungsfunktion der $\mathrm{N}(0, 1)$-Normalverteilung.

Parameters r nach oben. Gegenüber alternativen Ansätzen erweist sich diese Entscheidung zum einen deshalb als vorteilhaft, dass jeder indexierte Punkt nur die minimal notwendige Bitrate für einen vorgegebenen Approximationsfehler beansprucht. Zum anderen eröffnet uns dieses Indexierungsprinzip die Möglichkeit, die von den Signaturen induzierte Region nach Belieben zu gestalten und durch die so mögliche Adaptivität an die Datenverteilung nochmals deutliche Verbesserungen des Approximationsfehlers zu erreichen.

Diese Signaturen beschreiben Regionen im Vektorraum, deren Gestalt beliebig gewählt werden kann. Dazu wird ein Referenzpunkt c identifiziert, der nahe genug am indexierten Punkt p liegt und den Mittelpunkt der Region repräsentiert. Der passende Referenzpunkt wird in einem hierarchischen Partitionierungsverfahren bestimmt und kann später aus der Signatur rekonstruiert werden. Die Approximation eines Punktes $p \in [0,1]^d$ in $\mathbb{R}^2$ ist in Abbildung 2 illustriert. Der fixierte Referenzpunkt c_0 bildet den Ausgangspunkt des

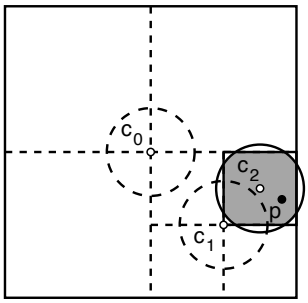

Abbildung 2: Approximationsprinzip der AV-Methode

hierarchischen Approximationsprinzips. Da p außerhalb der Region von c_0 liegt, wird ein nachgeordneter Referenzpunkt c_1 bestimmt, an dem diese Überprüfung erneut angewandt wird. Die Region eines Referenzpunktes ist in diesem Beispiel durch eine Hyperkugel bestimmt, die für alle Punkte mit dem identischen Radius r parametrisiert ist. Anhand des Referenzpunktes c_0 nehmen wir eine Zerlegung des Vektorraumes in 2^d Partitionen (gestrichelte Linien) vor. Der Referenzpunkt c_1 (1) wird so ausgewählt, dass er in der gleichen Partition wie p liegt (rechts unten). Diese Partition lässt sich mit einem Code aus d Bit eindeutig bestimmen, wobei in jeder Dimension i eine „0“ für $c_1[i] < c_0[i]$ und eine „1“ für $c_1[i] > c_0[i]$ vergeben wird[4]. In diesem Beispiel ergibt sich für die Partition von c_1 ein Bitcode „10“, der als Präfix in die Signatur von p eingeht. Die (2) exakte Position von c_1 stimmt in diesem Beispiel mit dem Zentrum der gewählten Partition überein. Da p weiterhin außerhalb der Region des aktuellen Referenzpunktes c_1 liegt, wird die Hierarchisierung fortgesetzt. Der so identifizierte Referenzpunkt c_2 ist gegenüber c_1 durch den Bitcode „11“ auffindbar, die Signatur verlängert sich dementsprechend. Da p nunmehr in der Region von c_2 liegt, ist die Approximation von p beendet, dessen Signatur sich aus „1011“ ergibt.

Die unteren und oberen Distanzschranken zwischen einem Anfragepunkt q und dem in-

[4] $c[i]$ bezeichnet die Koordinate von c in der Dimension i.

dexierten Punkt p sind zum einen durch die minimalen und maximalen Abstände zur Region des korrespondierenden Referenzpunktes (hier: c_2) gegeben. Zum anderen impliziert das Approximationsprinzip die Lage von p in einer rechteckigen Raumpartition, die sich mit jeder Hierarchiestufe verkleinert. Der entsprechende Überlappungsbereich ist in Abbildung 2 schattiert dargestellt. Die Raumpartition kann zur Berechnung alternativer Distanzschranken herangezogen werden, wobei wir aus beiden Alternativen (Region und Partition) die besseren Schranken auswählen können, um eine möglichst kleine Kandidatenmenge zu erhalten.

Ein weiteres Optimierungspotenzial hinsichtlich nochmals verringerter Bitrate und kleinerem Approximationsfehler eröffnet sich, wenn die Datenverteilung zur Adaption (1) des Partitionierungsschemas und (2) der Regionengeometrie herangezogen wird. Die AV-Methode gestattet eine variable Lage der Referenzpunkte, die von der eingeführten symmetrischen Aufteilung abweichen kann. So können in einem Vorverarbeitungsschritt die potenziellen Koordinaten der Referenzpunkte asymmetrisch so festgelegt werden, dass sie in den Schwerpunkt der Punkte einer Partition fallen. Die Regionengeometrie kann beliebig festgelegt werden, wobei angepasste Regionen eine Verringerung des Approximationsfehlers versprechen. Von den untersuchten Figuren [Bal04] konnte der Vorschlag eines unsymmetrischen Ellipsoides die deutlichsten Kostenverringerungen erzielen. Im Gegenzug gestaltet sich die Berechnung der Abstandsschranken in einem numerischen Verfahren komplexer. Der Minimalabstand zwischen Anfragepunkt q und Ellipsoid entspricht der Distanz zu einem Punkt s auf der Oberfläche des Ellipsoides, in dem der Gradient durch q verläuft. Die Formalisierung dieses Zusammenhangs mündet in einem Polynom $(2 \cdot d)$-ten Grades aus dessen größter reeller Nullstelle s bestimmt werden kann. Dazu muss das Polynom (1) zunächst in die Koeffi zientenform überführt werden, um anschließend die Nullstelle in einem numerischen Verfahren mit garantierter Konvergenz zu fi nden. Offenbar müssen diese Berechnungen so effi zient ausgeführt werden, dass die Signaturen weiterhin in sequenziellen Plattenzugriffen gelesen werden können. Durch eine einfache Maßnahme zur zweistufi gen Filterung der Signaturen anhand der Distanzschranken zur Partition und, falls notwendig, anschließender Berechnung der Distanzschranken zur Region, konnte der Anteil dieser aufwändigen Berechnungen in experimentellen Untersuchungen auf weniger als 1% reduziert werden.

4 Kostenmodellierung

Wir werden in diesem Abschnitt kurz die Prinzipien einer formalen Kostenmodellierung der AV-Methode skizzieren und anschließend in einem analytischen Kostenvergleich mit dem VA-File die Anfragekostenvorteile der AV-Methode illustrieren.

Der *filter-and-refinement*-Algorithmus zur Nearest-Neighbor-Anfragebearbeitung läuft in zwei Stufen ab, wobei in der ersten Stufe sämtliche Signaturen sequenziell von der Platte gelesen werden, um daraus eine Kandidatenmenge zu bestimmen, aus denen das endgültige Ergebnis durch randomisierte Zugriffe auf die exakten Daten bestimmt wird. Unsere Kostenmodellierung beschränkt sich auf die Betrachtung der I/O-Kosten, die als „Flaschenhals“ die Ausführungszeit bestimmen [BSS04]. Unsere Kostenmodellierung beruht

auf einer konkreten Hardware-Konfi guration, die durch die Festplattenparameter (1) Positionierungszeit des Lesekopfes $t_{\text{seek}} = 7{,}4\,\text{ms}$, (2) Latenzzeit für das Einrotieren des Blocks $t_{\text{latency}} = 4{,}17$ und (3) Übertragungszeit des Blocks $t_{\text{transfer}} = 0{,}0385$ bestimmt wird. Weiterhin nehmen wir eine Blockgröße von $\mathsf{blocksize} = 1\,\text{kByte}$ und eine Datenbankgröße von $N = 100000$ Punkten an.

Während der ersten Stufe werden sämtliche Signaturen $S_1, \ldots, S_N$ durchlaufen. Damit muss zunächst deren Speicherplatzbedarf bestimmt werden, um die Anzahl der belegten Festplattenblöcke zu ermitteln, die sequenziell gelesen werden müssen. Aufgrund der variablen Signaturlängen müssen wir zunächst für jede mögliche Signaturlänge $|S_i|$ deren relative Häufi gkeit $\text{P}(|S_i| = d \cdot t)$ bestimmen. Offenbar entsprechen die Signaturlängen dem Produkt aus Dimensionalität d und zutreffender Tiefe t der Approximationshierarchie. Wir können $\text{P}(|S_i| = d \cdot t)$ mittels der Hilfsfunktion $G(x,t) = F_{\text{norm}}(x)$ mit $w = 2^{-t}$ ausdrücken, wobei wir auf die detaillierte Herleitung an dieser Stelle verzichten wollen:

$$\text{P}(|S_i| = d \cdot t) = G(r, t+1) - G(r, t)$$

Aus diesem Zusammenhang lassen sich die absoluten Zahlen sehr einfach bestimmen:

$$\mathsf{scan}_{\text{AV}} = t_{\text{seek}} + t_{\text{latency}} + t_{\text{transfer}} \cdot \left\lceil \sum_{t=0}^{\infty} (N \cdot \text{P}(|S_i| = d \cdot t) \cdot d \cdot t) / \mathsf{pagesize} \right\rceil$$

Die zweite Anfragestufe (1) inspiziert die Kandidaten in aufsteigender Reihenfolge ihrer unteren Distanzschranken, (2) fordert die exakte Repräsentation in einem randomisierten Plattenzugriff an und (3) fügt diesen Punkt anhand seiner exakten Distanz erneut in die Kandidatenliste ein. Die Anfrage ist beendet, sobald das erste Element der Kandidatenliste, die beispielsweise über eine Prioritätswarteschlange realisiert werden könnte, ein zuvor wiedereingefügter Punkt ist, der damit als Ergebnis der Nearest-Neighbor-Anfrage feststeht. Offenbar müssen also lediglich jene Kandidaten inspiziert werden, deren untere Distanzschranken unterhalb der Nearest-Neighbor-Distanz liegen.

Wir nehmen an, dass (1) die Distanzschranken ausschließlich über den Abstand zu den Referenzpunktregionen bestimmt werden, (2) wobei wir uns auf die Betrachtung von Hyperkugeln beschränken. Durch beide Annahmen nehmen wir konservative Einschränkungen vor, die sich negativ auf die Anfragekosten der AV-Methode auswirken und somit einen fairen Kostenvergleich mit konkurrierenden Ansätzen gestatten. Der relative Anteil der Signaturen, dessen Minimaldistanz unterhalb der Nearest-Neighbor-Distanz liegt, kann so leicht mittels unserer Verteilungsfunktion $F_{\text{sphere}}(x)$ modelliert werden:

$$\mathsf{access}_{\text{AV}} = N \cdot F_{\text{sphere}}(\mathsf{dist}_{\text{NN}}) \cdot (t_{\text{seek}} + t_{\text{latency}} + t_{\text{transfer}})$$

Die Bestimmung der minimalen Gesamtkosten entspricht einem nichtlinearen Optimierungsproblem, für dessen Lösung verschiedene, schnell konvergierende Heuristiken bereit stehen. Der optimale Radius r^* ist für verschiedene Dimensionalitäten d unterschiedlich.

Die Kostenmodellierung des VA-Files gestaltet sich aufgrund der fi xen Bitrate b nochmals einfacher und soll hier aus Platzgründen nur angerissen werden:

$$\begin{aligned} \mathsf{scan}_{\text{VA}} &= t_{\text{seek}} + t_{\text{latency}} + \lceil (N \cdot d \cdot b) / \mathsf{pagesize} \rceil \cdot t_{\text{transfer}} \\ \mathsf{access}_{\text{VA}} &= N \cdot F_{\text{cube}}(\mathsf{dist}_{\text{NN}}) \cdot (t_{\text{seek}} + t_{\text{latency}} + t_{\text{transfer}}) \end{aligned}$$

Die Kantenlänge w der Regionen ergibt sich implizit aus der Anzahl der für die Approximation aufgebrachten Bits pro Dimension b, so dass $w = 2^{-b}$ gilt. In Abbildung 3 haben

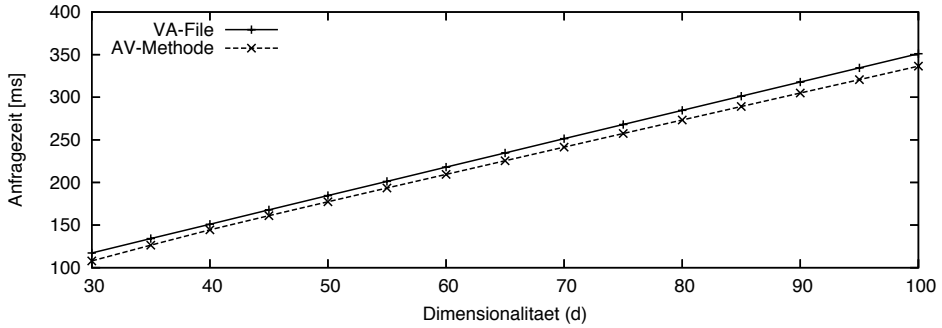

Abbildung 3: Analytischer Anfragekostenvergleich: *AV-Methode* vs. *VA-File*

wir die so ermittelten *minimalen* Anfragekosten beider Methoden für verschiedene Dimensionalitäten $d = 30, \ldots, 100$ abgetragen. Dabei wird ein verringerter Zeitbedarf für die Anfragebearbeitung mit der AV-Methode deutlich. Allein das Grundprinzip aus flexibilisierter Bitrate und Beschränkung des Approximationsfehler führt bei $d = 100$ zu um etwa $4{,}2\%$ verringerten Anfragekosten. Erst mit den besprochenen Anpassungen des Approximationsschemas an konkrete Datenverteilungen wird das volle Potenzial der AV-Methode ausgeschöpft, das nochmals deutlichere Kostenreduzierungen bewirkt.

5 Experimentelle Ergebnisse

Zur Evaluation unserer analytischen Ergebnisse, wollen wir im Folgenden die experimentelle Gegenüberstellung unseres Indexierungsvorschlags mit konkurrierenden Techniken auszugsweise aufführen. Namentlich untersuchen wir die Anfragekosten mittels des Naivverfahrens *SCAN* und vergleichen diese mit einer Indexunterstützung durch die AV-Methode, das VA-File [WSB98] sowie das LPC-File [CZPC02]. Wir verzichten auf die Betrachtung hybrider Ansätze aus hierarchischer Indexstruktur und Signaturverfahren [SYUK00, BBJ$^+$00], da diese Konzepte (1) die Existenz und Identifi zierung von Clustern voraussetzen und sich zudem (2) mit Signaturverfahren kombinieren lassen.

Wir messen die Anzahl sequenzieller und randomisierter Plattenzugriffe und bestimmen den Zeitbedarf anhand der in Abschnitt 4 eingeführten Hardware-Parameter. Im Einzelnen untersuchen wir das Kostenverhalten auf mehreren synthetischen und Realdatenmengen aus $N = 50000$ Punkten in 32 Dimensionen. Dabei handelt es sich um gleichverteilte Daten (U_{32}). Darüber hinaus betrachten wir explizit geclusterte Daten mit gleichverteilten Cluster-Zentren und Punkten, die mit einer Standardabweichung von $\sigma = 0{,}1$ normalverteilt darum streuen (C_{32}), sowie Farbhistogramme aus der COREL Bildsammlung (H_{32}). In Abbildung 4 sind die so ermittelten Anfragezeiten für die 32-dimensionalen Da-

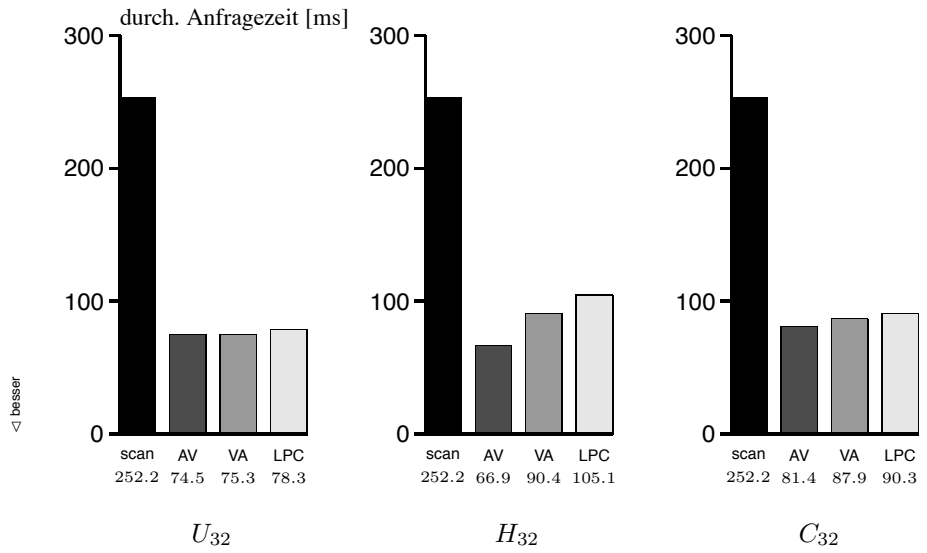

Abbildung 4: Experimenteller Kostenvergleich

tenmengen abgetragen. Durch deutliche Kostenreduzierungen aller drei Signaturverfahren gegenüber *SCAN* wird die Problematik hochdimensionaler Indexierung offenbar in allen Fällen erfolgreich bewältigt. Darüber hinaus benötigt die AV-Methode unter allen untersuchten Datenmengen die geringsten Anfragezeiten, wobei die Kostenunterschiede verschieden ausfallen. Während es bei gleichverteilten Daten wie erwartet nur geringe Kostenunterschiede gibt, fallen die Einsparungen bei den Feature-Daten mit $\approx 26\%$ erheblich größer aus. Offenbar wirken sich die flexiblen Bitraten sowie die angepasste Regionengeometrie des unsymmetrischen Ellipsoides hier sehr deutlich zu Gunsten der AV-Methode aus. Selbst im Falle der synthetisch geclusterten Daten liegen die Einsparungen noch bei $\approx 7\%$.

Weitere Experimente [BSS04, Bal04] untersuchen unabhängig von konkreten Hardware-Parametern die weite Bandbreite an Datenverteilungen. Die Ergebnisse belegen die universelle Einsetzbarkeit der AV-Methode als Indexierungstechnik mit geringen Anfragekosten und guter Adaptivität an die vorgefundenen Datenverteilungen.

6 Zusammenfassung

Die Hauptbeiträge aus [Bal04] können mit (1) der umfassenden Aufarbeitung hochdimensionaler Indexierungsansätze als Entwurfsentscheidung eines neuen Signaturverfahrens; (2) der formal-stochastischen Untersuchung von Distanzen im hochdimensionalen Raum; (3) einer genauen analytischen Abschätzung der Nearest-Neighbor-Distanz mittels expliziter Formeln; (4) dem Vorschlag der AV-Methode als neues Signaturverfahren mit flexiblen Bitraten und verbesserter Regionengeometrie; (5) der formalen Kostenmodellierung der

AV-Methode; (6) der Adaption der AV-Methode an Realdatenverteilungen und (7) der experimentellen Evaluierung und Gegenüberstellung verschiedener konkurrierender Indexierungsansätze benannt werden.

Literatur

[Bal04] S. Balko. *Grundlagen, Entwicklung und Evaluierung einer effi zienten Approximationstechnik für Nearest-Neighbor-Anfragen im hochdimensionalen Vektorraum*. DISDBIS 86. infix, 2004.

[BBJ+00] S. Berchtold, C. Böhm, H. V. Jagadish, H.-P. Kriegel und J. Sander. Independent Quantization: An Index Compression Technique for High-Dimensional Data Spaces. In *ICDE*, Seiten 577–588. IEEE Computer Society, 2000.

[BBK01] C. Böhm, S. Berchtold und D. A. Keim. Searching in High-Dimensional Spaces – Index Structures for Improving the Performance of Multimedia Databases. *ACM Computing Surveys*, 33(3):322–373, 2001.

[BBKK97] S. Berchtold, C. Böhm, D. A. Keim und H.-P. Kriegel. A Cost Model For Nearest Neighbor Search in High-Dimensional Data Space. In *PODS*, Seiten 78–86, 1997.

[BSS04] S. Balko, I. Schmitt und G. Saake. The Active Vertice method: a performant filtering approach to high-dimensional indexing. *DKE*, 51:369–397, 2004.

[CZPC02] G.-H. Cha, X. Zhu, D. Petkovic und C.-W. Chung. An Efficient Indexing Method for Nearest Neighbor Searches in High-Dimensional Image Databases. *IEEE Transactions on Multimedia*, 4(1):76–87, März 2002.

[FTAA00] H. Ferhatosmanoglu, E. Tuncel, D. Agrawal und A. E. Abbadi. Vector Approximation based Indexing for Non-uniform High Dimensional Data Sets. In *CIKM*, Seiten 202–209, 2000.

[HS95] G. R. Hjaltason und H. Samet. Ranking in Spatial Databases. In *4th Int. Symp. on Advances in Spatial Databases*, Jgg. 951 of *LNCS*, Seiten 83–95, 1995.

[SYUK00] Y. Sakurai, M. Yoshikawa, S. Uemura und H. Kojima. The A-tree: An Index Structure for High-Dimensional Spaces Using Relative Approximation. In *VLDB*, Seiten 516–526, 2000.

[WBS00] R. Weber, K. Böhm und H.-J. Schek. Interactive-Time Similarity Search for Large Image Collections Using Parallel VA-Files. In *ECDL*, Seiten 83–92, 2000.

[WSB98] R. Weber, H.-J. Schek und S. Blott. A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In *VLDB*, Seiten 194–205, 1998.

Sören Balko studierte von 1993 bis 1998 Informatik an der Universität Magdeburg. Er erhielt sein Diplom im September 1998 für seine Arbeit über die Behandlung von Integritätsbedingungen bei der Schemaintegration. Ab Oktober 1998 arbeitete als Doktorand in der Gruppe von Gunter Saake auf den Gebieten *Formale Spezifikation von Informationssystemen* und *Multimedia-Datenbanken*. Im September 2003 wechselte Sören Balko für einen Forschungsaufenthalt an das IPK Gatersleben, um dort im Bereich *Datenintegration in der Bioinformatik* zu arbeiten. Im März 2004 verteidigte er seine Dissertation über hochdimensionale Indexierung zur Unterstützung von Nearest-Neighbor-Anfragen. Im Juli 2004 wechselte er an die ETH Zürich, um in der Gruppe von Hans-Jörg Schek als PostDoc auf den Gebieten *Digitale Bibliotheken* und *Grid Computing* zu arbeiten. Seit April 2004 ist er als Projektleiter an der UMIT Innsbruck beschäftigt.