In: Gesellschaft für Informatik (Veranst.): Proceedings of the IR Workshop at LWA 2012 (Lernen, Wissen, Adaption 2012); Dortmund, 12.09.12-14.09.12). 2012.

Techniken der Ressourcenbeschreibung und -auswahl für das geographische Information Retrieval

Stefan Kufer **Daniel Blank** **Andreas Henrich**
Lehrstuhl für Medieninformatik, Universität Bamberg
D-96047 Bamberg
{daniel.blank|andreas.henrich}@uni-bamberg.de

Abstract

Die stete Zunahme der Menge an Medienobjekten und -kollektionen sowohl im WWW als auch auf privaten Endgeräten führt zu einem starken Bedarf nach adäquaten Indexierungs- und Suchtechnologien. Trends wie persönliche Medienarchive, soziale Netzwerke und mobile Geräte mit großer Speicherkapazität sowie Netzwerke mit hoher Bandbreite machen in diesem Kontext verteilte Lösungen und Peer-to-Peer (P2P) Technologien interessant. Hier dient eine Ressourcenselektion, die auf kompakten Beschreibungen der von den Ressourcen verwalteten Inhalte basiert, dazu, für eine bestimmte Anfrage vielversprechende Ressourcen zu ermitteln. Neben z.B. textuellen Informationen können diese Zusammenfassungen auch geographische Informationen der Medienobjekte eines Archivs beschreiben (z.B. wo verwaltete Bilder aufgenommen wurden).

Diese Arbeit präsentiert und evaluiert verschiedene Techniken zur Ressourcenauswahl, die auf der Beschreibung der geographischen Daten persönlicher Medienarchive basieren. Dabei wird nach in der Nähe eines bestimmten Anfragepunktes liegenden Medienobjekten gesucht.

1 Einleitung

In den letzten Jahren ist eine starke Zunahme (persönlicher) Medienobjekte im Web zu verzeichnen. Nutzer schreiben Blogs, „twittern“ oder nutzen Foto- und Videoplattformen. Neben der Speicherung von Medienobjekten tendieren Nutzer dazu, diese mit anderen zu teilen und zu interagieren (z.B. durch Tagging oder Kommentierung von Medienobjekten). Infolgedessen müssen Online-Ressourcen, die sich häufig bzgl. Größe, Medientyp und Aktualisierungshäufigkeit unterscheiden, verwaltet werden [Thomas und Hawking, 2009]. Dies erfordert effektive und effiziente Retrieval-Technologien.

In unserem Szenario werden persönliche Medienarchive in einem P2P-System verwaltet. Die Medienobjekte (z.B. Bilder) eines Archivs sind lokal auf einem Peer bzw. dem Endgerät eines Nutzers gespeichert, ohne dass sie auf Server von Dienstanbietern (wie z.B. *Flickr*) hochgeladen werden müssen. Um die Suche nach Medienobjekten zu unterstützen, können diese durch vier Kriterien beschrieben werden: 1) textuelle Inhalte, 2) low-level Inhaltseigenschaften, 3) Zeitstempel und 4) geographische Informationen (z.B. ein mit „Sonnenuntergang“ getagtes Bild, dessen Farbhistogramm, sowie Aufnahmeort und -zeit). Medienarchive mit mehreren Medienobjekten können demnach durch vier korrespondierende Zusammenfassungen beschrieben werden. Diese stellen eine Aggregation der Eigenschaften der durch die Ressource verwalteten Objekte dar. Die vorliegende Arbeit betrachtet Techniken zur Ressourcenbeschreibung und -auswahl für geographische Daten (s. Abschnitt 2). Verschiedene Beschreibungen werden dabei bzgl. k-Nächste-Nachbarn-Anfragen (kNN-Anfragen) evaluiert (s. Abschnitt 3). Abschnitt 4 ordnet die vorliegende Arbeit in den Kontext relevanter Arbeiten ein. Die Arbeit schließt mit einer kurzen Zusammenfassung und einem Ausblick auf zukünftige Arbeiten in Abschnitt 5.

Unser Ansatz basiert (ist aber keinesfalls beschränkt) auf Rumorama [Müller *et al.*, 2005], einem skalierbaren P2P-Protokoll, das Hierarchien von PlanetP-artigen [Cuenca-Acuna *et al.*, 2003] Netzwerken aufbaut. Bei PlanetP kennt jeder Peer die Ressourcenbeschreibung der anderen Peers im Netzwerk, anhand derer Routing-Entscheidungen während der Anfragebearbeitung getroffen werden. Die dafür notwendige Verteilung der Beschreibungen im Netzwerk wird durch randomisiertes „rumor spreading“ gewährleistet.

2 Techniken der Ressourcenbeschreibung und -auswahl für geographische Anfragen

In diesem Abschnitt werden Techniken der Ressourcenbeschreibung und -auswahl für geographische Anfragen vorgestellt. Dabei existiert ein Trade-off zwischen der Qualität einer Ressourcenbeschreibung und ihrer Größe, d.h. dem Speicherplatzbedarf für deren Repräsentation. Große Zusammenfassungen können mehr Informationen kodieren als kleine und sollten daher eine bessere Ressourcenauswahl ermöglichen. In unserem Szenario ist es erforderlich, einen Kompromiss zwischen diesen beiden gegenläufigen Zielen zu finden. Zusammenfassungen müssen speicherplatzeffizient, gleichzeitig aber auch selektiv genug sein, um eine effiziente Ressourcenauswahl zu unterstützen.

Bei der Ressourcenauswahl werden Peers auf Basis ihrer Ressourcenbeschreibungen, eines Anfragepunktes und ggf. zusätzlichen Informationen (wie z.B. Referenzpunkten) hinsichtlich ihrer potenziellen Relevanz gerankt (d.h., in eine Reihenfolge gebracht, in der sie bei der Anfragebearbeitung kontaktiert werden). Wenn nach den k nächsten Bildern zu einem Anfragepunkt gesucht wird, sollten Peers, die einen höheren Anteil der top-k Bilder verwalten, sich weiter oben im Ranking befinden als solche, die einen geringen Anteil der top-k Bilder besitzen.

Jeder Peer verwaltet eine Menge von Bildern, die jeweils durch ein Paar von Lat/Long-Koordinaten beschrieben werden. Diese geographischen Koordinaten werden

wie Punktdaten in einer zweidimensionalen Ebene behandelt, wodurch ein Peer letztlich zweidimensionale Punktmengen verwaltet. Für Abstandsmessungen wird die euklidische Distanz verwendet, da Untersuchungen in [Blank und Henrich, 2012] auf Basis der gleichen Dokumentkollektion gezeigt haben, dass die Verwendung anderer Distanzmaße wie etwa der Haversine- bzw. der Vincenty-Distanz zu keinen wesentlichen Veränderungen führt. Im Folgenden werden acht verschiedene Techniken zur Beschreibung und Auswahl von Ressourcen vorgestellt. Dabei können zwei generelle Klassen von Verfahren unterschieden werden. Zum einen kann die Punktmenge durch geometrische Formen beschrieben werden, die die Ausmaße der Punktwolke approximieren (Organisation des Datensatzes). Zum anderen lässt sich der Datenraum in eine Menge von Teilräumen (Subräume) zerlegen (Organisation des Datenraumes). In der Zusammenfassung eines Peers wird dann gespeichert, inwiefern er Daten in den einzelnen Subräumen vorliegen hat.

2.1 Geometrische Verfahren

Die Berechnung approximierter, knapper Repräsentationen von komplexen Formen ist ein Standardproblem in vielen Bereichen der Informatik [Becker $et\ al.$, 1991]. Für viele spezielle Probleme existieren Algorithmen, von denen einige für unser P2P-Szenario anwendbar sind. Im Folgenden werden geeignete Techniken vorgestellt.

Minimum Bounding Rectangle (MBR)
Bei Verwendung eines MBR als Ressourcenbeschreibung berechnet jeder Peer das minimale Rechteck, das sämtliche geographischen Koordinaten der von ihm verwalteten Bilder beinhaltet (s. Abb. 1.1). Zur Speicherung eines Paars von Lat/Long-Koordinaten werden 8 Byte benötigt, 4 für Latitude und 4 für Longitude. Zur Repräsentation eines MBR sind zwei Lat/Long-Paare zu speichern, wodurch für eine Peer-Beschreibung insgesamt $2 \cdot 8$ Byte Rohdaten benötigt werden.

Das Peer-Ranking wird folgendermaßen durchgeführt: Wenn Peer p_a den Anfragepunkt in seinem MBR enthält und Peer p_b nicht, so wird p_a vor p_b gerankt. Wenn hingegen sowohl p_a als auch p_b den Anfragepunkt im MBR beinhalten, wird die von den MBRs überdeckte Fläche als zusätzliches Kriterium herangezogen. Dabei wird der Peer, dessen MBR eine kleinere Fläche aufweist, weiter oben in der Rangfolge einsortiert. Wenn weder p_a noch p_b den Anfragepunkt in ihrem MBR enthalten, wird der Peer bevorzugt, dessen MBR eine geringere minimale Distanz zum Anfragepunkt besitzt.

Exakte konvexe Hülle (ECH)
Eine andere Repräsentationsform ist hingegen die exakte konvexe Hülle (ECH: exact convex hull) der Punktwolke eines Peers (s. Abb. 1.2). Diese ist oftmals wesentlich genauer als ein MBR. Zur Kodierung wird in der Regel aber auch mehr Speicherplatz benötigt (im schlechtesten Fall liegen sämtliche von einem Peer verwalteten Punkte auf dem Rand der konvexen Hülle). In der Zusammenfassung eines Peers wird für jeden Hüllenpunkt das Paar seiner Lat/Long-Koordinaten gespeichert, wodurch 8 Byte Speicher pro Hüllenpunkt benötigt werden.

Das Peer-Ranking funktioniert analog zu dem der MBR-Beschreibung.

kMeans++-basierte Zusammenfassung (kMeans++$_k$)
Weiterhin kann die Punktmenge eines Peers aufgeteilt und durch mehrere beschreibende Formen approximiert werden. Eine Möglichkeit ist, ein Clustering der Peer-Daten durchzuführen (s. Abb. 1.3). Clustering zielt allgemein hauptsächlich auf die Identifikation dichter Gruppen von Punkten ab, es lassen sich jedoch auch Informationen über die geometrische Form der Cluster extrahieren [Hershberger $et\ al.$, 2009]. Für unser Szenario wurde ein Verfahren analysiert, das ein kMeans++-Clustering [Arthur und Vassilvitskii, 2007] der Peer-Daten vornimmt[1]. In der Zusammenfassung eines Peers wird für jeden Cluster ein Cluster-Ball, bestehend aus einem Zentrum und einem Radius, gespeichert. Zu dessen Berechnung wird zunächst das MBR der zu dem Cluster gehörenden Punkte aufgespannt. Der Mittelpunkt des MBRs wird als Cluster-Zentrum verwendet, der Radius ergibt sich als maximale Entfernung eines Punktes des Clusters von diesem Mittelpunkt. Zur Speicherung eines Cluster-Balls werden 12 Byte Speicher benötigt (8 Byte für die Lat/Long-Koordinaten des Zentrums, 4 Byte für den Radius)[2].

Die maximale Zahl der berechneten Cluster für einen Peer ist abhängig von einem Parameter k (um die Genauigkeit der Beschreibungsform zu steigern, kann k erhöht werden). Gleichwohl wird unter Umständen eine Anpassung von k an den Datenbestand eines Peers vorgenommen:

i. Wenn ein Peer weniger als k Datenpunkte verwaltet, wird k für diesen Peer vor Beginn des Clusterings auf die Anzahl der Datenpunkte reduziert.

ii. Wenn die verwendete Cluster-Bibliothek einen Cluster zurückliefert, dem kein Datenpunkt zugeordnet ist, wird dieser Cluster nicht in die Zusammenfassung aufgenommen (was einer Reduktion von k entspricht).

Beim Ranking wird für jeden Cluster-Ball eines Peers die minimale Entfernung zum Anfragepunkt berechnet und zusammen mit der Fläche des Balls in einem sogenannten *REntry* gespeichert (liegt der Anfragepunkt innerhalb eines Balles, so beträgt die Distanz 0). Die *REntries* eines Peers werden aufsteigend nach ihrer Entfernung zum Anfragepunkt sortiert. Bei gleichem Abstand wird der *REntry* mit der geringeren Fläche bevorzugt. Zur Bestimmung des Rankings zwischen zwei Peers werden die sortierten *REntries* jeweils der Reihe nach verglichen. Wenn der erste *REntry* von p_a näher am Anfragepunkt liegt als der erste *REntry* von p_b, wird p_a vor p_b gerankt. Wenn beide *REntries* den gleichen Abstand aufweisen, wird der Peer bevorzugt, dessen *REntry* eine geringere Fläche aufweist. Sollte auch dies keine Entscheidung bringen, werden die jeweils zweiten *REntries* auf gleiche Weise verglichen, und so weiter[3].

Rekursive Berechnung von Minimum Area Rectangles (RecMAR$_k$)
Ein anderer Ansatz zur Beschreibung der Daten eines Peers durch mehrere geometrische Formen zerlegt das initiale

[1] Zum Clustering wird die Klasse `KMeansPlusPlusClusterer` der apache-commons-Bibliothek verwendet (http://commons.apache.org/).

[2] Die Verwendung eines Balls zur Repräsentation eines Clusters erwies sich bei gleichem Speicherplatzbedarf als besser als die Verwendung eines MBR, da bei gleicher Zusammenfassungsgröße tendenziell mehr Bälle als MBRs (Faktor 4/3) gespeichert werden können. Die geringere Abgrenzungsgenauigkeit einer Punktmenge durch einen Ball lässt sich somit durch eine größere Anzahl an Formen kompensieren.

[3] Aufgrund der möglichen Reduktion von k besitzen nicht alle Peers gleich viele *REntries*. In diesem Fall werden für denjenigen Peer, der weniger *REntries* besitzt, Dummy-Entries erzeugt, deren Werte sich möglichst ungünstig auf das Ranking auswirken.

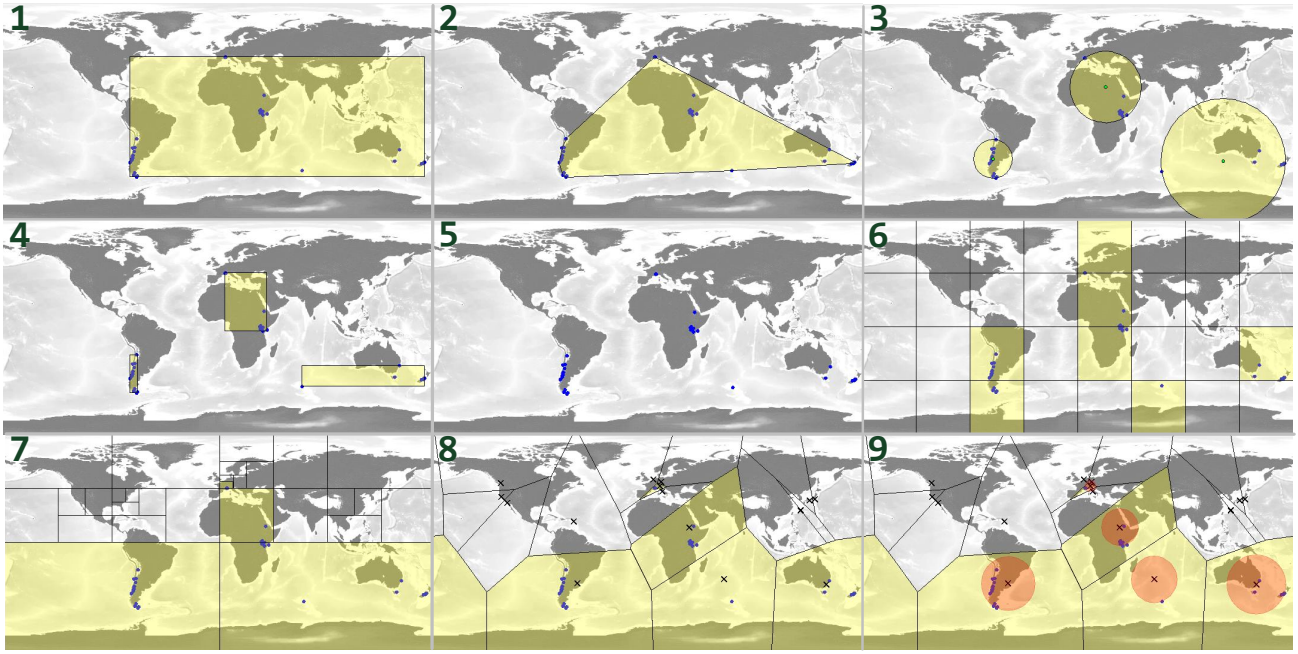

Abbildung 1: Visualisierung der verschiedenen Techniken zur Ressourcenbeschreibung. Die blauen Punkte entsprechen den Datenpunkten eines Peers, die schwarzen Kreuze bei 8 und 9 den Referenzpunkten für UFS/DFS. Die Lage der Datenpunkte des Beispiel-Peers im Datenraum ist in 5 noch einmal separat dargestellt.

MBR in sogenannte Minimum Area Rectangles (MARs) (s. Abb. 1.4). Das Verfahren basiert auf einem in [Becker *et al.*, 1991] vorgestellten Algorithmus zur Zerlegung eines MBR in zwei MARs, der zur Verwendung von Punktdaten angepasst wurde. Die zwei MARs liegen innerhalb des ursprünglichen MBR und beinhalten all dessen Datenpunkte bei minimaler Summe der beiden Teilflächen. Als Erweiterung des Algorithmus wird die Zerlegung rekursiv fortgeführt bis eine festgelegte Maximalzahl k an Rechtecken erreicht ist oder alle bereits berechneten Rechtecke ein bestimmtes Stopp-Kriterium erfüllen. Dabei wird für ein Rechteck der Mittelpunkt ermittelt. Ist die maximale Entfernung eines dem Rechteck zugeordneten Datenpunktes zu diesem Mittelpunkt kleiner als ein Schwellwert *dist*, wird das Rechteck nicht weiter aufgeteilt. Gilt dies für alle bereits berechneten Rechtecke vor Erreichung von k Rechtecken, wird die Rekursion abgebrochen, wodurch der Peer dementsprechend durch weniger als k Rechtecke beschrieben wird[4].

Um die Genauigkeit der Beschreibungsform zu steigern, kann k erhöht werden. Aufgrund des Stopp-Kriteriums können verschiedene Peers, je nach Verteilung ihrer Daten im Raum, dabei durch eine unterschiedliche Zahl an Rechtecken beschrieben werden. Wie beim MBR werden zur Speicherung eines Rechtecks $2 \cdot 8$ Byte Speicher benötigt.

Das Peer-Ranking funktioniert analog zu dem der kMeans++-Beschreibung.

[4]Ursprünglich haben wir eine Kombination von vier Stopp-Kriterien (Fläche, Fläche pro zugeordnetem Punkt, maximale Kantenlänge und Längenverhältnis der kürzeren zur längeren Seite der einzelnen Rechtecke) analysiert. Aufgrund ähnlich guter Ergebnisse wurde jedoch das vorgestellte Kriterium verwendet, u.a. deshalb, weil dessen Einsatz nur von einem Parameter abhängt, der leicht zu interpretieren ist. Wie dieser auf einfache Weise gewählt werden kann, wird in Abschnitt 3.2 beschrieben.

2.2 Raumzerlegende Verfahren

Bei den raumzerlegenden Verfahren wird der Datenraum in Teilräume zerlegt. In der Zusammenfassung eines Peers wird kodiert, in welchen Subräumen dieser Daten vorliegen hat.

Grid-basierte Zusammenfassung (Grid$_n$)

Hierbei wird der geographische Koordinatenraum als ein gleichförmiges Gitter repräsentiert (s. Abb. 1.6). Ein Parameter r spezifiziert die Zahl der Zeilen des Grids. Die Anzahl der Spalten beträgt $2r$ (und somit $n = r \cdot 2r$), da Longitude einen doppelt so großen Wertebereich wie Latitude umfasst. Die Kantenlänge einer Grid-Zelle (in Grad) ergibt sich als $\frac{180^\circ}{r} = \frac{360^\circ}{2r}$. Dies resultiert in uneinheitlich großen Gitterzellen auf der Erdkugel.

Durch Erhöhung der Anzahl der Zellen kann die Selektivität der Beschreibung erhöht, der damit einhergehende zusätzliche Speicherplatzbedarf durch Kompressionstechniken zum Teil kompensiert werden. Als Zusammenfassungen werden Bitvektoren verwendet, wobei jedes Bit eine Zelle repräsentiert. Falls ein Peer ein oder mehrere Bilder besitzt, dessen/deren Geokoordinaten in einer bestimmten Zelle liegen, ist in seiner Zusammenfassung an der entsprechenden Stelle eine 1, ansonsten eine 0 gesetzt[5].

Beim Peer-Ranking wird die den Anfragepunkt beinhaltende Zelle bestimmt. Wenn p_a mindestens ein Bild in dieser Zelle besitzt und p_b nicht, wird p_a vor p_b gerankt. Führt dies zu keiner Entscheidung (d.h., beide haben (k)einen Eintrag für die Anfragezelle), werden rekursiv die Nachbarzellen berücksichtigt. In einer ersten Runde wird die Anzahl der Einträge in dem ersten Ring der bis zu acht (direkten) Nachbarzellen der Anfragezelle verglichen. Hat p_a in diesem Ring mehr Einträge als p_b, so

[5]Dies gilt allgemein für alle raumzerlegenden Beschreibungsformen. Lediglich bei DFS werden für die einzelnen Zellen Distanzinformationen gespeichert.

wird p_a vor p_b gerankt[6]. Ist diese Anzahl ebenfalls gleich, wird der zweite Ring von bis zu 16 Nachbarzellen um die Anfragezelle berücksichtigt, und so weiter. Dieses Verfahren ist in [Blank und Henrich, 2010] beschrieben und dient in der vorliegenden Arbeit als Vergleichsbasis für alternative Verfahren, wenngleich Optimierungen dieses Verfahrens vorstellbar sind (z.B. iterative Berücksichtigung der jeweils zum Anfragepunkt nächstgelegenen Nachbarzelle anstelle aller benachbarter Zellen auf einmal, vgl. Ranking-Algorithmus GFBu$_n$).

GridFileBucket-basierte Zusammenfassung (GFBu$_n$)
Diese Beschreibungsart basiert auf der in Zusammenhang mit dem GridFile [Nievergelt $et\ al.$, 1984] vorgestellten Zerlegung des Datenraumes in Buckets. Dabei wird anhand von Trainingsdaten eine Zerlegung des Datenraumes in unterschiedlich große, rechteckige Zellen erlernt (s. Abb. 1.7). Zu Beginn besteht der Datenraum aus einem einzigen Bucket. In diesen werden sukzessive Trainingsdatenpunkte eingefügt, bis ein Überlauf eintritt und der Bucket in zwei neue Buckets aufgeteilt wird[7]. Anschließend werden weitere Punkte in die Datenstruktur eingefügt. Dies wird bis zur Erreichung der gewünschten Anzahl an Teilräumen fortgeführt. Durch dieses Vorgehen bilden sich in Regionen, in denen viele (Trainings-)Datenpunkte liegen, deutlich kleinere Zellen als in Gebieten, in denen wenig Datenpunkte liegen. Bei entsprechender Auswahl der Trainingsdaten entsteht so eine Raumzerlegung, bei der die Bilder der Kollektion (s. Abschnitt 3.1) gleichmäßiger auf die einzelnen Zellen aufgeteilt sind als bspw. beim Grid. Die grundlegende Raumaufteilung ist allen Peers bekannt und wird im Zuge von Software-Updates verteilt.

Beim Peer-Ranking wird erneut zunächst die Anfragezelle bestimmt. Wenn ein Peer p_a mindestens ein Bild in dieser Anfragezelle besitzt und p_b nicht, wird p_a vor p_b gerankt. Führt dies zu keiner Entscheidung, werden die Nachbarzellen gemäß ihrer minimalen Distanz zum Anfragepunkt aufsteigend in einer Liste sortiert (d.h., das erste Listenelement entspricht der Nachbarzelle mit minimalem Abstand zum Anfragepunkt, usw.). Besitzt p_a einen Eintrag für das erste Element der Liste und p_b nicht, so wird p_a vor p_b gerankt. Sind die Einträge gleich, werden sukzessive die nächsten Listenelemente überprüft, um eine Entscheidung herbeiführen zu können.

Ultra Fine-grained Summaries (UFS$_n$)
Bei dieser Beschreibungsform wird der Datenraum anhand von n Referenzpunkten so zerlegt, dass eine Voronoi-Diagramm-artige Aufteilung entsteht (s. Abb. 1.8). Ein Datenpunkt befindet sich dabei in der Zelle des ihm am nächsten liegenden Referenzpunktes. Je repräsentativer (für die zugrundeliegende Datenkollektion) die Referenzpunkte ausgewählt werden, desto gleichmäßiger verteilen sich auch die Bilder auf die einzelnen Zellen. Die Menge der Referenzpunkte ist allen Peers bekannt und kann ebenfalls mit Software-Updates transferiert werden.

Das Ranking funktioniert ähnlich wie bei der GFBu-Zusammenfassung. Der einzige Unterschied ist, dass (aufgrund der Komplexität der exakten Berechnung der nächsten Nachbarzelle) die Liste der sukzessive zu berücksichtigenden Teilräume nicht nach der minimalen Distanz der Zellengrenzen, sondern nach der minimalen Distanz der jeweiligen Referenzpunkte zum Anfragepunkt sortiert wird.

Distance Fine-grained Summaries (DFS$_n$)
Wie bei UFS wird bei dieser Beschreibungsform der Raum anhand von n Referenzpunkten zerlegt. Allerdings werden hier in den Zusammenfassungen Distanzinformationen (4 Byte je Zelle) erfasst. In der Beschreibung eines Peers wird dabei für jede Voronoi-Zelle gespeichert, wie groß die maximale Entfernung der sich in der Zelle befindlichen, vom Peer verwalteten Datenpunkte vom Referenzpunkt der Zelle ist. Besitzt der Peer in der Zelle kein Bild, so steht ein negativer Platzhalterwert an entsprechender Stelle in der Zusammenfassung. Mit diesen Informationen können Bälle (mit den Referenzpunkten als Zentren und den Distanzwerten als Radien), innerhalb derer sich alle Datenpunkte eines Peers befinden (s. Abb. 1.9, rote Kreise), dazu dienen, Peers während der Anfragebearbeitung auszuschließen.

Das Ranking funktioniert analog zu dem der UFS-Beschreibung (ist ein Eintrag in der Zusammenfassung nicht negativ, so bedeutet das, dass der Peer mindestens ein Bild in der entsprechenden Zelle besitzt). Mit den zusätzlichen Informationen lassen sich bei DFS im Rahmen der Anfragebearbeitung (s. Abschnitt 3.2) allerdings z.B. auch Peers, die für die Anfragezelle Einträge in ihrer Zusammenfassung besitzen, als für eine Anfrage irrelevant klassifizieren (wenn sich die alle Datenpunkte des Peers beinhaltenden Bälle und der Anfrageball nicht überlappen).

3 Evaluation

Im Folgenden wird die zur Evaluation verwendete Datenkollektion vorgestellt (Abschnitt 3.1) sowie der Aufbau der Experimente erläutert (Abschnitt 3.2). In Abschnitt 3.3 werden die Ergebnisse der Experimente analysiert.

3.1 Datenkollektion

Im Jahr 2007 haben wir eine große Anzahl öffentlich verfügbarer, georeferenzierter Bilder, die bei *Flickr* hochgeladen wurden, gecrawlt. Zur Simulation unseres Szenarios betreibt jeder Nutzer einen eigenen Peer, d.h., die Bilder werden anhand der *Flickr*-UserID den Peers zugeordnet. Insgesamt besteht die Datenkollektion aus 406.450 georeferenzierten Bildern von 5.951 verschiedenen Nutzern (bzw. Peers).

Abb. 2 zeigt links die Verteilung der Anzahl der Bilder auf die Peers der Kollektion. Die Verteilung ist windschief (d.h., es gibt viele Peers, die sehr wenige Bilder verwalten und einige wenige Peers, die sehr viele Bilder besitzen), was typisch für P2P-Netzwerke ist [Cuenca-Acuna $et\ al.$, 2003]. 1% der Peers verwalten 42,0% der Bilder, der größte Peer alleine 8,8%. Im Gegensatz dazu bestehen die Archive von 20,7% der Peers aus nur einem Bild. Abb. 2 illustriert rechts die Verteilung der Bilder im Datenraum, bei der sich eine Konzentration auf Nordamerika, Westeuropa und Japan feststellen lässt.

[6] Wenn sich die Anfragezelle nahe der Pole befindet, sind unter Umständen weniger als die erwähnten Anzahlen von Nachbarzellen vorhanden, da die Lat-Dimension „oben“ und „unten“ abgeschlossen ist. Im Fall der Long-Dimension wird deren Nicht-Abgeschlossenheit berücksichtigt, die Nachbarzellen ggf. über die Datumsgrenze hinweg bestimmt.

[7] Dabei wurden verschiedene Splitstrategien getestet. Aus der Wahl der Splitdimension (zyklisch zweimal Long und einmal Lat bzw. abwechselnd Long und Lat) sowie der Splitposition (Mitte der Zelle, Median bzw. Mittelwert der in der Zelle liegenden Datenpunkte in der Splitdimension) resultieren sechs Splitstrategien. In den Experimenten wurde stets die Strategie mit den besten Ergebnissen verwendet.

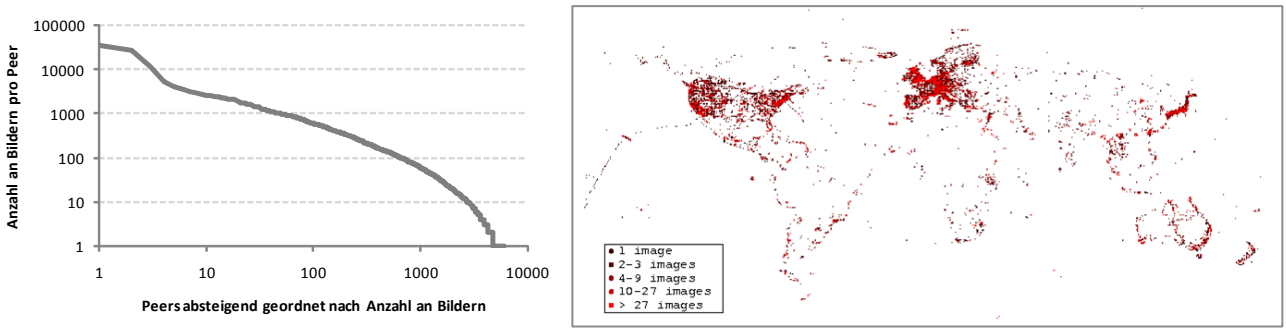

Abbildung 2: Verteilung von Bildern auf Peers (links), geographische Verteilung der Kollektionsdaten (rechts).

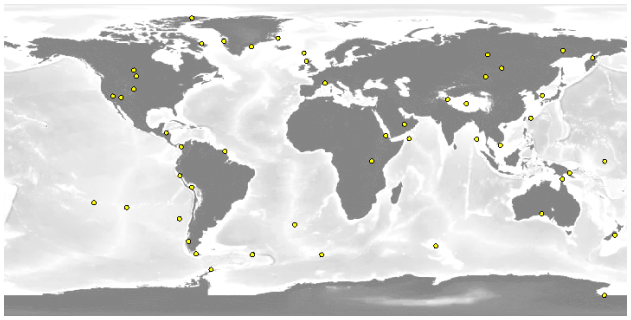

Abbildung 3: Geographische Verteilung der Anfragepunkte für *queryMode* = 2.

3.2 Experimentaufbau

Es werden zwei verschiedene Modi verwendet, um die geographischen Positionen der Anfragepunkte zu bestimmen. Beim ersten (*queryMode* = 1) wird zunächst ein zufälliger Peer aus der Datenkollektion ausgewählt. Anschließend wird wiederum zufällig ein von diesem Peer verwaltetes Bild ausgewählt. Dessen Geokoordinaten werden als Anfragepunkt verwendet. Die Anfragepunkte werden nicht aus der Kollektion entfernt. Insgesamt werden auf diese Weise 500 Anfragen bestimmt. Beim zweiten Modus (*queryMode* = 2) stammen die Geopositionen hingegen aus einer externen, verschiedene reale Orte spezifierenden Quelle. Die Anfragepunkte liegen bei *queryMode* = 1 in der Regel in Regionen hoher Punktdichte. Um Anfragepunkte aus Regionen geringer Punktdichte zu erhalten, wurden für *queryMode* = 2 daher 50 Geokoordinaten aus verschiedenen Listen „abgelegener Orte“ gewonnen (vgl. Abb. 3)[8].

Für die parametrisierbaren geometrischen Verfahren (kMeans++$_k$ und RecMAR$_k$) wird die Maximalzahl der zu berechnenden Formen für die Zusammenfassung der Daten eines Peers mit $k = 3$, $k = 6$ und $k = 9$ variiert. Zur Bestimmung von *dist* bei RecMAR führen wir 500 separate kNN-Anfragen durch und erfassen jeweils die Entfernung des k-nächsten Dokuments. In den Experimenten setzen wir *dist* auf den Median (RecMAR_M$_k$) bzw. das 0,75-Quantil (RecMAR_Q$_k$) dieser 500 ermittelten Werte. Hiermit simulieren wir eine adaptive Bestimmung eines geeigneten *dist*-Wertes während der Anfragebearbeitung.

Für die raumzerlegenden Verfahren wird die Subraumzahl mit $n \in \{512, 2.048, 8.192\}$ variiert. Bei Verwendung geeigneter Trainingsdaten bzw. Referenzpunkte kann die Raumaufteilung bei GFBu bzw. den Voronoi-Diagramm-basierten Zusammenfassungen (UFS/DFS) an die Datenverteilung der zugrundeliegenden Kollektion angepasst werden[9]. Wir verwenden dafür Daten aus zwei verschiedenen Quellen. Bei ersterer werden für die Trainingsdaten- bzw. Referenzpunkte die Geokoordinaten zufällig gewählter Datenpunkte aus der Kollektion verwendet. Bei zweiterer stammen die Daten jeweils aus einer externen Quelle, wobei sie in ihrer Verteilung derjenigen des Netzwerkes angenähert sind[10]. Konkret werden dazu Statistiken zum Bruttoinlandsprodukt (BIP) der Länder von Worldmapper (http://www.worldmapper.org/) genutzt. Proportional zu diesen Statistiken werden zufällige, aus den verschiedenen Ländern stammende Geokoordinaten als Trainingsdaten- bzw. Referenzpunkte aus dem Geonames Gazetteer (http://www.geonames.org/) ausgewählt (d.h., wenn ein Land x% des weltweiten BIP erwirtschaftet, stammen x% der Datenpunkte aus diesem Land). Diese Strategie erwies sich für die Auswahl von Referenzpunkten aus einer externen Quelle bereits in [Blank und Henrich, 2010] im Vergleich mit anderen Verfahren als erfolgversprechend, da die Flickr-Bilder einer ähnlichen geografischen Verteilung wie das BIP folgen (vgl. Abb. 2, rechts).

Die Speicherplatzeffizienz der Verfahren wird durch Analyse der Zusammenfassungsgrößen gemessen (s. Abschnitt 3.3). Zur Kompression der Zusammenfassungen der raumzerlegenden Verfahren wird die Java gzip-Implementierung verwendet (für die geometrischen Verfahren resultiert eine Kompression in größeren Beschreibungen). Die Messungen beinhalten einen Serialisierungsoverhead, der zur Verteilung der Zusammenfassungen im Netzwerk erforderlich ist. Zur Bewertung der Retrieval-Leistung bzw. der Selektivität der Verfahren wird der durchschnittliche Anteil der Peers erfasst, der zur Ermittlung der top-50 Bilder bzgl. eines gegebenen Lat/Long-Paares als Anfragepunkt angefragt werden muss.

Zur Bestimmung der 50 nächsten Nachbarn wird ein kNN-Algorithmus verwendet, der als Bereichsanfrage mit sich reduzierendem Anfrageradius implementiert wurde. Zunächst werden sämtliche Peers anhand ihrer Beschreibungen gemäß der in Abschnitt 2 vorgestellten Ranking-

[8] Verwendete Quellen: List25: Most Remote Places, Open Travel: Remotest Places, Lonely Planet: The World's Best Secret Islands, Supertightstuff: Most Isolated Inhabited Locations, Weburbanist: Remotest Abondened Wonders, Lonely Planet: Best 'Middle of Nowhere' Places, Forbes: Places to Hide, Yahoo UK Travel: Most Remote Places on Earth, letzter Abruf: 13.07.2012.

[9] D.h., der Raum wird dort, wo viele Bilder liegen, in kleinere Zellen zerlegt. In Gebieten mit wenigen Bildern bleibt die Raumauflösung hingegen grob.

[10] Zur Unterscheidbarkeit werden die Verfahren im Fall der Verwendung externer Daten um ein „_e“ ergänzt. UFS wird so z.B. zu UFS_e, UFS(_e) steht für beide Methoden.

Algorithmen initial sortiert. Von den zehn bestgerankten Peers werden jeweils die 50 nächsten Bilder zum Anfragepunkt angefragt, um daraus die 50 besten Treffer aller bisher betrachteten Peers zu ermitteln. Die jeweils bereits betrachteten Peers werden anschließend aus der Menge der noch zu analysierenden Peers entfernt. Die Distanz des Anfragepunktes zur Geokoordinate des am weitesten entfernten der 50 bis dahin ermittelten Bilder ist der Anfrageradius der nächsten „Anfragerunde“, in der wiederum zehn Peers kontaktiert werden[11]. Peers, die aufgrund ihrer Beschreibung keine Bilder in Anfragereichweite besitzen können, werden dabei aus der Menge der zu betrachtenden Peers entfernt. Ein erneutes Ranking findet nicht statt. Die zehn höchstgerankten Peers der Restmenge werden auf relevante Bilder geprüft, und so weiter. Dies wird fortgeführt, bis die Menge der zu analysierenden Peers leer ist und die 50 nächsten Nachbarn somit ermittelt sind. Bei den geometrischen Verfahren können irrelevante Peers anhand der Distanz ihrer beschreibenden Formen zum Anfragepunkt bestimmt werden. Für die raumzerlegenden Verfahren sind zunächst die Subräume in Anfragereichweite zu ermitteln. Peers, die keine entsprechenden Einträge in ihren Zusammenfassungen haben, können aus der weiteren Betrachtung ausgeschlossen werden. Für DFS können zusätzlich noch die gespeicherten Distanzwerte zur Identifikation irrelevanter Peers verwendet werden.

3.3 Experimente

Tabelle 1 stellt die Ergebnisse der Experimente dar. Zunächst werden die geometrischen Verfahren analysiert.

Die Größe der Zusammenfassungen auf Basis des MBR beträgt konstant 43 Byte (16 Byte für die Koordinaten plus 27 Byte Serialisierungsoverhead). Bei der ECH schwanken die Größen sehr stark, mit deutlichen Ausreißern in der maximalen Beschreibungsgröße eines Peers. Da viele Peers jedoch nur ein Bild verwalten und für diese 35 (8 + 27) Byte zur Beschreibung ausreichen, sind die Zusammenfassungen im Durchschnitt nur ca. 42% größer als beim MBR. Für kMeans++ und RecMAR sorgt die ggf. vorgenommene Reduktion der Anzahl der beschreibenden Formen dafür, dass die mittleren Beschreibungsgrößen in moderatem Ausmaß steigen. Bei RecMAR_M sind die Zusammenfassungen durchschnittlich größer als bei RecMAR_Q, demgegenüber ist bei ersterem die Retrieval-Leistung besser. Allgemein führt eine Erhöhung des *dist*-Parameterwertes zu einer Verschlechterung der Retrieval-Leistung bei verringertem Speicherplatzbedarf. Eine Abwägung zu Gunsten der Retrieval-Leistung bzw. des Speicherplatzbedarfs kann somit anwendungsabhängig vorgenommen werden. Die kMeans++$_9$-Beschreibungen sind im Mittel größer als die bei RecMAR$_9$ (bei beiden Parametrisierungen von *dist*), obwohl für einen Ball weniger Speicherplatz benötigt wird als für ein Rechteck. Die Reduktion der Anzahl der verwendeten Formen anhand des Datenbestandes eines Peers greift bei RecMAR demnach häufiger als bei kMeans++. Daraus lässt sich ableiten, dass bei kMeans++ die Gesamtheit der Peers durch wesentlich mehr Formen beschrieben wird als bei RecMAR.

[11] Die Simulation in Runden soll der Ausnutzung der Parallelität in verteilten Szenarien Rechnung tragen. Die Ermittlung des „Auffindungszeitpunkts“ eines relevanten Dokumentes (als wievielter Peer der verwaltende Peer kontaktiert wird) erfolgt jedoch auf „Einzelpeerbasis“, d.h., die Ranking-Position eines Peers wird in Einzel- und nicht in Zehnerschritten erfasst.

Bei der Retrieval-Leistung zeigt sich unabhängig vom Anfragemodus, dass diese mit den durch die beschreibenden Formen überdeckten Flächen korreliert (s. Tabelle 1, Spalte rechts). MBR erweist sich als das ungünstigste Verfahren. Schon die ECH ermöglicht eine wesentlich selektivere Auswahl von Peers. Durch Verwendung mehrerer Formen in einer Ressourcenbeschreibung lassen sich nochmals bessere Leistungen erzielen. Die überdeckten Flächen sind durch mehrere einfache Formen stärker reduzierbar als bei Verwendung einer einzelnen, komplexen Form. Sowohl bei kMeans++$_k$ als auch bei RecMAR$_k$ verbessert sich die Retrieval-Leistung durch Erhöhung von k degressiv. Die Punktmengen werden bei RecMAR mit insgesamt weniger Formen in kleinere Flächen eingegrenzt als bei kMeans++. Der starke Fokus auf die Flächenreduzierung bei RecMAR macht sich hier bemerkbar. Dadurch überlappen sich weniger beschreibende Formen verschiedener Peers, was zu einem besseren Ranking der Peers führt.

Bei Abwägung zwischen Retrieval-Leistung und Zusammenfassungsgrößen erscheint RecMAR als der vielversprechendste geometrische Ansatz. Bei nicht wesentlich speicherplatzintensiveren Zusammenfassungen zeigt RecMAR eine deutlich bessere Performance als MBR, gegenüber kMeans++ und ECH ist RecMAR letztlich auch mit kleineren mittleren Beschreibungsgrößen leistungsfähiger.

Dank der Kompression fällt der Anstieg der mittleren Zusammenfassungsgrößen bei den raumzerlegenden Verfahren auch bei starker Erhöhung der Subraumzahl gering aus. Allgemein sind die Zusammenfassungen beim Grid bei vergleichbarer Subraumzahl am kleinsten, da hier durchschnittlich am wenigsten Zellen belegt sind. Für GFBu(_e), UFS(_e) und DFS(_e) zeigt sich jeweils, dass die Beschreibungen bei Verwendung von Kollektionsdaten zum Erlernen der Raumpartitionierung größer sind, als wenn die Daten aus externer Quelle stammen. UFS(_e) zeigt dabei leicht besseres Kompressionspotenzial als GFBu(_e). Die Distanzwerte bei DFS(_e) sind deutlich schlechter komprimierbar, da (vor der Kompression) pro Zelle 4 Byte Daten (statt nur 1 Bit) anfallen und auch stärker variierende Werte in den Beschreibungen stehen. Folglich sind bei DFS(_e)$_{8192}$ Zusammenfassungen im Durchschnitt über dreieinhalb mal, bei UFS(_e)$_{8192}$ und GFBu(_e)$_{8192}$ hingegen nur ca. eineinhalb mal größer als bei MBR. Letztere sind somit kleiner als bei RecMAR$_9$ und insbesondere kMeans++$_9$.

Allgemein arbeiten die raumzerlegenden Methoden besser, je weniger Geokoordinaten und damit Bilder in der Anfragezelle (bzw. deren Nachbarzellen) liegen, da dann die relevanten Bilder und die sie verwaltenden Peers schnell gefunden werden können. Befinden sich hingegen viele Bilder in der Anfragezelle, besitzen in der Regel auch viele Peers entsprechende Einträge in ihren Zusammenfassungen, was das Ranking erschwert.

Bei *queryMode* = 1 liegen die Anfragepunkte in der Regel in Regionen hoher Punktdichte. Die Methoden mit angepasster Raumzerlegung schlagen das Grid deutlich, dessen Auflösung in diesen Gebieten zu grob ist. Bei *queryMode* = 2 liegen die Anfragepunkte hingegen in der Regel in Gebieten geringer Punktdichte. Hier rückt das Grid näher an die anderen Verfahren heran und kann bei hohen Subraumzahlen sogar GFBu(_e) schlagen. UFS(_e)/DFS(_e) bleiben jedoch auch bei hohen Subraumzahlen leicht leistungsstärker als das Grid. UFS(_e)/DFS(_e) verhalten sich unter vergleichbaren Bedingungen stets etwas besser als GFBu(_e). Die Voronoi-Diagramm-artige Raumzerlegung

Verfahren	*queryMode* = 1	*queryMode* = 2	$S_\emptyset$	S_{min}	S_{max}	Ø abgedeckte Fläche
MBR	4,15	1,85	43	43	43	624,27
ECH	2,07	0,74	61,43	35	347	203,63
kMeans++$_3$	1,54	0,48	55,25	39	63	114,76
kMeans++$_6$	0,68	0,17	73,80	39	99	20,80
kMeans++$_9$	0,44	0,12	88,74	39	135	9,19
RecMAR_M$_3$	1,02	0,26	58,37	43	75	41,03
RecMAR_M$_6$	0,49	0,13	71,25	43	123	7,95
RecMAR_M$_9$	0,35	0,11	79,31	43	171	2,91
RecMAR_Q$_3$	1,03	0,26	55,64	43	75	41,03
RecMAR_Q$_6$	0,52	0,13	64,59	43	123	7,96
RecMAR_Q$_9$	0,38	0,11	69,31	43	171	2,93
Grid$_{512}$	7,50	0,96	52,35*	51*	88*	nicht berechnet
Grid$_{2048}$	4,00	0,59	53,97*	51*	118*	nicht berechnet
Grid$_{8192}$	2,32	0,31	58,14*	51*	147*	nicht berechnet
GFBu$_{512}$	1,39	0,72	55,68*	48*	107,6*	nicht berechnet
GFBu$_{2048}$	0,80	0,48	60,38*	48*	195,8*	nicht berechnet
GFBu$_{8192}$	0,63	0,44	68,41*	48*	470,5*	nicht berechnet
UFS$_{512}$	1,07	0,72	55,42*	48*	109,9*	nicht berechnet
UFS$_{2048}$	0,50	0,34	59,55*	48*	194,9*	nicht berechnet
UFS$_{8192}$	0,27	0,19	66,88*	48*	467,4*	nicht berechnet
DFS$_{512}$	0,82	0,45	83,00*	67,3*	474*	nicht berechnet
DFS$_{2048}$	0,42	0,24	118,75*	86,9*	804,4*	nicht berechnet
DFS$_{8192}$	0,25	0,15	161,14*	109,1*	2130,9*	nicht berechnet
GFBu_e$_{512}$	1,80	0,78	53,82*	48*	93*	nicht berechnet
GFBu_e$_{2048}$	1,06	0,48	56,94*	48*	154,9*	nicht berechnet
GFBu_e$_{8192}$	0,79	0,49	63,44*	50*	252,8*	nicht berechnet
UFS_e$_{512}$	1,59	0,65	53,31*	48*	92,8*	nicht berechnet
UFS_e$_{2048}$	0,97	0,43	55,70*	48,2*	145,3*	nicht berechnet
UFS_e$_{8192}$	0,66	0,25	60,46*	49,8*	295,3*	nicht berechnet
DFS_e$_{512}$	1,20	0,42	79,42*	68*	433,9*	nicht berechnet
DFS_e$_{2048}$	0,77	0,27	110,21*	88,7*	701,9*	nicht berechnet
DFS_e$_{8192}$	0,55	0,19	142,74*	111*	1617*	nicht berechnet

Tabelle 1: Ergebnistabelle mit den jeweiligen Retrieval-Leistungen für *queryMode* = 1 und *queryMode* = 2 (Angaben in % der kontaktierten Peers zur Ermittlung der 50 nächsten Nachbarn), den durchschnittlichen, minimalen und maximalen Zusammenfassungsgrößen S (in Byte) sowie den durch die geometrischen Formen durchschnittlich überdeckten Flächen. Durch Kompression erreichte Werte der Zusammenfassungsgrößen sind durch * gekennzeichnet.

führt zu einer leicht besseren Adaption der Raumaufteilung an die Datenkollektion. Mit den in den DFS(_e)-Zusammenfassungen gespeicherten Distanzwerten lassen sich ggf. auch Peers, die sehr weit oben im Ranking stehen, als für eine Anfrage irrelevant klassifizieren (z.B. solche, die zwar Bilder in der Anfragezelle besitzen, welche aber alle außerhalb des Anfrageballs liegen). Deshalb werden bei DFS(_e) die relevanten Bilder trotz des gleichen initialen Rankings schneller gefunden als bei UFS(_e). Mit steigender Subraumzahl werden die Unterschiede jedoch geringer. Sowohl für GFBu(_e) als auch UFS(_e)/DFS(_e) führt bei der Wahl der Referenzpunkte die Verwendung von Kollektionsdaten erwartungsgemäß zu besseren Ergebnissen als das Zurückgreifen auf externe Daten.

Allgemein lässt sich die Retrieval-Leistung bei allen raumzerlegenden Verfahren durch Erhöhung der Teilraumzahl verbessern. Unter Berücksichtigung des Speicherplatzbedarfs erscheint UFS als beste raumzerlegende Alternative, da UFS bei deutlich kleineren Zusammenfassungen eine fast ebenso gute Retrieval-Leistung zeigt wie DFS sowie bei ähnlicher Zusammenfassungsgröße dem Grid deutlich und GFBu leicht überlegen ist. Dank der starken Anpassung der Raumzerlegung an die Datenverteilung der Kollektion ist UFS$_{8192}$ bei *queryMode* = 1 sogar besser als RecMAR$_9$. Bei *queryMode* = 2 werden bei UFS$_{8192}$ hingegen fast doppelt so viele Peers angefragt wie bei RecMAR$_9$, allerdings sind die absoluten Unterschiede sehr gering und betragen nur ca. 6 Peers.

4 Verwandte Arbeiten

In dieser Arbeit werden verschiedene Techniken zur Beschreibung räumlicher Daten analysiert, um Kollektionen zweidimensionaler Geokoordinaten zusammenzufassen. Vorgestellte Alternativen orientieren sich dabei an den aus räumlichen Indexstrukturen bekannten Beschreibungstechniken (vgl. [Samet, 2005]) wie bspw. dem R-Baum [Guttman, 1984], dem k-d-Baum [Bentley, 1975], dem GridFile [Nievergelt $et\ al.$, 1984], Verfahren auf Basis einer Voronoi-Zerlegung (z.B. [Kolahdouzan und Shahabi, 2004]), bzw. Beschreibungen auf Basis der konvexen Hülle (vgl. [Preparata und Shamos, 1985]). Ebenso werden Techniken aus dem Bereich metrischer Indexstrukturen aufgegriffen, bspw. bei den Beschreibungsformen UFS bzw. DFS (vgl. auch hierfür [Samet, 2005]).

In zukünftigen Arbeiten sollen auch hybride Techniken analysiert werden, wie sie bereits im Kontext räumlicher Indexstrukturen vorgeschlagen wurden. Dabei werden z.B. geometrische Beschreibungsverfahren auf Basis von MBRs mit raumzerlegenden Verfahren (z.B. Voronoi-Diagramm-basiert) kombiniert. Ein solcher Ansatz ist im Bereich der Indexstrukturen z.B. in [Sharifzadeh und Shahabi, 2010] beschrieben.

Das Szenario eines zusammenfassungsbasierten P2P-Systems stellt im Kontext unserer Arbeit den allgemeinen Rahmen dar, um verschiedene Zusammenfassungstechniken gegeneinander zu evaluieren. Alternativ dazu könnte es gerade im Kontext zweidimensionaler Geodaten attrak-

tiv sein, auch andere P2P-Systeme einzusetzen. Strukturierte Systeme könnten sich eignen. Hierbei ist jeder Peer für einen bestimmten Datenbereich zuständig. Neu im P2P-Netz zu verwaltende Indexdaten werden entsprechend dieser Zuständigkeiten transferiert. Dies reduziert die Autonomie der Peers [Doulkeridis *et al.*, 2009]. Gleichzeitig ist bei strukturierten Ansätzen eine Anfragebearbeitung mit logarithmischem Aufwand möglich [Doulkeridis *et al.*, 2009]. Ein umfassender Überblick über alternative P2P-Technologien ist in [Shen *et al.*, 2009] gegeben.

Im aufgezeigten Szenario werden derzeit Anfragen mit geographischem Bezug auf Basis unabhängiger Zusammenfassungstechniken bearbeitet. Es könnte lohnenswert sein, Zusammenfassungstypen nicht isoliert zu betrachten, um bspw. Anfragen nach einem Bild, das einen Sonnenuntergang in den Alpen zeigt, besser beantworten zu können. Derzeit würden je eine Zusammenfassung für low-level Inhaltsinformationen der Bilder und eine Zusammenfassung der geografischen Informationen zu zwei Rankings der Ressourcen führen. Diese beiden Rankings müssten anschließend mit entsprechenden Techniken kombiniert werden (vgl. z.B. [Belkin *et al.*, 1995]). [Hariharan *et al.*, 2008] präsentieren hierzu einen Ansatz, der mehrere Zusammenfassungstypen in einer Zusammenfassung aggregiert, um Wechselwirkungen zu berücksichtigen.

5 Zusammenfassung

Wir haben ein P2P-System motiviert, das auf der Beschreibung und Auswahl persönlicher Medienarchive basiert. Medienobjekte werden durch textuelle Inhalte, low-level Inhaltseigenschaften, Zeitstempel und geographische Informationen beschrieben. Gegenstand der Arbeit ist die auf geographischen Informationen basierende Ressourcenauswahl. In den Experimenten zeigt sich, dass eine selektive Auswahl sowohl mit auf einer Organisation der Daten als auch auf einer Organisation des Datenraumes beruhenden Beschreibung für unterschiedliche Anfrageszenarien speicherplatzeffizient möglich ist.

In zukünftigen Arbeiten wollen wir hybride Zusammenfassungstechniken und geeignete Auswahlverfahren analysieren, die vorgestellte Techniken kombinieren. Ferner scheint es spannend, die in dieser Arbeit analysierten Techniken auf anderen Kollektionen zu testen, die weniger stark an P2P-Szenarien angelehnt sind und bei denen die Dokumente in ihrer Anzahl gleichmäßiger auf die Ressourcen verteilt sind.

Literatur

[Alani *et al.*, 2001] H. Alani and C.B. Jones and D. Tudhope. Voronoi-based region approximation for geographical information retrieval with gazetteers. In *Intl. Journal of Geographical Information Science*, 15(4), pages 287–306, 2001.

[Arthur und Vassilvitskii, 2007] D. Arthur and S. Vassilvitskii. The Advantages of Careful Seeding. In *Proc. of the 18th annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.

[Becker *et al.*, 1991] B. Becker and P.G. Franciosa and S. Gschwind and T. Ohler and G. Thiemt and P. Widmayer. An Optimal Algorithm for Approximating a Set of Rectangles by Two Minimum Area Rectangles. In *Intl. Workshop on Computational Geometry*, Vol. 553 of LCNS, Springer, pages 22–29, 1991.

[Belkin *et al.*, 1995] N.J. Belkin and P. Kantor and E.A. Fox and J.A. Shaw. Combining the evidence of multiple query representations for information retrieval. In *Inf. Processing and Management*, 31(3), pages 431–448, 1995.

[Bentley, 1975] J.L. Bentley. Multidimensional binary search trees used for associative searching. In *Commun. ACM*, 18(9), pages 509–517, September 1975.

[Blank und Henrich, 2010] D. Blank and A. Henrich. Description and Selection of Media Archives for Geographic Nearest Neighbor Queries in P2P Networks In *Information Access for Personal Media Archives at ECIR 2010*, pages 22–29, 2010.

[Blank und Henrich, 2012] D. Blank and Andreas Henrich. Describing and Selecting Collections of Georeferenced Media Items in Peer-to-Peer Information Retrieval Systems. In *Diaz, Laura ; Granell, Carlos ; Huerta, Joaquin (eds.): Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications*, pages 1–20, IGI Global, 2012.

[Cuenca-Acuna *et al.*, 2003] F. Cuenca-Acuna and C. Peery and R.P. Martin and T.D. Nguyen. PlanetP: Using gossiping to build content addressable peer-to-peer information sharing communities. In *IEEE Intl. Symp. on High Performance Distributed Computing*, pages 236–246, Seattle, WA, USA, 2003..

[Doulkeridis *et al.*, 2009] C. Doulkeridis and A. Vlachou and K. Nrvag and M. Vazirgiannis. Part 4: Distributed Semantic Overlay Networks. In *Handbook of Peer-to-Peer Networking.*, Springer Science+Business Media, 1st edition, 2009.

[Guttman, 1984] A. Guttman. R-trees: a dynamic index structure for spatial searching. In *SIGMOD Rec.*, 14(2), pages 47–57, June 1984.

[Hariharan *et al.*, 2008] R. Hariharan and B. Hore and S. Mehrotra. Discovering gis sources on the web using summaries. In *JCDL '08: Proc. of the 8th ACM/IEEE joint Conf. on Digital libraries,*, pages 94–103, Pittsburgh, PA, USA, 2008, ACM.

[Hershberger und Suri, 2004] J. Hershberger and S. Suri. Adaptive Sampling for Geometric Problems over Data Streams. In *Proc. of the 23rd ACM sigmod-sigactsigart symposium on principles of database systems*, pages 252–262, 2004.

[Hershberger *et al.*, 2009] J. Hershberger and N. Shrivastava and S. Suri. Summarizing Spatial Data Streams Using ClusterHulls. In *J. Exp. Algorithms*, Vol. 13, pages 26–40, 2009.

[Kolahdouzan und Shahabi, 2004] M. Kolahdouzan and C. Shahabi. Multidimensional binary search trees used for associative searching. In *Proc. of the 30th international conference on very large data bases*, pages 840–851, 2004.

[Müller *et al.*, 2005] W. Müller and M. Eisenhardt and A. Henrich. Scalable summary based retrieval in P2P networks. In *Intl. Conf. on Information and Knowledge Management*, pages 586–593, Bremen, Germany, 2005.

[Nievergelt *et al.*, 1984] J. Nievergelt and H. Hinterberger and K.C. Sevcik. The Grid File: An Adaptable, Symmetric Multi-Key File Structure. In *ACM Transactions on Database Systems*, pages 38–71, 1984.

[Preparata und Shamos, 1985] F.P. Preparata and M.I. Shamos. *Computational Geometry: an Introduction*, Springer-Verlag New York, Inc., New York, NY, USA, 1985.

[Samet, 2005] H. Samet. *Foundations of Multidimensional and Metric Data Structures.*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[Sharifzadeh und Shahabi, 2010] M. Sharifzadeh and C. Shahabi. VoR-tree: R-trees with Voronoi diagrams for efficient processing of spatial nearest neighbor queries. In *Proc. VLDB Endow.*, 3(1-2), pages 1231–1242, September 2010.

[Shen *et al.*, 2009] X. Shen and H. Yu and J. Buford and M. Akon. *Handbook of Peer-To-Peer Networking*, Springer Publishing, 1st edition, 2009.

[Thomas und Hawking, 2009] P. Thomas and D. Hawking. Server selection methods in personal metasearch: a comparative empirical study. In *Information Retrieval*, 12(5): pages 581–604, 2009.