Bachelorarbeit

Clustering von Aufnahmedaten digitaler Fotos zur Erkennung von Ereignissen für ein Bottom-Up Gazetteer

Clustering capture times of digital photos to identify events for a bottom-up gazetteer

Raimund Schnürer

Abgabedatum:

2. September 2010

Betreuer:

Dipl.-Geoinf. Patrick Maué

Prüfer:

Prof. Dr. Werner Kuhn
Dr. Carsten Keßler

Hiermit erkläre ich, dass ich die Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Für meine Oma

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

Fotoalben im Internet stellen eine beliebte Anwendung des Web 2.0 dar, bei der eine Vielzahl digitaler Fotos veröffentlicht wird. Zusätzlich zu den Bildern werden Metadaten wie Schlagwörter oder Freitextbeschreibungen gespeichert. Diese ermöglichen, Fotos zu kategorisieren und für textbasierte Suchmaschinen zu indizieren. Aus den Metadaten können ebenfalls Einträge für geographische Namenslexika gewonnen werden. In diesem Fall spricht man auch von sogenannten Bottom-Up Gazetteers (vgl. Keßler et al. 2009b). Zum Beispiel können die räumlichen Koordinaten aller Fotos mit der Bezeichnung „Münster“ zu einer Punktwolke zusammengefasst werden. Anschließend lassen sich daraus Aussagen über die geometrische Form und die geographische Lage des Ortes ableiten.

Die vorliegende Arbeit betrachtet die zeitliche Dimension von Fotos. Es wird dabei untersucht, ob sich mithilfe der Aufnahmedaten die Dauer von Ereignissen ermitteln lässt. Hierbei werden Veranstaltungen (engl. Events) als eine bestimmte Ausprägung eines Ereignisses fokussiert. Deren Analyse erfolgt mit einem Clustering-Algorithmus, einer gängigen Klassifikationsmethode in der Geoinformatik. Das Verfahren wird anhand fünf ausgewählter Beispiele für Events und vergleichend mit drei Beispielen für Orte getestet.

Zur Visualisierung des Datenmaterials wurde eine graphische Benutzerschnittstelle in Form einer interaktiven Karte implementiert. Diese kann nun mit existierenden Gazetteers verbunden werden. Die Anwendung wurde speziell für Smartphones entwickelt, die sich in den letzten Jahren zunehmender Popularität erfreuen. Aufbauend auf den Ergebnissen der Analyse und der Implementierung werden mögliche Einsatzgebiete und Erweiterungen in Aussicht gestellt.

In dem ersten Kapitel werden zunächst alle für die weiteren Abschnitte notwendigen Grundlagen eingeführt.

1.1 Zeit in der Geoinformatik

Einen grundlegenden Bestandteil der Geoinformatik bilden Geoinformationen. Geoinformationen werden als Tupel <x, z> definiert, wobei x eine Position in Raum und Zeit beschreibt. z formuliert eine Menge von Eigenschaften an dieser Position, sogenannte Attribute (vgl. Goodchild 2003, S. 4). Ein Beispiel für eine Geoinformation ist eine Temperaturmessung, die an einem bestimmten Ort und zu einer bestimmten Zeit vorgenommen wurde. Die Geoinformatik berücksichtigt dabei besonders den räumlichen Bezug von Informationen (vgl. de Lange 2006, S. 4). Zunehmend etablieren sich aber auch die beiden anderen Komponenten Zeit und Thematik. Da sich diese Bachelorarbeit vorwiegend mit dem zeitlichen Aspekt von Geoinformationen beschäftigt, sollen hierfür vorab einige Grundbegriffe erläutert werden.

Das Konzept Zeit lässt sich anschaulich anhand eines Zeitstrahls beschreiben (siehe Abbildung 1). Man unterscheidet zwischen einzelnen Zeitpunkten und Zeitintervallen, die aus dem Abstand zweier Zeitpunkte gebildet werden (vgl. Šumrada 2003, S. 5). Die Zeitachse ist eindimensional und es wird angenommen, dass Zeit auf dieser in positiver Richtung verläuft. Da zeitliche Daten kontinuierlich sind, müssen diese für eine Erfassung zunächst diskretisiert werden (vgl. de Lange 2006, S. 199). Das kleinste Intervall für die Messung eines Zeitabschnitts und für die Definition der Skalenintervalle auf der Zeitachse wird dabei als Chronon bezeichnet. (vgl. Ott u. Swiaczny 2001, S. 59). Dieses legt die zeitliche Auflösung (auch Granularität genannt) fest, welche von Millisekunden bis Jahrtausenden reichen kann. Dabei kann Zeit auf Ordinal- oder Intervallskalenniveau angegeben werden. Auf der Ordinalskala weisen Zeitpunkte oder -intervalle eine chronologische Aufeinanderfolge auf (z.B. in der Geologie). Bei der Intervallskala können zusätzlich die zeitlichen Abstände bestimmt werden (vgl. Šumrada 2003, S. 5). Die beiden Skalen werden entweder in linearer oder aber auch in zyklischer Form (z.B. bei den Jahreszeiten) angegeben (vgl. Ott u. Swiaczny 2001, S. 60).

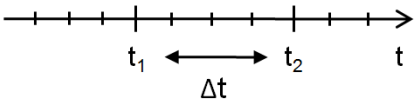

Abbildung 1: Zwei Zeitpunkte t_1 und t_2 und ein Zeitintervall Δt auf einem Zeitstrahl

Um zeitliche Messungen zu vergleichen, benötigt man ein gemeinsames Referenzsystem. Das einfachste benötigt lediglich einen Ursprung und ein Chronon. Komplexere zeitliche Referenzsysteme (z.B. Kalender) folgen bestimmten Regeln für Tage, Monate, usw. (vgl. Chrisman 2002, S. 19). Ein weltweit akzeptiertes, zeitliches Referenzsystem ist die UTC (Universal Time Coordinated). Dieses wird aus einer Kombination des Datums (z.B. 2010-08-09), einer Uhrzeit (z.B. 17:37:00) und einer lokalen Zeitzone (z.B. +02:00) gebildet (vgl. Wolf u. Wicksteed 1998). Ein solch standardisiertes Referenzsystem ist notwendig, um Uhrzeiten in verschiedenen Ländern zu synchronisieren, so auch etwa bei Ereignissen.

1.2 Definition und Klassifikation von Ereignissen

Experten aus unterschiedlichen Fachgebieten - wie der Philosophie, der Mathematik, der Physik, der Chemie, der Medizin und der Informatik - kennen den Begriff eines „Ereignisses“ jeweils unter einer anderen Bedeutung. In dieser Arbeit werden Ereignisse im Sinne von Veranstaltungen, d.h. in Anlehnung an die Freizeit- und Tourismusgeographie, aufgefasst. Im weiteren Verlauf werden daher Ereignis und die heutzutage ebenso gängige Bezeichnung „Event“ synonym verwendet. Eine Definition lautet hierbei wie folgt: “A special event is a onetime or infrequently occurring event outside the normal program or activities of the sponsoring or organizing body. To the costumer, a special event is an opportunity for a leisure, social or cultural experience outside the normal range of choices or beyond everyday experience” (Getz 1991, S. 44). Ergänzend ließe sich anbringen, dass Events meist innerhalb eines zeitlich begrenzten Rahmens stattfinden (vgl. Freyer 2009, S. 639) und dass eine Vielzahl von Menschen daran teilnehmen.

Events lassen sich nach verschiedenen Kriterien klassifizieren (vgl. Metzler u. Job 2007, S. 515). Einen Überblick über unterschiedliche Ereignistypen bietet Abbildung 2. Je nach Größe unterteilt man diese in Mikro-, Medium und Mega-Events. Als wichtiger Indikator für die Größe dient dabei die Anzahl der Teilnehmer (vgl. Dießl 2009, S. 18). Auch zeitlich lassen sich Events einordnen: So unterscheidet man hier nach der Häufigkeit des Auftretens

(einmalig – mehrmalig), der Dauer des Ereignisses (wenige Stunden/eintägig, wenige Tage, zwei bis vier Wochen, mehrere Monate (Metzler u. Job 2007, S. 515)), dem Abstand mehrmaliger Events zueinander (oftmals jährlich (Freyer 2009, S. 639)) und der Regelmäßigkeit des Auftretens (regelmäßig – unregelmäßig). Bei mehrfach auftretenden Ereignissen handelt es sich dabei stets um Beispielausprägungen (Instanzen) eines Ereignistyps (vgl. Galton 2008, S. 915). Ebenfalls in Analogie zu dem Ereignisbegriff in der Geoinformatik kann ein Event aus mehreren Teilereignissen, engl. Subevents, bestehen (vgl. Allen et al. 1995, S. 405). Events sind räumlich auf ein bestimmtes Gebiet lokalisiert, sodass es während eines Ereignisses meist zu keinen größeren Ortsverlagerungen kommt (Ausnahmen sind z.B. einige Sportveranstaltungen wie der Tour de France). Bei mehrmalig stattfindenden Events ist hingegen ein Ortswechsel durchaus möglich (z.B. bei Tourneen). Aus Tradition und zur besseren Identifikation sind Ereignisse oft mit einem bestimmten Ort verbunden (z.B. die CeBIT in Hannover). So können „auch die Veranstaltungsorte und Aufführungsstätten selbst (z.B. der Eiffelturm in Paris) in der Folgezeit als (dauerhafte) Touristenattraktion dienen“ (Freyer 2009, S. 639). Aufgrund des nicht alltäglichen Charakters werden die meisten Fotos jedoch zu den Ereignissen selbst gemacht.

Abbildung 2: Klassifikation verschiedener Eventtypen (nach Freyer 2009, S. 638 u. Getz 2007, S. 404)

1.3 Digitale Fotos im Web 2.0

Digitale Fotos lassen sich als implizite geographische Informationen auffassen. Denn neben dem Bild an sich gibt es noch eine Reihe von Metadaten, die Hintergrundinformationen zu dem Bild liefern. Dazu gehört zunächst ein textueller Kommentar, bestehend aus einem Titel, einer kurzen Beschreibung des Bildes und Schlagwörtern - sogenannten Tags. Erst diese Daten ermöglichen eine gezielte Suche nach bestimmten Fotos. „Wenn viele Nutzer innerhalb eines Systems gemeinsam Tags erzeugen und diese Tags auch anderen Nutzern zugänglich machen, spricht man von einer Folksonomie“ (Stadler 2007, S. 5) oder Collaborative Tagging. Abgesehen von den Tags sind außerdem der Aufnahmeort und das Aufnahmedatum von besonderem Interesse. Nachdem schon seit längerem die Aufnahmezeit eines Fotos automatisch abgespeichert wird, kann nun durch die Entwicklung von Digitalkameras und Mobiltelefonen mit integriertem GPS ebenso der Aufnahmeort erfasst werden. Nach wie vor kann der Aufnahmeort aber auch manuell auf einer digitalen Karte eintragen werden. Ort, Zeit und Schlagwörter genügen, um digitale Fotos als Geoinformationen zu betrachten. Darüber hinaus gibt es noch weitere Attribute, wie der Besucheranzahl eines Fotos, Informationen über den Fotografen und dem Kameramodell, mit dem das Foto aufgenommen wurde.

In Analogie zu den Fotoalben, in denen Fotos eingeklebt und beschriftet werden, bieten nun Web-Fotoalben - wie Flickr[1], Picasa[2] und Panoramio[3] - die passende Oberfläche, um digitale Fotos im Internet zu präsentieren. Eine benutzerfreundliche Bedienung ermöglicht dabei das einfache Hochladen und Verwalten der Fotos. Mit durchschnittlich über 4000 neuen Bildern in der Minute verzeichnen Web-Fotoalben - in diesem Fall Flickr - ein rasches Anwachsen des Datenbestandes (vgl. Lux 2010). Web-Fotoalben sind ein gutes Beispiel für „User-Generated Content“. Durch eine vereinfachte Anwendbarkeit der Technologien ist es heute nahezu für jeden möglich, Inhalte auf Webseiten zu veröffentlichen (vgl. Stein 2007, S. 14). Dadurch ergeben sich positive Netzwerkeffekte, die im Zusammenhang mit dem Begriff des „Web

[1] http://www.flickr.com/
[2] http://picasaweb.google.com/
[3] http://www.panoramio.com/

2.0“ genannt werden (vgl. O'Reilly 2005). Weitere bekannte Beispiele für User-Generated Content und das Web 2.0 sind persönliche Blogs (z.B. Blogger[4] und Twitter[5]), soziale Netzwerke (z.B. Facebook[6] und StudiVZ[7]), Lexika (z.B. Wikipedia[8]), Kartendienste (z.B. OpenStreetMap[9]) und Videoportale (z.B. YouTube[10]). Wenn die bereitgestellten Daten einen geographischen Bezug aufweisen, spricht man auch von „Volunteered Geographic Information“ (VGI) (vgl. Goodchild 2007, S. 212ff.). Bei OpenStreetMap ist dies offensichtlich und ebenfalls bei Wikipedia, da hier Ortsnamen durch die Angabe von geographischen Koordinaten ergänzt werden. Durch den vermehrten Einsatz von Smartphones werden auch die anderen Beispiele, wie Twitter und die Web-Fotoalben, mit freiwilligen Geoinformationen angereichert. VGI stellen für Geoinformatiker ein interessantes Datenmaterial dar, welches man u.a. für Gazetteers verwenden kann (vgl. Keßler et al. 2009a, S. 92).

1.4 Gazetteers

Ein Gazetteer (oder auch Ortslexikon) ist ein „Verzeichnis von geographischen Namen, einschließlich Angaben zur Lage, Objektart und gegebenenfalls weiteren relevanten Informationen zu jedem Eintrag“ (Sievers 2001, S. 301). Dies ermöglicht u.a. zu einem Ortsnamen (z.B. Berlin) dessen Position (z.B. 52.5186°, 13.4081° in WGS84-Koordinaten) und dessen Typ (z.B. Stadt) herauszufinden. „Die früher [für ein Ortslexikon] übliche Veröffentlichung in Buchform wird heute zunehmend ersetzt durch Bereitstellung digitaler Datensätze oder als Datenbank“ (Sievers 2001, S. 301f.). Man spricht auch von „Digital Gazetteers“, die dieselben Kernkomponenten wie ihre Vorgänger besitzen. Beispiele für online-verfügbare Gazetteers sind GeoNames[11], das Getty Thesaurus of Geographic Names[12] (TGN) und das

[4] http://www.blogger.com/
[5] http://twitter.com/
[6] http://www.facebook.com
[7] http://www.studivz.net/
[8] http://de.wikipedia.org
[9] http://www.openstreetmap.org/
[10] http://www.youtube.com
[11] http://www.geonames.org/
[12] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

Geographic Names Information System[13] (GNIS) (vgl. Hill 2000, S. 1f.). Die darin enthaltenen Geodaten werden normalerweise von administrativen Behörden oder von Benutzern manuell eingetragen. Dies bezeichnet man auch als Top-Down Vorgehen. Im Gegensatz dazu versucht ein Bottom-Up Gazetteer die erforderlichen Daten automatisch aus benutzergenerierten Inhalten zu gewinnen, beispielsweise aus den Schlagwörtern und geografischen Koordinaten von digitalen Fotos in Webalben. Hieraus ergeben sich vielfältige Möglichkeiten für Forschungsansätze. So können zum Beispiel geometrische Repräsentationen für geographische Entitäten ermittelt (vgl. Keßler et al. 2009b, S. 83ff.) oder mundartliche Bezeichnungen für nicht genau definierte Regionen (z.B. Innenstadt) untersucht werden (vgl. Wilske 2008, S. 179ff. u. Hollenstein u. Purves 2010, S. 21ff.).

In dieser Bachelorarbeit soll die Frage geklärt werden, ob sich anhand der Aufnahmedaten von Fotos größere, kulturelle Veranstaltungen herausfinden lassen. Hierfür ist eine Erweiterung eines Gazetteers um die zeitliche Dimension nötig. Jedem geographischen Namen wird dabei ein Zeitraum zugeordnet, in welchem dieser gültig ist. Auch Zeiträume können Namen besitzen (z.B. das Mittelalter). In der Informatik hingegen benötigt man aber eher den Start- und Endzeitpunkt eines Intervalls. Dies unterscheidet sich vom Räumlichen, wo meist der Einfachheit halber der Mittelpunkt für einen Ort angegeben wird. Gemeinsam ist jedoch, dass sowohl räumliche als auch zeitliche Angaben stets mit einer gewissen Unsicherheit behaftet sind. Mit Zeit als zusätzliches Attribut in Gazetteers ist es nun möglich, Veränderungen darzustellen und nicht nur statische Informationen zu liefern (vgl. Hill 2006, S. 121f.). Anwendung findet dies beispielsweise bei historischen Orten (vgl. (Kauppinen 2010, S. 15ff.) und historischen Ereignissen (vgl. Allen 2004, S. 72f.). Hier sollen allerdings aktuellere Ereignisse betrachtet werden, die mithilfe eines Clustering-Verfahrens identifiziert werden.

[13] http://gnis.usgs.gov/

1.5 Clustering-Verfahren

In der Geoinformatik wird Clustering als Bestandteil der Clusteranalyse zur Klassifikation von Geodaten verwendet. So wird zum Beispiel das Natural breaks (Jenks) Verfahren in der digitalen Kartographie eingesetzt, um Attributwerte (z.B. Bevölkerungsdichte) in einzelne Klassen einzuteilen (vgl. ArcGIS 9.3 Desktop Help 2008). In der Fernerkundung werden Pixel von Satellitenbildern geclustert, um Geoobjekte voneinander zu unterscheiden (z.B. Nadelwäldern von Laubwäldern) (vgl. Prinz 2007). Ebenso benötigt man Clustering-Algorithmen in der Geostatistik zur Analyse von multivariaten Daten (z.B. Bodenproben) (vgl. Pebesma 2008). In allen drei Fällen ist das Ziel dasselbe: Es sollen Objekte zu Gruppen (Clustern) zusammengefasst werden, wobei Objekte innerhalb einer Gruppe möglichst ähnlich sein sollen. Zwischen einzelnen Gruppen soll möglichst keine Ähnlichkeit bestehen (siehe Abbildung 3). Zur Bestimmung der Ähnlichkeit nutzt man Ähnlichkeits- oder Distanzmaße (vgl. Backhaus et al. 2005, S. 490ff.). Im dreidimensionalen Raum könnte man z.B. dafür die Euklidische Distanz verwenden.

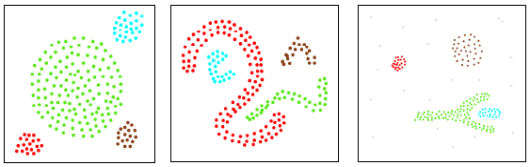

Abbildung 3: Ergebnis eines räumlichen Clustering-Verfahrens (Ester et al. 1996, S. 231)

Im Gegensatz zur überwachten Klassifikation, bei der zunächst manuell Trainingsdaten erzeugt werden (vgl. Buchroithner u. Margraf 2002, S. 57), handelt es sich bei der Cluster-Analyse um eine unüberwachte Klassifkationsmethode, bei dem die Klassen rein statistisch berechnet werden (vgl. Will 2001, S. 122). Man unterscheidet zwischen hierarchischen und partitionierenden Klassifkationsverfahren. Bei der hierarchischen Klassifikation werden nach und nach die zueinander ähnlichsten Objekte in Gruppen zusammengefasst. Dadurch nimmt schrittweise die Clusteranzahl ab, dafür steigt die Objektanzahl in den Gruppen an (vgl. de Lange 2006, S. 99). Ein Beispiel hierfür wäre das Nearest-Neighbor Clustering, bei dem Punkte in einer vorher definierten Umgebung gruppiert werden (vgl. de Smith et al. 2009). Demgegenüber gehen partitionierende Klassifikationsverfahren schon von einer anfänglichen Zerlegung aus und versuchen diese mit einer Zielfunktion zu minimieren (vgl. de Lange 2006, S. 99). Als Beispiel sei das k-means Cluste-

ring genannt, bei dem der Parameter k die Anzahl der Gruppen spezifiziert. Ausgehend von zufällig erzeugten Startpunkten werden anschließend die Cluster iterativ berechnet (vgl. de Smith et al. 2009). Neben den angeführten gibt es eine Vielzahl weiterer Clustering-Algorithmen mit steigender Komplexität. Dazu gehört zum Beispiel das dichte-basierte Verfahren „dbscan“ (vgl. Ester et al. 1996, S. 226). Zur Bestimmung der Distanzmaße können darüber hinaus stochastische Indikatoren, wie der Kullback-Leibler-Divergenz, verwendet werden (vgl. Allan et al. 2000, S. 2).

2 Verwandte Arbeiten

Auch in anderen Bereichen der Informatik werden Clustering-Verfahren eingesetzt. Beispielsweise kann man Webseiten-Suchergebnisse zeitlich clustern, um ältere Nachrichten leichter aufzufinden (vgl. Alonso et al. 2009). Außerdem lassen sich mithilfe von Clustering-Algorithmen alle wichtigen Nachrichten in den Medien innerhalb eines Zeitraumes (z.B. eines Jahres) ermitteln, die man dann überblicksartig auf Zeitleisten darstellen kann (vgl. Swan u. Allan 2000).

Eine andere Anwendungsmöglichkeit von Clustering besteht darin, digitale Fotoalben automatisch zu organisieren (vgl. Graham et al. 2002, Gargi 2003 u. Cooper et al. 2005). Zielsetzung ist hierbei die Entwicklung eines Fotobrowsers, mit dem man einfach und schnell durch eine Kollektion navigieren kann. Dafür ist es notwendig, die Aufnahmedaten der Fotos zeitlich zu clustern und ein repräsentatives Foto für jedes Cluster zu finden. Als Zielgruppen werden hierfür professionelle Fotografen und Urlauber betrachtet.

Auf den Ergebnissen der oben beschriebenen Arbeiten aufbauend, wurde in den letzten Jahren untersucht, wie man aus Multimedia-Inhalten - wie Fotos, Videos, Musik und Texten - Zusammenfassungen von Medienereignissen erstellen kann (vgl. Xie et al. 2008). Aus georeferenzierten Fotos lassen sich zum Beispiel sogenannte Tag Maps zu erstellen, auf denen die bedeutendsten Schlagwörter einer Region angezeigt werden (vgl. Jaffe et al. 2006). Zudem können Geotags für eine räumliche Suche verwendet werden (vgl. Heuer u. Dupke 2007). Darüber hinaus soll die Analyse von Ähnlichkeitsmetriken helfen, die Ergebnisse von Clustering-Verfahren im Umgang mit multimedialen Daten zu verbessern (vgl. Becker et al. 2010).

Zuletzt seien noch einige entferntere Einsatzgebiete von Clustering genannt. So lassen sich etwa in der Genetik durch Clustering-Algorithmen bestimmte DNA-Sequenzen herausfinden (vgl. Schmid et al. 2007). In der Geologie ermöglicht Clustering, endogene Vorgänge wie Erdbeben zu rekonstruieren (vgl. Mukhopadhyaya et al. 2004).

3 Experiment

In diesem Abschnitt wird experimentell untersucht, ob es möglich ist, den Start- und Endzeitpunkt von Events mithilfe von Aufnahmedaten von Fotos zu bestimmen. Hierbei wird zunächst der Algorithmus beschrieben, der diese Aufgabe erfüllen soll. Im Anschluss daran werden ausgewählte Ereignisse präsentiert, welche vom Algorithmus identifiziert werden sollen. Die Ergebnisse des Verfahrens werden schließlich ausgewertet und diskutiert.

3.1 Algorithmus

Um den Beginn und das Ende von Events zu bestimmen, lässt sich die Eigenschaft nutzen, dass während eines Events viele Fotos in kurzen Zeitintervallen aufgenommen werden. Dagegen werden im Zeitraum zwischen zwei Events weniger und in größeren Zeitabständen Fotos gemacht. Unter diesen beiden Annahmen bietet sich zur Erkennung von Events ein Clustering-Algorithmus an, bei dem die Differenz zwischen zwei aufeinanderfolgenden Aufnahmedaten als Distanzmaß verwendet wird. Somit können Fotos innerhalb von Events zu Gruppen zusammengefasst und Fotos zwischen einzelnen Events herausgefiltert werden. Da es sich bei Clustering um ein unüberwachtes Verfahren handelt, kann die Berechnung automatisch erfolgen. Ein partionierender Algorithmus, z.B. das k-means Clustering, kann jedoch nicht verwendet werden, da die Anzahl der Events im Vorhinein nicht bekannt ist. Komplexere Verfahren wurden aufgrund der begrenzten Zeitvorgabe der Bachelorarbeit nicht weiter fokussiert. Es wurde daher ein einfacher Clustering-Algorithmus entwickelt, der sich an hierarchischen Verfahren, wie dem Nearest-Neighbor Clustering, orientiert. Der Algorithmus soll in erster Linie funktional sein, d.h. versuchen, möglichst genau die zeitliche Begrenzung von Ereignissen zu ermitteln. Als „Trade-Offs“ folgen an zweiter und dritter Stelle Laufzeit und Speicherplatz (vgl. Müller-Olm 2008). Weiterhin sollte der Algorithmus mit größeren Mengen von Fotos umgehen können und berücksichtigen, dass Fotos in Webalben nicht nur in chronologischer Reihenfolge hinzugefügt werden.

3.1.1 Beschreibung

1. Ordne die Aufnahmedaten von Fotos chronologisch.

2. Bestimme alle Intervalle zwischen zwei aufeinanderfolgenden Aufnahmedaten und sortiere diese aufsteigend nach ihrer Größe.

3. Ermittle die q% größten Intervalle und ordne diese nach ihrem Endzeitpunkt.

4. Bestimme alle Intervalle, die sich zwischen zwei der q% größten Intervallen befinden, und prüfe jedes davon, ob es sich dabei um ein Teilereignis handelt:
Dies ist der Fall, wenn während des Intervalls eine Mindestanzahl von Fotografen F ein Foto gemacht hat und die Aufnahmerate innerhalb des Intervalls r_i um ein Vielfaches kleiner ist als die durchschnittliche Aufnahmerate r_d.

5. Fasse zwei Teilereignisse zu einem Ereignis zusammen, wenn sie nicht länger als ein bestimmtes zeitliches Intervall I auseinander liegen.

Eine Beschreibung des Algorithmus ist im nebenstehenden Kasten zu finden. Schritte 1 bis 3 wurden dabei mit drei sortierten, balancierten Binärbäumen realisiert, um Einfüge-, Such- und Löschoperationen in logarithmischer Laufzeit zu bewältigen (vgl. Kuchen 2008). Schritte 4 und 5 des Algorithmus werden linear bearbeitet. Neben den Aufnahmedaten selbst ist noch ein Primärschlüssel für das Foto notwendig - im Falle zweier gleicher Aufnahmedaten. Außerdem wird eine Kennung (ID) des Fotografen benötigt, um die Anzahl der Fotografen während eines Teilereignisses bestimmen zu können. Um vorab die Cluster zu bilden, wird ein gewisser Prozentsatz (hier: q = 10) der größten Intervalle entfernt. Die Festlegung einer Mindestanzahl von Fotografen (hier: F >= 3) soll bewirken, dass keine Ereignisse registriert werden, bei denen lediglich ein einzelner Fotograf viele Fotos innerhalb kurzer Zeit macht. Bei einem richtigen Event dagegen ist es umso wahrscheinlicher, dass mehrere Personen fotografieren. Die in Abschnitt 3.1 genannte Eigenschaft wird durch den Vergleich der Aufnahmeraten (hier: $21*r_i <= r_d$) berücksichtigt. Der letzte Punkt ist für mehrtägige Ereignisse vorgesehen (hier: I = 27 Tage), da zwei dicht beieinander liegende Teilereignissen in den meisten Fällen zu demselben Event gehören. Alle Variablen des Algorithmus wurden empirisch ermittelt.

3.1.2 Beispiel

Zur Veranschaulichung des Algorithmus soll folgendes Beispiel dienen. Hierbei wurden vereinfacht die Binärbäume als Listen und die Aufnahmedaten als Zahlen aufgefasst. Der Foto-Primärschlüssel ist über den Aufnahmedaten und die Fotografen-Kennung darunter dargestellt.

Im Schritt 1 werden zuerst die Aufnahmedaten der Größe nach geordnet.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
24	27	28	115	138	143	165	196	199	213	219	234	296	340	357	439	441	442	442	443	444
a	a	b	c	d	d	e	f	f	g	g	e	h	i	i	j	k	k	l	l	l

Nun werden die Abstände zwischen benachbarten Aufnahmedaten berechnet und ebenfalls der Größe nach sortiert. Zur Referenzierung wird darüber die Foto-ID des jeweiligen Intervallendes gespeichert.

S	C	R	T	U	Q	B	I	F	K	J	L	O	G	E	H	N	M	P	D
0	1	1	1	1	2	3	3	5	6	14	15	17	22	23	31	44	62	82	87

Im nächsten Schritt werden die 10% größten Intervalle herausgesucht. Bei 20 Intervallen sind dies genau zwei. Anhand der Referenz können die dazugehörigen Aufnahmezeiten gefunden werden, sodass auch diese sortiert werden können.

D	P
115	439

Jetzt ist es möglich, durch die erste und letzte Liste zu iterieren und Teilereignisse nach den beiden Kriterien herauszusuchen. Zum Vergleich wird vorher noch die durchschnittliche Aufnahmerate berechnet (= (letztes Aufnahmedatum – erstes Aufnahmedatum) / Anzahl aller Intervalle). Diese beträgt hier: (444 – 24) / 20 = 21. Eine Prüfung des ersten Teilereignisses A-C fällt negativ aus, da die Anzahl der Fotografen (= 2) kleiner ist als die vorgegebene Anzahl von 3. Auch das Teilereignis D-O wird herausgefiltert. Die Anzahl der Fotografen ist zwar höher (= 7), dafür liegt die Aufnahmerate während des Intervalls mit 22 sogar über dem Durchschnitt. Lediglich das Subevent P-U erfüllt beide Vorgaben. Hier ist sowohl die Mindestanzahl an Fotografen (= 3) erreicht, als auch ist die Aufnahmerate mit 1 genau 21-mal kleiner als die durchschnittliche.

Da sich keine weiteren Teilereignisse in Nachbarschaft des ermittelten befinden, entfällt Schritt 5, sodass das Teilereignis als Ereignis ausgegeben werden kann.

Hinweis: In der Implementierung werden bei Veränderungen (z.B. dem Hinzufügen eines Fotos) nicht alle Elemente in den Schritten 1 bis 3 neu sortiert, sondern lediglich die betreffenden Stellen reorganisiert.

3.2 Datenmaterial

Um den beschriebenen Algorithmus zu testen, wurden folgende fünf Events betrachtet:

- Raumfährenstarts in Cape Canaveral
- die Eröffnung des „Burj Khalifa“, des momentan höchsten Gebäudes der Welt
- das Musikfestival „Wacken Open Air“
- die beiden letzten Fußball-Weltmeisterschaften in Deutschland und Südafrika
- die vergangene Expo in Spanien und die derzeitige Weltausstellung in China

Diese Ereignisse wurden beispielhaft ausgewählt, da sie nach verschiedenen Kriterien klassifiziert werden können (siehe Tabelle 1).

Ereignis	Cape Canaveral	Burj Khalifa	Wacken	FIFA World Cup	Expo
Dauer	wenige Minuten	ein paar Stunden	drei Tage	ein Monat	mehrere Monate
Regelmäßigkeit	unregelmäßig	-	regelmäßig	regelmäßig	unregelmäßig
Häufigkeit	mehrmalig	einmalig	mehrmalig	mehrmalig	mehrmalig
Abstand	ca. 1-6 Monate	-	einmal im Jahr	alle vier Jahre	ca. 2-4 Jahre
Ort	gleichbleibend	-	gleichbleibend	wechselnd	wechselnd
Größe	Mikro-Event	Medium-Event	Mega-Event	Mega-Event	Mega-Event

Tabelle 1: Zeitliche und räumliche Klassifikation der Beispiel-Events

Als Gegenbeispiele wurden Aufnahmedaten von Fotos an Orten untersucht, die permanent vorhanden sind:

1. der Amazonas, einem der längsten Flüsse der Erde und Name von Bundesstaaten in mehreren südamerikanischen Ländern
2. Mombasa, die zweitgrößte Stadt Kenias
3. der Uluru (ehemals Ayers Rock), eine bekannte Felsformation in Australien

Alle Daten wurden mithilfe der Flickr-API[14] gesammelt. Diese ermöglicht es unter anderem, Fotos gezielt nach Schlagwörtern zu suchen und dabei den Aufnahmezeitraum einzugrenzen. Um die Stichproben zusätzlich für ein Bottom-Up Gazetteer zu verwenden, wurden nur georeferenzierte Fotos extrahiert. Als Rückgabe liefert die Flickr-API ein XML-Dokument, welches die

[14] http://www.flickr.com/services/api/

angeforderten Parameter erfüllt. Aufnahmedaten[15] werden im UTC-Format zurückgegeben, allerdings ohne Angabe einer Zeitzone. Dadurch liegen alle Zeiten in Ortszeit vor. Insgesamt handelt es sich um intervallskalierte Zeiten, bei denen das Chronon eine Sekunde beträgt.

3.3 Ergebnisse

Nachfolgend sind sowohl die Start- und Endzeitpunkte der tatsächlich stattgefundenen Events aufgeführt als auch diejenigen, welche der Clustering-Algorithmus identifiziert hat. Ebenfalls sind die Suchkriterien ersichtlich, die für eine Anfrage an die Flickr-API[14] verwendet wurden. Einige Rahmendaten, wie der Anzahl der Fotos, geben Auskunft über die Größe der Stichprobe.

Raumfährenstarts in Cape Canaveral			
Zeitraum	01.01.2010 – 01.09.2010	**Suchbegriff**	Cape Canaveral
Tatsächliche Events[16]	08.02.2010, 4.14 Uhr: Raumfähre „Endeavour“	**Identifizierte Events**	06.02.2010, 8.31 Uhr – 08.02.2010, 5.13 Uhr
	05.04.2010, 6.21 Uhr: Raumfähre „Discovery“		04.04.2010, 23.32 Uhr – 05.04.2010, 13.04 Uhr
	14.05.2010, 14.20 Uhr: Raumfähre „Atlantis“		13.05.2010, 9.39 Uhr – 15.05.2010, 0.52 Uhr
Fotos	523	**Fotografen**	56

Tabelle 2: Raumfährenstarts in Cape Canaveral

Wie man sieht, entspricht die Anzahl der registrierten Events der Anzahl der tatsächlichen Events. Zudem sind alle Raumfährenstarts im gefundenen Zeitraum inkludiert. Die Starts liegen relativ am Ende des Zeitraums, was den Verdacht äußert, dass Zuschauer schon ein bis zwei Tage früher zu dem Event anreisen. Dafür machten sich Fotos von einem in der Nähe befindlichen Museum, dem Kennedy Space Center[17], nicht bemerkbar.

Eröffnungsfeier des Burj Khalifa			
Zeitraum	01.09.2004 – 01.09.2010	**Suchbegriffe**	Burj, Opening
Tatsächliche Events[18]	04.01.2010, abends	**Identifizierte Events**	04.01.2010, 20.29 Uhr – 04.01.2010, 20.49 Uhr
Fotos	45	**Fotografen**	4

Tabelle 3: Eröffnungsfeier des Burj Khalifa

[15] http://www.flickr.com/services/api/misc.dates.html
[16] http://spaceflightnow.com/tracking/launchlog.html
[17] http://www.kennedyspacecenter.com/
[18] http://www.presseportal.de/pm/62377/1539746/government_of_dubai_department_of_tourism_and_commerce_marketing

Trotz der sehr kleinen Stichprobe wurde die Zeit dieses einmaligen Ereignisses richtig erkannt. Der Suchzeitraum entspricht in etwa der Bauzeit des Burj Khalifa. Allerdings hatte dies in positiver Hinsicht keine Auswirkungen auf den Clustering-Algorithmus.

Musikfestival „Wacken Open Air“			
Zeitraum	01.01.2007 – 15.08.2010	**Suchbegriff**	Wacken
Tatsächliche Events[19]	02.08.2007 – 04.08.2007	**Identifizierte Events**	02.08.2007, 17.07 Uhr – 04.08.2007, 22.54 Uhr
	31.07.2008 – 02.08.2008		30.07.2008, 19.02 Uhr – 03.08.2008, 00.02 Uhr
	30.07.2009 – 01.08.2009		29.07.2009, 19.56 Uhr – 02.08.2009, 00.30 Uhr
	05.08.2010 – 07.08.2010		04.08.2010, 14.30 Uhr – 07.08.2010, 22.08 Uhr
Fotos	2779	**Fotografen**	60

Tabelle 4: Musikfestival „Wacken Open Air“

Analog zu Cape Canaveral scheinen die Zuschauer zu diesem Event schon einen Tag früher anzureisen. Ansonsten ist das Ergebnis aber sehr zufriedenstellend. Außerdem wurde das letzte Festival bereits richtig erkannt, obwohl es erst eine Woche vor der Anfrage stattgefunden hat.

Fußball-Weltmeisterschaften in Deutschland und Südafrika			
Zeitraum	01.01.2006 – 01.09.2010	**Suchbegriff**	FIFA World Cup
Tatsächliche Events[20]	09.06.2006 – 09.07.2006: Fußball-WM in Deutschland	**Identifizierte Events**	12.06.2006, 15.59 Uhr – 10.07.2006, 02.37 Uhr
	11.06.2010 – 11.07.2010: Fußball-WM in Südafrika		12.06.2010, 12.41 Uhr – 11.07.2010, 23.26 Uhr
Fotos	3416	**Fotografen**	141

Tabelle 5: Fußball-Weltmeisterschaften in Deutschland und Südafrika

Alles in allem wurde auch der Zeitrahmen dieses Ereignis recht gut erkannt. Der Anfang ist zwar in beiden Fällen zu spät, dafür ist das Ende durch das Finale sehr genau. Eine mögliche Beeinträchtigung des Ergebnisses durch die Vorrundenspiele der Weltmeisterschaften kam nicht zustande.

[19] http://www.wacken.com/de/woa2011/main-history/history/
[20] http://www.fifa.com/worldcup/archive/index.html

Weltausstellungen in Spanien und China			
Zeitraum	01.01.2008 – 01.08.2010	**Suchbegriffe**	Expo, Spain + Expo, China
Tatsächliche Events	14.06.2008 – 14.09.2008: Expo in Spanien[21]	**Identifizierte Events**	18.06.2008 01.08.2008 – 13.09.2008
	01.05.2010 – 31.10.2010: Expo in China[22]		24.04.2010 – 22.07.2010
Fotos	2340 + 4020 = 6360	**Fotografen**	300

Tabelle 6: Weltausstellungen in Spanien und China

Bei diesem Event war der Algorithmus nur mäßig erfolgreich. Die Ursache liegt dafür wahrscheinlich an der höheren Anzahl von Fotos bei der Expo in China. Dies hat eine geringere Aufnahmerate zur Folge, wodurch mehrere Subevents während der Expo in Spanien herausgefiltert wurden. Die zeitliche Begrenzung liegt in Anbetracht dieses mehrmonatigen Events noch im Rahmen. Dieses Beispiel wurde unter anderem auch ausgewählt, da die Expo in China noch andauert.

Amazonas			
Zeitraum	01.07.2008 – 01.07.2010	**Suchbegriff**	Amazonas
Erwartete Events	keine	**Identifizierte Events**	26.07.2008 10:59 – 26.07.2008 19:06
			05.09.2008 07:39 – 05.09.2008 17:37
			30.10.2009 11:35 – 31.10.2009 18:54
			10.01.2010 03:05 – 13.01.2010 05:02
Fotos	3707	**Fotografen**	241

Tabelle 7: Fluss und Bundesstaat Amazonas

Als erstes Beispiel für einen Ort, an dem keine Ereignisse vermutet wurden, fand der Algorithmus dennoch vier mögliche Events heraus. Eine genauere Analyse der Fotos ergab jedoch, dass es sich dabei um keine Events handelte. Meist hatte einer der Fotografen sehr viele Fotos (teilweise über 100) auf einmal aufgenommen und alle Fotos unter derselben Bezeichnung veröffentlicht. Zur gleichen Zeit haben zufällig noch zwei bis drei andere Leute etwas fotografiert, was den Identifikationsmechanismus des Algorithmus auslöste.

[21] http://expomuseum.com/2008/
[22] http://expomuseum.com/2010/

Mombasa			
Zeitraum	01.07.2008 – 01.07.2010	**Suchbegriff**	Mombasa
Erwartete Events	keine	**Identifizierte Events**	keine
Fotos	2124	**Fotografen**	99

Tabelle 8: Die Stadt Mombasa

Im Gegensatz zum Beispiel des Amazonas wurden in Mombasa keine Ereignisse registriert. Auffällig bei beiden Gegenbeispielen ist allerdings, dass es mehr verschiedene Fotografen gibt als bei den richtigen Events.

Uluru			
Zeitraum	01.07.2008 – 01.07.2010	**Suchbegriff**	Uluru
Erwartete Events	keine	**Identifizierte Events**	02.01.2009 12:35 – 02.01.2009 22:36
			03.04.2009 08:39 – 04.04.2009 11:15
Fotos	2906	**Fotografen**	253

Tabelle 9: Die Felsformation Uluru

Das Ergebnis des Ulurus liegt in der Mitte von den beiden vorangegangen. Die identifizierten Ereignisse lassen sich analog zum Amazonas-Beispiel erklären. Es sei bemerkt, dass bei allen drei Gegenstichproben der Suchzeitraum über zwei Jahre gewählt wurde, um möglicherweise Regelmäßigkeiten bei auftretenden Events zu finden.

3.4 Beurteilung

Insgesamt sind die Ergebnisse des Clustering-Algorithmus akzeptabel. Die Anzahl der identifizierten Events stimmt mit den tatsächlich stattgefundenen fast immer überein. Zudem entsteht bei Start- und Endzeitpunkt meist nur eine Abweichung von wenigen Tagen. Allerdings muss die Ermittlung von Events mit unterschiedlichen Fotodichten noch verbessert werden. Auch die Abgrenzung von fortwährend existierenden Orten hat nicht in allen Fällen funktioniert. Hier wurden Ereignisse registriert, die es in Wirklichkeit nicht gegeben hat.

3.4.1 Verbesserungsvorschläge

Ein erster Ansatz zur besseren Differenzierung zwischen Orten und Events besteht in der Skalierung oder Normalisierung von Fotos eines Benutzers, sodass der Einfluss von sogenannten „Bursts“ (Graham et al. 2002, S. 328)

verringert werden kann. Hierbei sei allerdings zu beachten, dass keine Informationen über die tatsächlichen Events verloren gehen. Des Weiteren könnten auch die Verteilungen der Fotos während eines Ereignisses im Vergleich zu denen von Orten untersucht werden. Man könnte annehmen, dass die Fotos bei Events generell eher einer Normalverteilung folgen bzw. mit einer gewissen Schiefe oder einem bestimmten Exzess (vgl. Sachs u. Hedderich 2006, S. 155) behaftet sind (wie im Falle der Raketenstarts in Cape Canaveral). Bei Orten hingegen würde man eher eine Gleichverteilung vermuten. Zur Behebung des Problems der verschiedenen Fotodichten könnte man einen anderen Clustering-Algorithmus ausprobieren. Beispielsweise wäre „OPTICS“ (vgl. Ankerst et al. 1999, S. 49ff.) ein guter Kandidat dafür. Schließlich könnte eine Verbesserung bei Events mit wechselnden Aufführungsstätten erzielt werden (wie der Fußball-Weltmeisterschaft), indem man räumlich-zeitliches Clustering anwendet (vgl. de Smith et al. 2009). Ebenso sind höher-dimensionale Clustering-Varianten unter Einbeziehung weiterer Variablen wie Tags (semantisches Clustering), Fotografen (soziales Clustering) bis hin zu den Farben eines Fotos denkbar.

3.4.2 Fazit

Alles in allem eignet sich der Algorithmus für eine Erweiterung eines Bottom-Up Gazetteers. Gemeinsam mit den Ergebnissen des räumlichen Clusterings (vgl. Keßler et al. 2009b, S. 97ff.) können nun zu einem Ereignis mit eindeutigem Ortsbezug folgende Angaben gemacht werden:

Name	Geographische Position	Geometrische Repräsentation	Anfangszeitpunkt	Endzeitpunkt

Zusätzlich lässt sich leicht über die Anzahl der Fotos und Fotografen die Größe eines Events oder die Bedeutung eines Ortes ermitteln. Weitere Ergänzungsmöglichkeiten zu Bottom-Up Gazetteers sind im Abschnitt 5.2 beschrieben.

4 Anwendung

Zur Kombination des räumlichen Clusterings (vgl. Keßler et al. 2009b, S. 97ff.) und des hier entwickelten zeitlichen Clustering-Algorithmus wurde eine graphische Benutzeroberfläche für ein Bottom-Up Gazetteer implementiert. Hierbei können eingetragene Orte und Ereignisse auf einer interaktiven Karte visualisiert werden. Dadurch ist eine anschaulichere und leichtere Zugriffsmöglichkeit auf die Daten gegeben, als dies in tabellarischer Form der Fall wäre. So können zum Beispiel Touristen, als eine potentielle Benutzergruppe, interessante Orte und Events in ihrer Umgebung finden.

Die Anwendung ist für Smartphones konzipiert, wie dem iPhone[23] von Apple, dem HTC Desire[24] oder dem Xperia X10[25] von Sony-Ericsson. Jedes dieser Mobilfunkgeräte zeichnet sich neben einem leistungsstarken Prozessor und einem großem Display noch durch weitere Funktionen, wie Internetzugang und einem GPS-Empfänger, aus (vgl. Cassavoy o.J.). Das handliche Format eines Smartphones erlaubt dabei vielfältige Einsatzmöglichkeiten. Eine davon ist die Darstellung von Karten, auf die im folgenden Abschnitt eingegangen wird.

4.1 Kartenentwicklung für Smartphones

Trotz der relativen Neuartigkeit gibt es bereits einige Möglichkeiten, benutzerspezifische Karten für Smartphones anzufertigen. Zum Einen bietet Google auf verschiedene Zwecke ausgelegte Varianten der Kartenentwicklung an. So können mit der dritten Version der Google Maps API[26] erstellte Karten nicht nur wie gewohnt auf Websites eingebunden werden, sondern auch auf Smartphones abgerufen werden. Ein Vorteil dieser API ist die ausführliche Dokumentation der darin enthaltenen Kartenobjekte sowie deren Methoden und Eigenschaften. Ebenso lassen sich bei der Applikation Google Maps[27] des Betriebssystems Android Karten individuell anpassen. Allerdings werden

[23] http://www.apple.com/de/iphone/
[24] http://www.htc.com/de/product/desire/overview.html
[25] http://www.sonyericsson.com/cws/products/mobilephones/overview/xperiax10
[26] http://code.google.com/intl/de-DE/apis/maps/documentation/v3/
[27] http://www.google.com/intl/de_ALL/mobile/maps/

hier nur wenige Funktionen zur Verfügung gestellt, sodass nur einfache Kartenelemente hinzugefügt werden können. Für Programmierer hingegen hat Google eine Erweiterung der Entwicklungsumgebung für Android vorgesehen[28], um Karten in Anwendungen für Smartphones zu integrieren. Eine Schnittstelle mit ähnlichem Umfang bieten die Ericsson Labs mit Mobile Maps[29] an. Im Gegensatz zu Google werden jedoch OpenStreetMap-Daten für die Kartendarstellung verwendet. Weitere Alternativen zur Anzeige von Karten auf Smartphones werden von Cloudmade[30] und Nutiteq[31] bereitgestellt.

4.2 Technische Details der Implementierung

Von den oben beschriebenen Möglichkeiten wurde für die Implementierung der graphischen Benutzeroberfläche die Google Maps API V3[22] ausgewählt. Hauptkriterien waren dabei die flexible Einsatz, der Funktionsumfang und die einfache Bedienbarkeit. Generell definiert eine API (= Application Programming Interface) „einen Satz von Funktionen (mit ihrer Semantik) und ihre Schnittstellen (Parameter). Die API-Definition ist oft auf eine Programmiersprache bezogen“ (Rechenberg u. Pomberger 2006, S. 701). So basiert zum Beispiel die API von Google Maps auf Javascript und HTML. Um die Position des Benutzers per GPS zu bestimmen, wurden zusätzlich sogenannte Geolocation APIs verwendet. Zum einen wird eine solche vom World Wide Web Consortium[32] spezifiziert und zum anderen Google Gears[33]. Weiterhin hat sich zum Auslesen der Daten des Gazetteers, welche im GeoJSON-Format[34] übermittelt werden, die jQuery API[35] bewährt. Um die Anwendung im Internet zu testen, wurde die Google-App-Engine[36] verwendet. Hier können bis zu einer gewissen Größe und Auslastung selbst entwickelte Applikationen kostenfrei veröffentlicht werden.

[28] http://code.google.com/intl/de-DE/android/add-ons/google-apis/index.html
[29] https://labs.ericsson.com/apis/mobile-maps/
[30] http://cloudmade.com/
[31] http://www.nutiteq.com/
[32] http://dev.w3.org/geo/api/spec-source.html
[33] http://code.google.com/intl/de-DE/apis/gears/api_geolocation.html
[34] http://geojson.org/geojson-spec.html
[35] http://api.jquery.com/jQuery.getJSON/
[36] http://code.google.com/intl/de-DE/appengine/

4.3 Allgemeine Beschreibung der Funktionalität

Nach dem Erscheinen eines Begrüßungsbildschirms bestimmt die Anwendung nach Einverständnis des Nutzers dessen gegenwärtige Position und zeigt diese auf der Karte an (siehe Abbildung 4). Die Position wird entweder per GPS oder über die bestehende Internetverbindung ermittelt. Im weiteren Verlauf aktualisiert sich der Standort des Benutzers automatisch, wenn dieser sich bewegt. Innerhalb des Kartenausschnitts wird eine bestimmte Anzahl von Orten aus dem Bottom-Up Gazetteer herausgesucht und als Standortmarker repräsentiert. Die Marker erscheinen dabei umso größer, je mehr Fotografen an einem Ort fotografiert haben. Beim Zoomen oder Verschieben der Karte werden neue Orte dem Gazetteer entnommen. Wenn man auf einen Ortsmarker klickt, öffnet sich eine Sprechblase mit dessen Namen und möglichen Events, die mithilfe des entwickelten Clustering-Algorithmus identifiziert wurden. Zur geometrischen Repräsentation des Ortes wird ein Polygon als zusätzliche Ebene über die Karte gelegt (siehe Abbildung 5). Darüber hinaus werden Fotos des Ortes ausgegeben, sodass sich der Benutzer einen besseren Eindruck verschaffen kann. Die Fotos werden ebenfalls zunächst als Standortmarker angezeigt, wobei die Größe der Marker proportional zur Besucheranzahl der Fotos ist. Klickt man einen Fotomarker an, so öffnet sich eine Sprechblase mit dem Foto (siehe Abbildung 6). Dieses ist mit der ursprünglichen Seite im Web-

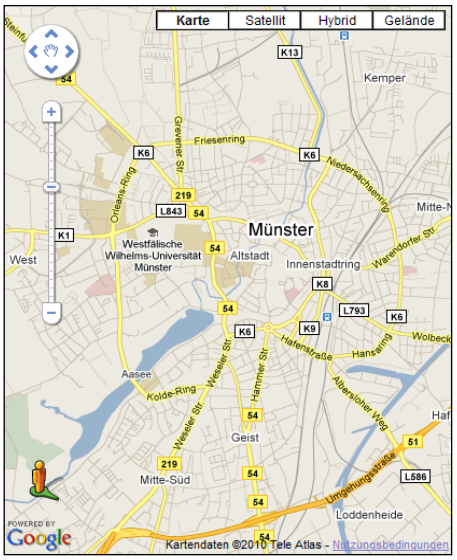

Abbildung 4: Bestimmung und Anzeige der gegenwärtigen Position

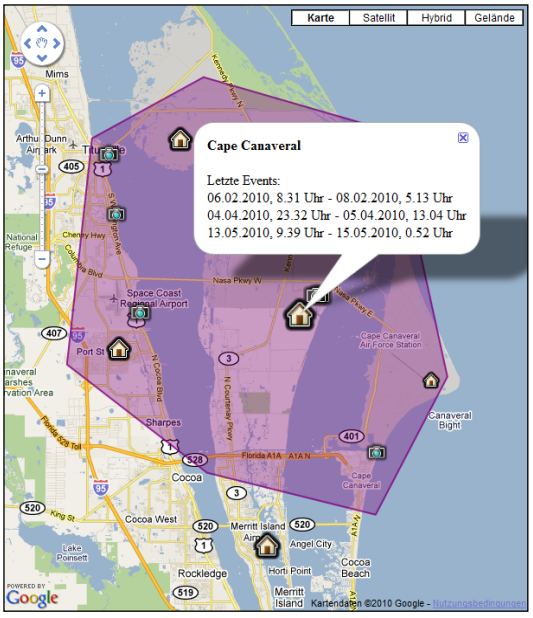

Abbildung 5: Darstellung der Geometrie und Events eines Ortes

Fotoalbum verlinkt, damit sich der Nutzer über Titel oder Aufnahmedatum des Bildes informieren kann.

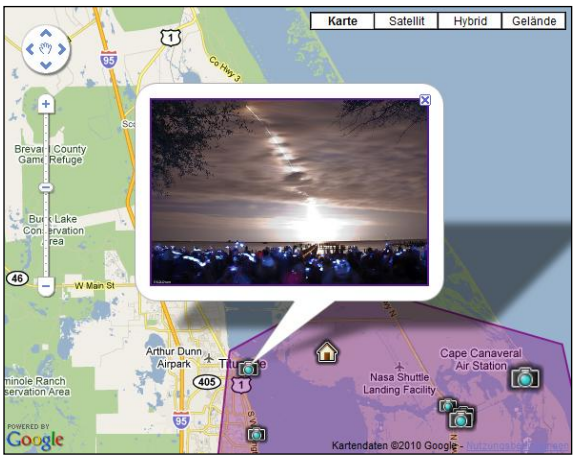

Abbildung 6: Öffnen eines Fotomarkers auf der interaktiven Karte

4.4 Zukünftige Arbeit

Zum Testen der Anwendung, die unter http://emergent-places.appspot.com abrufbar ist, wurden beispielhafte Daten verwendet. Die volle Funktionalität wird erreicht, wenn ein Webservice des Bottom-Up Gazetteers bereit steht. Auf Seiten der graphischen Benutzeroberfläche ist hierfür lediglich nötig, die Schnittstellen entsprechend zu konfigurieren.

Es ist denkbar, die Anwendung durch eine Suchfunktion zu erweitern. Diese würde dem Benutzer gestatten, schneller zu einem bestimmten Ort zu gelangen, als es über die Kartennavigation der Fall wäre. Zur Interaktion mit der zeitlichen Dimension von Fotos und Events könnte zusätzlich eine Zeitleiste (z.B. Simile[37]) eingebunden werden. Bei beiden Vorschlägen muss berücksichtigt werden, diese auf die Größe von Smartphones anzupassen.

[37] http://www.simile-widgets.org/timeline/

5 Zusammenfassung und Ausblick

Diese Arbeit zeigt, dass durch Anwendung eines Clustering-Verfahrens auf Aufnahmedaten digitaler Fotos Anfangs- und Endzeitpunkte von Ereignissen bestimmt werden können. Der dafür entwickelte Algorithmus liefert bereits gute Resultate, die für eine Erweiterung eines Bottom-Up Gazetteers um die zeitliche Dimension genutzt werden können. Mit den angesprochenen Verbesserungsvorschlägen kann der Algorithmus weiter verfeinert werden. Somit können in Bottom-Up Gazetteers zusätzlich zu Orten, die über längere Zeiträume existieren, auch weitaus kürzer andauernde Ereignisse erfasst werden. Zur Visualisierung dieses Datenmaterials wurde eine interaktive Karte erstellt, die über das Internet abgerufen werden kann. Dies ermöglicht nicht nur die Ergebnisse des räumlichen und zeitlichen Clustering anschaulich zu validieren, sondern auch zu weiterführenden Anwendungen, die nachfolgend kurz beschrieben werden.

5.1 Einsatzgebiete

Eine sehr naheliegende Anwendung eines Bottom-Up Event Gazetteers besteht darin, Ereignisse der nahen Vergangenheit zu recherchieren. Dies könnte z.B. für Historiker von Nutzen sein, die das Gazetteer als eine alternative Quelle zu den üblichen Medien heranziehen können. Bei Events, die in bestimmten zeitlichen Mustern auftreten, kann eine Vorhersage für zukünftige Ereignisse getroffen werden. Beispielsweise ist es beim Musikfestival in Wacken sehr wahrscheinlich, dass dieses auch im nächsten Jahr gegen Ende Juli oder Anfang August stattfindet. Die Untersuchung der zeitlichen Abfolge könnte gegebenenfalls vom Computer vorgenommen werden. Zusammen mit der Identifizierung von sich gerade abspielenden Ereignissen würde dies eine gute Grundlage für einen Dienst, vergleichbar mit einem Veranstaltungsführer, schaffen, bei dem sich Benutzer über Events in ihrer Nähe informieren können. Beim „on-the-fly“ Herausfinden von Events anhand von Fotos muss allerdings beachtet werden, dass die meisten Fotos erst mit einer gewissen zeitlichen Verzögerung in Web-Fotoalben hochgeladen werden. Dadurch könnten nur länger andauernde Ereignisse in Echtzeit erfasst

werden, kürzere jedoch eher nicht. Als Alternative zu Fotos in Webalben bieten sich Einträge bei Twitter an, da diese ebenfalls zunehmend mit Smartphones erstellt werden. Durch eine entsprechende Auswertung der Schlagwörter und des Textes können so auch auf diesem Weg Ereignisse detektiert werden. In eine andere Forschungsrichtung geht die Kombination von Fotoaufnahmen mit Messungen von Sensoren, um so natürliche Ereignisse (z.B. Naturkatastrophen) zu registrieren. Ob natürliche Events in Gazetteers aufgenommen werden, ist aufgrund ihres räumlich sehr dynamischen Charakters jedoch fraglich (vgl. Hill 2000, S. 287).

5.2 Erweiterungsmöglichkeiten von Bottom-Up Gazetteers

Neben der zeitlichen Komponente können noch weitere Elemente zu Bottom-Up Gazetteers hinzugefügt werden. Ein wichtiger, dafür etwas schwieriger zu ermittelnder Bestandteil ist der Typ des Ortes oder in diesem Falle von Events. Hierbei könnte eine semi-automatische Herangehensweise mithilfe von Tags hilfreich sein (vgl. Keßler et al. 2009a, S. 96f.). Von ähnlicher Bedeutung ist es, für dasselbe Event oder denselben Ort synonyme Bezeichnungen zu finden (wie bei Ayers Rock und Uluru) bzw. den Namen in verschiedenen Sprachen anzugeben (vgl. Hill 2000, S. 285). Auch eine Analyse des sozialen Aspekts (z.B. anhand von Benutzerstudien) erscheint interessant, um die Frage „Wer ist wann wo?“ zu beantworten (vgl. Keßler et al. 2009b, S. 99). Eine weitere Ergänzung im Räumlichen ist die Bestimmung von Hierarchien und Beziehungen zwischen Orten (z.B. „Mombasa liegt in Kenia“). Dies ließe sich mithilfe der geometrischen Darstellung ermitteln (vgl. (Hill 2006, S. 120f.). Zuletzt sei die Frage der Qualität der Ergebnisse angesprochen, welche räumlichen oder zeitlichen Abweichungen vom tatsächlichen Ort oder Event zu erwarten sind.

Anhang

- Quellcode und Beispieldateien des implementierten Clustering-Algorithmus
- Quellcode und Beispieldateien der implementierten graphischen Benutzerschnittstelle zu einem Bottom-Up Gazetteer

Literaturverzeichnis

Allan, J. et al. (2000): Detections, Bounds, and Timelines: UMass and TDT-3. In: Proceedings of Topic Detection and Tracking Workshop (TDT-3). Wien, S. 167-174

Allen, E. et al. (1995): Qualitative Causal Modeling in Temporal GIS. In: Frank, A. u. W. Kuhn (Hrsg.) (1995): Spatial Information Theory A Theoretical Basis for GIS. International Conference COSIT '95 Semmering, Austria, September 21–23, 1995 Proceedings. Berlin, Heidelberg, New York, S. 397–412

Allen, R. (2004): A Query Interface for an Event Gazetteer. In: Chen, H. et al. (Hrsg.) (2004): Global Reach and Diverse Impact. Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries. Tucson, Arizona, June 7 - 11, 2004. New York, S. 72–73

Alonso, O. et al. (2009): Clustering and Exploring Search Results using Timeline Constructions. In: Cheung et al. (Hrsg.) (2009): Conference on Information and Knowledge Management. Proceeding of the 18th ACM conference on Information and knowledge management. New York, S. 97–106

Ankerst, M. et al. (1999): OPTICS: Ordering Points To Identify the Clustering Structure. In: Davidson, S. u. C. Faloutsos (Hrsg.) (1999): ACM SIGMOD Record. Volume 28, Issue 2. New York, S. 49–60

ArcGIS 9.3 Desktop Help (2008): Natural breaks (Jenks). Online unter: http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?topicname=natural_breaks_%28jenks%29 (abgerufen am: 04.08.2010)

Backhaus, K. et al. (2005): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung (11. Auflage). Berlin, Heidelberg

Becker, H. et al. (2010): Learning Similarity Metrics for Event Identification in Social Media. In: Davison, B. et al. (Hrsg.) (2010): Web Search and Web Data Mining. Proceedings of the third ACM international conference on Web search and data mining. New York, S. 291–300

Buchroithner, M. u. O. Margraf (2002): Klassifikation. In: Bollmann, J. u. W. Koch (Hrsg.) (2002): Lexikon der Kartographie und Geomatik. Karto bis Z. Heidelberg, Berlin, S. 57

Cassavoy, L. (o.J.): What Makes a Smartphone Smart? We tackle the question: Just what is a smartphone, and why is it so smart? Online unter: http://cellphones.about.com/od/smartphonebasics/a/what_is_smart.htm (abgerufen am: 27.08.2010)

Chrisman, N. (2002): Exploring geographic information systems (2. Auflage). New York

Cooper, M. et al. (2005): Temporal Event Clustering for Digital Photo Collections. In: Georganas, N. (Hrsg.) (2005): ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP). New York, S. 269–288

Dießl, E. (2009): Management von Sportgroßveranstaltungen: Unter besonderer Berücksichtigung des Stakeholdermanagements. Hamburg

Ester, M. et al. (1996): A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E. et al. (Hrsg.) (1996): Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). Menlo Park, S. 226–231

Freyer, W. (2009): Tourismus-Marketing. Marktorientiertes Management im Mikro- und Makrobereich der Tourismuswirtschaft (6. Auflage). München

Galton, A. (2008): Processes and Events. In: Shekhar, S. u. H. Xiong (Hrsg.) (2008): Encyclopedia of GIS. New York, S. 913–917

Gargi, U. (2003): Consumer Media Capture: Time-Based Analysis and Event Clustering. Online unter: www.hpl.hp.com/techreports/2003/HPL-2003-165.pdf (abgerufen am: 26.06.2010)

Getz, D. (1991): Festivals, special events and tourism. New York

Getz, D. (2007): Event tourism: Definition, evolution, and research. In: Tourism Management 29 (2008), S. 403–428

Goodchild, M. (2003): The Nature and Value of Geographic Information. Online unter: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.161.9130&rep=rep1&type=pdf (abgerufen am 05.08.2010)

Goodchild, M. (2007): Citizens as sensors: the world of volunteered geography. In: GeoJournal 69 (4), S. 211–221

Graham, A. et al. (2002): Time as Essence for Photo Browsing Through Personal Digital Libraries. In: Hersh, W. (Hrsg.) (2002): International Conference on Digital Libraries. Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries. New York, S. 326–335

Heuer, J. u. S. Dupke (2007): Towards a Spatial Search Engine Using Geotags. In: Probst, F. u. C. Keßler (Hrsg.) (2007): GI-Days 2007 - Young Researchers Forum. Proceedings of the 5th Geographic Information Days. Münster, S. 199–204

Hill, L. (2000): Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In: Borbinha, J. u. T. Baker (Hrsg.) (2000): Research and Advanced Technology for Digital Libraries. 4th European Conference, ECDL 2000. Proceedings. Berlin, Heidelberg, S. 280–290

Hill, L. (2006): Georeferencing: The Geographic Associations of Information. Cambridge, Mass.
Hollenstein, L. u. R. Purves (2010): Exploring place through user-generated content: Using Flickr tags to describe city cores. In: Journal of Spatial Information Science 1 (2010), S. 21–48

Jaffe, A. et al. (2006): Generating Summaries and Visualization for Large Collections of Geo-Referenced Photographs. In: Wang, J. et al. (Hrsg.) (2006): International Multimedia Conference. Proceedings of the 8th ACM international workshop on Multimedia information retrieval. New York, S. 89–98

Kauppinen, T. (2010): Methods for Creating and Using Geospatio-temporal Semantic Web. In: TKK Dissertations 210. Espoo

Keßler, C. et al. (2009a): An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In: Agrawal, D. et al. (Hrsg.) (2009a): Geographic Information Systems. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Seattle, Washington, S. 91–100

Keßler, C. et al. (2009b): Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags. In: Janowicz, K. et al. (Hrsg.) (2009b): GeoSpatial Semantics. Third International Conference. Proceedings. Berlin, Heidelberg, S. 83–102

Kuchen, H. (2008): Informatik II: Datenstrukturen und Algorithmen. Kapitel 3: Dynamische Datenstrukturen (Listen, Baumstrukturen). Online unter: http://www-wi.uni-muenster.de/pi/lehre/ss08/info2/folien/info2k3.pdf (abgerufen am 16.08.2010)

Lange, N. de (2006): Geoinformatik in Theorie und Praxis (2. Auflage). Berlin

Lux, M. (2010): Flickr uploads per minute – mean uploads per hour. Online unter: http://www.semanticmetadata.net/2010/03/12/flickr-uploads-per-minute-mean-uploads-per-hour/ (abgerufen am 23.04.2010)

Metzler, D. u. H. Job (2007): Events und ihr Beitrag zur Regionalökonomie - die BUGA 05. In: Raumforschung und Raumordnung 65 (6), S. 514–530

Mukhopadhyaya, B. et al. (2004): Clustering of Earthquake Events in the Himalaya - Its Relevance to Regional Tectonic Set-up. In: Gondwana Research 7 (4), S. 1242–1247

Müller-Olm, M. (2008): Software Engineering. Kapitel 5: Entwurf. Online unter: http://cs.uni-muenster.de/sev/teaching/ws0809/se/SE0809-Kap5.pdf (abgerufen am 24.08.2010)

O'Reilly, T. (2005): What is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software. Online unter: http://oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=2 (abgerufen am 23.04.2010)

Ott, T. u. F. Swiaczny (2001): Time-Integrative Geographic Information Systems. Management and Analysis of Spatio-Temporal Data. Berlin

Pebesma, E. (2008): Introduction to Geostatistics. Looking forward: multivariate and geostatistics. Online unter: http://ifgi.uni-muenster.de/~epebe_01/Geostatistics/lec12.pdf (abgerufen am 04.08.2010)

Prinz, T. (2007): Digitale Fernerkundungsmethodik in den Geowissenschaften. Klassifikationen. Online unter: http://ivvgeo.uni-muenster.de/Vorlesung/FE_Script/3_7.html (abgerufen am 04.08.2010)

Rechenberg, P. u. G. Pomberger (2006): Informatik-Handbuch (4. Auflage). München

Sachs, L. u. J. Hedderich (2006): Angewandte Statistik. Methodensammlung mit R (12. Auflage). Berlin, Heidelberg

Schmid, C. et al. (2007): MADAP, a flexible clustering tool for the interpretation of one-dimensional genome annotation data. In: Nucleic Acids Research 35, S. 201–205

Sievers, J. (2001): Geographisches Namenbuch. In: Bollmann, J. u. W. Koch (Hrsg.) (2001): Lexikon der Kartographie und Geomatik. A bis Karti. Heidelberg, Berlin, S. 301–302

Smith, M. de et al. (2009): Geospatial Analysis - a comprehensive guide. Online unter: http://www.spatialanalysisonline.com/output/ (abgerufen am 04.08.2010)

Stadler, D. (2007): Tagging vs. Ontologies. In: Grob, H. u. G. Vossen (Hrsg.) (2007): Entwicklungen im Web 2.0 aus technischer, ökonomischer und sozialer Sicht. Internetökonomie und Hybridität 51. Münster, S. 3–12

Stein, A. (2007): Semantic Web vs. Web 2.0. In: Grob, H. u. G. Vossen (Hrsg.) (2007): Entwicklungen im Web 2.0 aus technischer, ökonomischer und sozialer Sicht. Internetökonomie und Hybridität 51. Münster, S. 13–21

Šumrada, R. (2003): Temporal Data and Temporal Reference Systems. In: Proceedings of the FIG Working Week 2003. Paris, TS10.3

Swan, R. u. J. Allan (2000): Automatic Generation of Overview Timelines. In: Yannakoudakis, E. et al. (Hrsg.) (2000): Annual ACM Conference on Research and Development in Information Retrieval. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. New York, S. 49–56

Will, H. (2001): Clusteranalyse. In: Bollmann, J. u. W. Koch (Hrsg.) (2001): Lexikon der Kartographie und Geomatik. A bis Karti. Heidelberg, Berlin, S. 122

Wilske, F. (2008): Approximation of Neighborhood Boundaries Using Collaborative Tagging Systems. In: Pebesma, E. et al. (Hrsg.) (2008): GI-Days 2008 - Interoperability and spatial processing in GI applications. Proceedings of the 6th Geographic Information Days. Münster, S. 179–187

Wolf, M. u. C. Wicksteed (1998): Status for Date and Time Formats. Online unter: http://www.w3.org/TR/NOTE-datetime (abgerufen am 10.08.2010)

Xie, L. et al. (2008): Event Mining in Multimedia Streams. In: Hanjalic, A. et al. (Hrsg.) (2008): Proceedings of the IEEE. Advances in Multimedia Retrieval. Los Alamitos, S. 623–647

Danksagung

Mein Dank gilt im Speziellen meinem Betreuer, Patrick Maué, für seine vielen Inspirationen zu dieser Arbeit.

Auch meiner Familie möchte ich an dieser Stelle danken, die mich nicht nur während dieser Arbeit, sondern seit Beginn meines Studiums auf vielen Wegen unterstützt.