Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme

Der Nutzen von Webkommentaren für das keyword-basierte Multimedia-Retrieval

Masterarbeit

Steffen Becker
Geboren am 27.11.1978 in Görlitz

Matrikelnummer 51448

1. Gutachter: Prof. Dr. Benno Stein
Betreuer: Martin Potthast

Datum der Abgabe: 27. Mai 2011

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weimar, den 27. Mai 2011

.....
Steffen Becker

Zusammenfassung

Das annotationsbasierte Multimedia-Retrieval ist in der Praxis die wichtigste Methode, um Multimedia-Dokumente inhaltsorientiert im Internet zu finden. Im Hinblick auf qualitative und quantitative Suchergebnisse sind Annotationen aber aufgrund des Mehraufwands häufig beim Erstellen zu kurz. Dies wird mit einem selbst erstellten Korpus mit einer großen Anzahl von Mediendokumenten sechs populärer Intenetportale nachgewiesen wird. Um zu überprüfen, ob Kommentare das Retrieval der Medienobjekte verbessern können, werden fünf verschiedene Experimente auf diesem Korpus ausgewertet. Es wird nachgewiesen, dass Kommentare ähnliche Retrieval-Eigenschaften wie Annotationen besitzen, durch ihre erheblich größere Textmenge die Quantität der Trefferdokumente sogar enorm verbessern können. Die Qualität der Suchergebnisse hingegen kann durch den Mangel einer Vergleichsbasis nicht eindeutig bewertet werden, da auch zwei populäre Suchmaschinen auf einer großen Anzahl von Suchanfragen vollkommen verschiedene Relevanzsortierungen liefern. Eine manuelle Bewertung einer Stichprobe von Suchtreffern zeigt, dass über Kommentare relevante Dokumente aufzufinden sind und sie somit einen wichtigen Beitrag zum Retrieval von Multimedia-Dokumenten leisten können.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Kommentare im Internet sind eine der ältesten und am weitesten verbreiteten Formen nutzergenerierter Inhalte, also Inhalte, die kostenfrei und nicht professionell von Internetnutzern publiziert werden. Sie sind auf vielen verschiedenen Internetseiten zu finden: Nachrichtenportale und Internetshops beispielsweise bitten Nutzer um Feedback zu ihren Artikeln; soziale Netzwerke, Blogs sowie Multimediaportale beruhen meist sogar vollständig auf solchen Inhalten. Es werden die unterschiedlichsten Objekte im Internet kommentiert. Dazu zählen Texte, Bilder, Videos, Musik, Produkte, aber auch persönliche Profile, Internetseiten und Links. Dadurch sind Kommentare für das Multimedia-Retrieval von besonderem Interesse. Sie könnten dabei helfen, Mediendokumente im Internet besser als bisher zu finden.

Internetsuchmaschinen sind in der Praxis auf Text angewiesen, um darin das Vorhandensein der Suchwörter einer Anfrage zu prüfen. Bei Dokumenten wie Bildern und Videos sind dafür Beschreibungen (Annotationen) in Textform notwendig. Auf den meisten Internetportalen, die auf die Präsentation von Medieninhalten spezialisiert sind, existieren solche vom jeweiligen Autor eines Mediendokuments verfassten Annotationen meist in Form von Titeln, Zusammenfassungen und Kurzbeschreibungen der Dokumente in Stichwörtern. Für das Auffinden eines Mediendokuments ist es besonders wichtig, dass dessen Annotationen treffend und ausführlich sind, da sie die wichtigste Ressource für das Retrieval darstellen. Jedoch ist dies nicht immer gegeben, da das Annotieren einen entscheidenden Mehraufwand für den Autor des Dokuments

bedeutet.

Das Kommentieren von Mediendokumenten auf Internetportalen ist sehr beliebt. Für populäre Dokumente sind tausende Kommentare keine Seltenheit. Nutzerkommentare sind, wie die anderen Annotationsformen ebenfalls in Textform vorhanden und prinzipiell für die Suche geeignet. Trotzdem werden Kommentare hierfür nicht genutzt. Sucht man beispielsweise bei `Google` Wörter, die in den Kommentaren zu einem bestimmten Mediendokument enthalten sind, aber nicht in Titel, Kurzbeschreibung oder Stichwörtern, so wird das Dokument nicht gefunden. Kommentare werden allgemein als zu unspezifisch betrachtet, da der inhaltliche Bezug eines Kommentars zu dessen Mediendokument meist sehr gering oder gar nicht vorhanden ist. Schaut man sich Kommentare beispielsweise bei `YouTube` an, kann man feststellen, dass meist mit sehr wenigen Worten größtenteils persönliche Meinungen ausgetauscht werden, zum Beispiel: „Nice clip!!!“, „hahahahahaha so crazy“, „ish ok but nt the best vid ive evr seen“. Verkürzte und oft falsche Schreibweisen sind ebenfalls problematisch.

In der vorliegenden Arbeit wird untersucht, ob Webkommentare dennoch einen Beitrag zum Keyword-Retrieval leisten können. Ein einzelner Kommentar kann dabei kaum als repräsentativ für ein Mediendokument betrachtet werden. Die Fragestellung ist vielmehr: Generiert die Gesamtmenge aller Kommentare zu einem Mediendokument eine inhaltliche Beschreibung dieses Dokuments, die für eine Suche geeignet ist? In verschiedenen Experimenten wird gezeigt, dass Kommentare für das Keyword-Retrieval grundsätzlich geeignet sind und dass sie die Suchergebnisse quantitativ und qualitativ verbessern können.

In Kapitel 2 werden zunächst Methoden des Information-Retrieval im Allgemeinen und das Keyword-basierte Multimedia-Retrieval im Speziellen erläutert. In Kapitel 3 werden sechs verschiedene Multimedia-Portale vorgestellt und das Kommentierverhalten auf diesen Portalen analysiert. Die Annotationen und Kommentare dieser sechs Portale werden anschließend in Kapitel 4 für verschiedene Retrieval-Experimente genutzt. Abschließend wird auf Basis der Experimentergebnisse in Kapitel 5.1 der Beitrag von Webkommentaren zum Keyword-Retrieval bewertet.

Kapitel 2

Information-Retrieval

In diesem Kapitel werden Aspekte und Methoden des Information-Retrieval vorgestellt, die für diese Forschungsarbeit von Bedeutung sind:

2.1 Allgemeines Retrieval-Modell Im ersten Abschnitt wird die grundlegende Problemstellung des Information-Retrieval aufgezeigt. Ein allgemeines Modell für Retrieval-Systeme wird eingeführt und Begriffe wie Relevanz und Indizierung erklärt.

2.2 Multimedia-Retrieval Hier wird der Begriff Multimedia im Information-Retrieval erklärt und die beiden Ansätze des inhaltsbasierten und annotationsbasierten Retrieval vorgestellt.

2.3 Text-Retrieval In diesem Abschnitt wird das Retrieval von Textdokumenten als traditionelle Domäne des Information-Retrieval erläutert. Verschiedene Modelle werden vorgestellt.

2.4 Retrieval mit Webkommentaren Webkommentare können als spezielle Informationsquelle für das Information-Retrieval genutzt werden. Verschiedene Arbeiten werden ausgewertet, die dies vorschlagen.

2.5 Evaluierung von Retrieval-Systemen Die grundlegenden Methoden zur Ermittlung der Qualität von Suchergebnissen und der Evaluierung von Retrieval-Systemen werden in diesem Abschnitt vorgestellt.

2.1 Allgemeines Retrieval-Modell

Sucht ein Internetnutzer ein Bild, so formuliert er sein Informationsbedürfnis in Form von Schlagwörten und stellt diese als Anfrage an eine Suchmaschine. Der Nutzer erwartet von den Bildern, dass diese seine Schlagwörter inhaltlich widerspiegeln und so sein Informationsbedürfnis befriedigt wird. Die Suchmaschine muss also in irgendeiner Weise den Inhalt verschiedener Bilder einer Bildersammlung mit der semantischen Bedeutung der gegebenen Schlagwörter vergleichen und dem Nutzer die relevanten Bilder liefern. Für diesen Vergleich müssen der Inhalt der Bilder, die als strukturierte Pixelanordnungen gespeichert sind und der Schlagwörter, die als Listen von Buchstaben vorliegen, in eine durch den Computer vergleichbare Form gebracht werden, um daraus zu bestimmen, welche Bilder zur Anfrage passen. Dieses Beispiel zeigt das Kernproblem des Information-Retrieval: die Transformation von Information aus verschiedenen Medien in adäquate Computerrepräsentationen. Ein konzeptionelles Modell für IR-Systeme kann wie folgt formuliert werden (siehe Abbildung 2.1):

Jedes Dokument d einer Dokumentsammlung D wird in eine Computerrepräsentation $\mathbf{d}$, die den Inhalt geeignet darstellt, überführt. Der Informationsbedarf wird als Anfrage q formuliert und ebenfalls in eine geeignete Computerrepräsentation $\mathbf{q}$ gewandelt. Aus $\mathbf{d}$ und $\mathbf{q}$ wird ein Wert ρ bestimmt, der die inhaltliche Relevanz des Dokuments zur Anfrage darstellt. Dieser Vergleich wird mit allen Dokumenten aus D durchgeführt. Eine nach ρ sortierte Liste der Dokumente wird als Ergebnis geliefert.

Die Relevanz bezeichnet die vom anfragenden Nutzer subjektiv empfundene Nützlichkeit, die er einem Dokument in Hinblick auf die Befriedigung seines konkreten Informationsbedürfnisses beimisst. Der vom IR-System berechnete Relevanzwert ist ein Schätzwert für diese Relevanz und wird zur Sortierung der Ergebnismenge genutzt. Diese Sortierung wird auch als Relevanz-Ranking bezeichnet. Die Dokumente mit der höchsten berechneten Relevanz werden dem Nutzer als erstes präsentiert, da sie das Informationsbedürfnis potentiell am besten befriedigen. Die Qualität der Suchergebnisse drückt sich somit einerseits darin aus, dass möglichst viele für den Nutzer relevante Dokumente

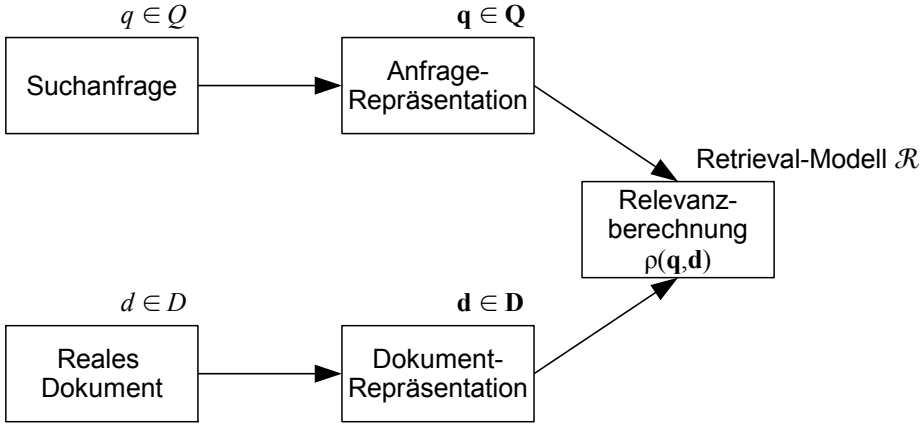


Abbildung 2.1: Konzeptionelles Model von IR-Systeme.

gefunden werden und andererseits, dass diese im Relevanz-Ranking richtig sortiert werden. Zur Bestimmung der Güte von Retrieval-Systemen müssen diese beiden Aspekte quantifiziert und gemessen werden. Das Vorgehen einer solchen Evaluierung mit entsprechenden Gütemaßen wird in Abschnitt 2.5 erläutert.

Unabhängig von der Art des Mediums der Dokumente lassen sich IR-Systeme hinsichtlich der Anfrageformulierung unterscheiden. Anfragen können beispielsweise in Form ausformulierter Fragen, als Schlagwörter („Query by Keywords“), als Beispiele („Query by Example“) oder auch als Anfragen mit spezifischen Strukturmerkmalen gestellt werden [Baeza-Yates et al., 1999]. Die bekannteste und am meisten verbreitete Form der Anfrageformulierung sind Schlagwörter. Die Suche wird dann auch als Keyword-Retrieval bezeichnet. Schlagwörter sind Wörter oder Phrasen, die das Informationsbedürfnis kurz und prägnant umschreiben. Zur genaueren Spezifizierung der Anfrage wird von einigen Retrieval-Systemen auch eine Verknüpfung der Schlagwörter mit booleschen Operatoren wie UND, ODER und NICHT unterstützt. Das Retrieval mit Schlagwörtern wird von den meisten Internetsuchmaschinen verwendet, da es sehr intuitiv und schnell zu formulieren sowie zu ändern ist. Erzielt eine

Suchanfrage nicht die erwarteten Ergebnisse, so kann sie schnell durch Ändern und Hinzufügen einzelner Schlagwörter präzisiert und neu gestellt werden.

Ein weiterer Aspekt von IR-Systemen ist die Organisation der Dokumentsammlung. Zusätzlich zu den digital gespeicherten Dokumenten besitzt ein Retrieval-System zur Beschleunigung der Suche einen Index. Im Index sind alle Daten der Computerrepräsentationen invertiert gespeichert, das heißt es werden den Dokumenten nicht die Ausprägungen ihrer Eigenschaften zugeordnet, sondern umgekehrt. Jeder existierenden Ausprägung einer Eigenschaft werden alle Dokumente der Sammlung zugeordnet, die diese Eigenschaft besitzen. Im Keyword-Retrieval sind dies beispielsweise die Schlagwörter der Dokumente. Der Index ist dort ein Schlagwortkatalog. Für eine konkrete Suchanfrage werden aus dem Index nur die Dokumente abgerufen, die auch die angefragten Schlagwörter besitzen. Nur diese Dokumente werden dann der Relevanzberechnung und dem anschließenden Ranking zugeführt. Alle anderen Dokumente, in denen keines der Schlagwörter vorkommt besitzen per Definition keine Relevanz für diese Anfrage. Dadurch ist es möglich, den Vergleich von Anfrage und allen Dokumenten der Kollektion auf wenige potentiell relevante Dokumente zu reduzieren und so eine effiziente Suche auf sehr großen Dokumentkollektionen zu ermöglichen.

Die Methoden zur Bildung der Computerrepräsentation unterscheiden sich stark für unterschiedliche Medienarten. Daher wird in die Teilgebiete Text-, Audio-, Bild-, Video- und Multimedia-Retrieval unterschieden. In den nächsten beiden Abschnitten werden das Multimedia-Retrieval und das Text-Retrieval vorgestellt, da diese Teilgebiete für das Retrieval mit Kommentaren von besonderer Bedeutung sind.

2.2 Multimedia-Retrieval

Der Begriff Multimedia wird allgemein verwendet, wenn mehrere Medien als Kommunikationsmittel betrachtet werden. Zu diesen Medien gehören Text, Bild, Bewegtbild, Audio und Video. Video beziehungsweise Film stellen in sich schon ein Multimedium dar, da darin Bewegtbild und Audio gemischt werden. Im Information-Retrieval wird der Begriff Multimedia unterschiedlich verwen-

det. Die inhaltliche Suche in gemischten Dokumentsammlungen von Text-, Bild-, Audio- und Videodokumenten wird beispielsweise von Baeza-Yates et al. [1999, S. 325] als Multimedia-Retrieval bezeichnet. Eine andere Eingrenzung trifft Stock [2006, S. 509-512]: Unterscheiden sich Anfragemedium und Medium der Dokumentkollektion, so wird dies als Multimedia-Retrieval bezeichnet. Hierzu zählt das Keyword-Retrieval auf Dokumenten, die keine Textform besitzen. Werden sich ähnelnde Dokumente gesucht, indem dem Retrieval-System ein Beispieldokument zum Vergleich gegeben wird, ist dies immer ein monomediales Retrieval, da sich hier das Medium von Anfrage und Dokumentsammlung per Definition nicht unterscheidet. Häufig werden in der Literatur aber auch die Teilgebiete Bildretrieval, Audio- und Musikretrieval sowie Videoretrieval als Multimedia-Retrieval eingeordnet, da meist ein Keyword-Retrieval auf Dokumenten im entsprechenden Medium erforscht wird. Allen Einordnungen gemein ist jedoch die Hauptaufgabe: die Erforschung von Methoden, die verschiedene Medientypen in für das Retrieval geeignete, den Inhalt abbildende Computerrepräsentationen wandeln können.

Um ein Keyword-Retrieval auf nicht-textuellen Medien zu ermöglichen, werden hauptsächlich zwei Ansätze verfolgt: das inhaltsbasierte (content-based) und das annotationsbasierte (annotation-based) Retrieval. Im inhaltsbasierten Retrieval wird die Computerrepräsentation $\mathbf{d}$ wie im formalen Modell direkt aus dem Dokument d erzeugt. Es wird beispielsweise versucht, den Inhalt eines Bildes anhand dessen Eigenschaften wie Farbe, Form und Textur automatisch mit Schlagwörtern zu beschreiben. Dabei gilt es die semantische Lücke (Semantic Gap) zwischen wenig komplexen, maschinell erfassbaren Eigenschaften und hoch komplexen inhaltlichen Interpretation zu überbrücken. Wichtige Forschungsschwerpunkte sind beispielsweise die Bilderkennung, um Dokumente in Bildern und Videos automatisch segmentieren zu können sowie maschinelles Lernen, das die automatische Klassifikation und die Zuordnung von Schlagwörtern für die segmentierten Dokumente ermöglicht. Da sich dieses Problem als äußerst schwierig erwiesen hat, existieren in der Praxis bislang keine robusten und leistungsfähigen Systeme, die auf sehr großen Datensammlungen wie dem Internet eingesetzt werden können.

Ein einfacherer und in der Praxis weit verbreiteter Ansatz für das Re-

trieval nicht-textueller Medien ist das annotationsbasierte Retrieval. Hierbei wird die Computerrepräsentation **d** nicht aus d selbst erzeugt, sondern aus den, dem Dokument beigefügten Annotationen a, die gleichsam als Stellvertreter für d dienen. Annotationen sind von Menschen erzeugte Inhaltsangaben in Textform. In der Praxis sind Annotationen meist Dokumenttitel, Kurzbeschreibungstexte und Schlagwörter. Verschiedene digitale Medienformate wie beispielsweise MPEG7 sehen Textfelder dafür vor, die Inhaltsbeschreibungen und Metadaten auf struktureller und semantischer Ebene aufnehmen können. Weiterhin wird im annotationsbasierten Retrieval versucht, Annotationen aus der Umgebung des publizierten Mediums, beispielsweise aus Internetseiten zu gewinnen. Hier setzt die vorliegende Arbeit an, indem sie Webkommentare auf Internetseiten als neuartige, bisher nicht genutzte Quelle für das annotationsbasierte Retrieval untersucht. Dabei werden keine Annotationen aus den Kommentaren extrahiert, sondern die Kommentare selbst als annotierender Text genutzt, um aus ihnen Computerrepräsentationen zu erzeugen. Abschnitt 2.4 erläutert das Retrieval mit Webkommentaren genauer und stellt verwandte Arbeiten vor.

Da im annotationsbasierten Ansatz jedes Dokument der Dokumentsammlung Annotationen in Textform besitzt, ist das Retrieval unabhängig vom Medientyp des jeweiligen Dokuments. Dokumentsammlungen können dadurch auch aus verschiedenen Medientypen bestehen, da für alle Dokumente auf Text zurückgegriffen wird. Das Retrieval gleicht damit dem Text-Retrieval beziehungsweise wird in dieses überführt.

2.3 Text-Retrieval

Text-Retrieval ist die am intensivsten erforschte Domäne des IR. Es umfasst die Wissenschaftsbereiche Informationswissenschaft, Informatik und Computerlinguistik. In der Vergangenheit wurde IR in sehr speziellen Anwendungsgebieten wie Recherchesystemen für Literatur und Patentdokumentationen eingesetzt [Stock, 2006, S. 43-46]. Mit dem Aufkommen und der großen Verbreitung des Internets wuchs die Bedeutung des Information-Retrieval. Das Internet stellt eine gigantische, kaum strukturierte Sammlung von Dokumenten, vorwiegend

in Textform, dar. Internetsuchmaschinen nutzen überwiegend Methoden des Text-Retrievals und sind für die Nutzung des Internets von zentraler Bedeutung. Erst durch sie wird der breite Zugang zu diesen Dokumenten und dem darin enthaltenen Wissen möglich.

Im Text-Retrieval bestehen die Dokumente der Dokumentkollektion sowie Nutzeranfragen aus natürlichsprachigem Text. Für die Bildung der Computerrepräsentationen **d** und **q** wird der so genannte Bag-of-Words Ansatz genutzt. Texte werden dabei als unsortierte Menge von Indextermen, unabhängig von Grammatik und Reihenfolge, dargestellt. Ein Indexterm ist ein Schlagwort, dessen Semantik den spezifischen Inhalt eines Dokuments zu beschreiben hilft. Indexterme sind häufig Substantive, da diese meist eine Bedeutung tragen und weniger stark vom Kontext des Satzes abhängig sind als andere Wortarten. Mit solchen Schlagwörtern ist es möglich Dokumentkollektionen zu indizieren und auch Dokumente kurz zusammenzufassen. Im Folgenden werden verschiedene wichtige Aspekte und Modelle des Text-Retrieval erläutert.

Dekomposition

Die Dokumente sind digital als Liste von Einzelbuchstaben gespeichert. Um daraus Indexterme zu gewinnen, werden die Texte tokenisiert. Dafür wird die Liste anhand von Leer- und Satzzeichen sowie verschiedener Heuristiken in kleinere Listen zerlegt, die Einzelwörtern entsprechen. Außerdem werden meist Stoppwörter entfernt. Stoppwörter sind sehr oft verwendete Wörter wie *und*, *oder* und *nicht*, die selbst wenig semantische Bedeutung besitzen und eher eine syntaktische Funktion haben. Indexterme müssen unabhängig von ihrer Zeitform und Beugung vergleichbar sein. Dazu wird nur der Wortstamm jedes Wortes betrachtet. Eine sehr häufig verwendete Methode zur Wortstammreduktion ist der Porter-Stemmer-Algorithmus [Porter, 1980]. Dabei wird durch Anwendung verschiedener Verkürzungsregeln ein Wort zu einer Minimallänge von Vokal-Konsonant-Sequenzen reduziert. Die so erzeugten Wortstämme entsprechen dabei nicht linguistischen Wortstämmen, genügen aber für die weitere Verarbeitung, da verwandte Wörter auf den selben Stamm reduziert werden.

Boolesches Retrieval-Modell

Das einfachste Modell im Text-Retrieval ist das Boolesche Retrieval-Modell.

Es nutzt ausschließlich Indexterme und basiert auf der Mengenlehre sowie der booleschen Algebra. Eine Anfrage wird darin aus der Verknüpfung von Schlagwörtern durch die logischen Operatoren UND, ODER und NICHT formuliert. Die Computerrepräsentation der Anfrage besteht aus diesem logischen Ausdruck. Die Repräsentation des Dokuments ist die unsortierte Menge der Indexterme. In der Relevanzberechnung eines Dokuments zu einer Anfrage wird der logische Ausdruck auf der Menge der Indexterme geprüft. Der Relevanzwert ist binär, das Dokument ist in Bezug auf die Anfrage entweder relevant, wenn der logische Ausdruck zutrifft, oder nicht relevant, wenn der Ausdruck nicht zutrifft. Partielle Treffer sowie eine Ordnung der Ergebnismenge sind in diesem Modell nicht möglich.

$tf{\cdot}idf$-Gewichtung

Um eine Ordnung zu ermöglichen, werden den Indextermen in anderen Modellen numerische Gewichte zugewiesen. Der Wert wird aus der Wichtigkeit des Terms für das entsprechende Dokument bestimmt. Terme, die in dem Dokument häufig vorkommen, in der gesamten Dokumentkollektion aber nur selten, bekommen das höchste Gewicht, da sie für dieses Dokument von besonderer Bedeutung sind. Die am häufigsten verwendete Gewichtungsfunktion ist die $tf{\cdot}idf$-Gewichtung:

$$tf{\cdot}idf_{i,j} = tf_{i,j} \cdot idf_i \quad \text{mit}$$
$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad \text{und} \quad idf_i = \log \frac{|D|}{n_i}$$

Das $tf{\cdot}idf$-Gewicht des Terms t_i aus dem Dokument d_j wird aus der normalisierten Termhäufigkeit $tf_{i,j}$ und der inversen Dokumentfrequenz idf_i bestimmt. Die Termhäufigkeit $tf_{i,j}$ ist der Quotient der Auftrittshäufigkeit des Terms in d_j und der Gesamtzahl aller Terme in d_j. Die inverse Dokumentfrequenz idf_i ergibt sich aus dem Logarithmus der Anzahl der Dokumente in D und Anzahl n_i der Dokumente in denen t_i vorkommt. Sie spiegelt die Wichtigkeit des Terms in der gesamten Kollektion wider. Je häufiger ein Term vorkommt, desto weniger diskriminativ ist er.

Vektorraummodell

Das von Salton and McGill [1983] vorgestellte Vektorraummodell (VSM) ist ein sehr gebräuchliches Retrieval-Modell und nutzt gewichtete Termvektoren. Die Computerrepräsentationen $\mathbf{d}$ und $\mathbf{q}$ sind hier die Vektoren der Termgewichte. Eine Dimension der Vektoren entspricht genau einem Wort. Als Relevanzmaß wird die Ähnlichkeit dieser hochdimensionalen Vektoren berechnet. Hierfür eignet sich der Kosinus des Winkels der beiden Vektoren besonders gut:

$$\rho(\mathbf{q}, \mathbf{d}) = \cos \measuredangle(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\| \|\mathbf{d}\|}$$

Der Kosinus des Winkels zwischen $\mathbf{d}$ und $\mathbf{q}$ wird aus dem Skalarprodukt der beiden Vektoren und deren Länge berechnet. Der Kosinus bildet nichtlinear auf den Bereich zwischen 0 und 1 ab, wobei 0 bei vollständiger Ungleichheit und 1 bei exakter Gleichheit der Vektoren erreicht wird. Aus der Ähnlichkeit der Termvektoren wird auf die inhaltliche Textähnlichkeit geschlossen. Diese Textähnlichkeit wiederum wird als Relevanzmaß genutzt. Demzufolge hat ein Dokument eine Relevanz von 0, wenn keines der Anfragewörter enthalten ist und eine Relevanz von 1, wenn alle Wörter der Anfrage mit der gleichen Wichtigkeit im Dokument vorkommen. Das Vektorraummodell eignet sich besonders gut, wenn Anfrage und Dokument eine ähnliche Länge besitzen, wie im Retrieval mit Anfragen, die als Beispiel formuliert sind. Beim Keyword-Retrieval hingegen sind Anfragen sehr kurz. Das Vektorraummodell ist aber auch hier einsetzbar. Die Schlagwörter werden dafür zusätzlich gewichtet, beispielsweise über den Rocchio-Algorithmus.

Weitere Modelle

Das Vektorraummodell gehört zur Klasse der algebraischen Retrieval-Modelle. Ein weiteres Modell dieser Klasse ist beispielsweise Latent-Semantic-Indexing (LSI). Dabei wird nicht wie beim Vektorraummodell von einer statistischen Unabhängigkeit der Indexterme ausgegangen. LSI kann dadurch auch mit Synonymen und Homonymen umgehen. Eine andere Klasse sind die so genannten probabilistischen Modelle. In diesen Modellen wird die Wahrscheinlichkeit bestimmt, mit der ein Dokument für eine Anfrage relevant ist. Diese Wahrscheinlichkeit wird über die Auftrittshäufigkeit beziehungsweise die Termhäufigkeit

der Indexterme und Wahrscheinlichkeitstheoreme, wie das Bayestheorem berechnet. Der wichtigste Vertreter dieser Klasse ist Okapi BM25, das oft in Websuchmaschinen zum Einsatz kommt.

In der Klasse der eigenschaftsbasierten Modelle werden Dokumente nicht mehr nur als Menge von Indextermen betrachtet, sondern als Menge abstrakter Eigenschaften (Features) dargestellt. Mit Methoden des maschinellen Lernens werden Ranking-Funktionen trainiert, die anhand der Eigenschaften Dokumente nach ihrer Relevanz zur Anfrage sortieren. Als Features werden Termgewichtsvektoren, aber auch Struktur- und Metaidaten der Dokumente genutzt. Diese Modelle sind relativ neu im Information-Retrieval und wurden speziell für die Internetsuche entwickelt. Häufig werden sie im Retrieval-Prozess als zweite Stufe genutzt. In einer ersten Stufe werden potentiell relevante Dokumente mit den traditionellen Modellen bestimmt. Die Dokumente mit dem höchsten Relevanzwert werden in einer zweiten Stufe mit diesen neuen Methoden noch einmal sortiert. Dabei fließen Informationen wie Click-Statistiken und Linkstrukturen als Features in die Sortierung ein. Der von Brin and Page [1998] entwickelte Page-Rank-Algorithmus berechnet Gewichte für jede Internetseite auf Basis der Hyperlinks von und zu anderen Seiten. Page-Rank ist ein wichtiges Maß für die Popularität einer Internetseite, das auch die Relevanz bei der Suche beeinflusst. Die Suchmaschine von Google verwendet unter anderem Page-Rank für die Sortierung der Suchtrefferseiten. Die exakte Funktionsweise kommerzieller Suchmaschinen wie Google ist nicht bekannt. Es existieren aber verschiedene frei verfügbare Suchmaschinen. Für die Experimente im Rahmen dieser Arbeit wurde beispielsweise die Suchmaschine Lucene[1] der Apache Software Foundation genutzt. Lucene ist in Java implementiert und verwendet das Vektorraummodell zum Keyword-Retrieval. Durch einen leistungsfähigen und frei anpassbaren Indexer ist die effiziente Suche in einer sehr großen Dokumentsammlung somit möglich.

[1] `http://lucene.apache.org`

2.4 Retrieval mit Webkommentaren

Kommentare stellen eine neuartige Informationsquelle für das annotationsbasierte Multimedia-Retrieval, aber auch für das Text-Retrieval im Internet dar. Ein Kommentar bezieht sich nämlich immer eindeutig auf ein bestimmtes Dokument beziehungsweise auf einen vorher abgegebenen Kommentar. Damit kann der Kommentartext C aller Kommentare c, die zu einem Dokument d abgegeben wurden, für das Keyword-Retrieval des Dokuments genutzt werden. Die Art des Mediums von d spielt keine Rolle, wenn zur Bildung der Computerrepräsentation $\mathbf{d}$ auf C zurückgegriffen wird.

Webkommentare können den Mehraufwand und die daraus resultierende, oft mangelnde Sorgfalt und Unvollständigkeit beim Annotieren von Multimedia-Dokumenten mindern. Die Güte der Retrieval-Ergebnisse im annotationsbasierten Multimedia-Retrieval hängt maßgeblich von der Qualität und Quantität der Annotationen ab. Je genauer und ausführlicher annotiert wird, desto besser können auch die Suchergebnisse sein. Außerdem kann die Subjektivität eines einzelnen Annotationsautors reduziert werden, da Kommentare von vielen verschiedenen Autoren verfasst werden und so auch verschiedene Sichtweisen, Aspekte und Zusatzinformationen für das Dokument beinhalten. Dagegen steht allerdings die Erfahrung, dass Kommentare oft sehr kurz sind und nur die persönliche Meinung des Kommentatierenden ausdrücken. Reine Meinungsäußerungen haben für die Suche nach einem Mediendokument aber meist keine Relevanz. Wird die Meinung allerdings mit objektiven Begründungen belegt, können darin wiederum wichtige Informationen für das entsprechende Dokument enthalten sein.

Verschiedene Forschungsarbeiten zeigen, dass Kommentare für das Information-Retrieval nützlich sein können. In einer früheren Arbeit [Potthast et al., 2011] untersuchten wir Aufgaben des Information-Retrieval innerhalb der Kommentarsphäre, das heißt innerhalb der Kommentarmenge eines Dokuments. Wir identifizierten darin drei Hauptaufgaben: Kommentarfilterung nach Qualität abgegebener Kommentare, Kommentarsortierung nach inhaltlicher Relevanz zum kommentierten Dokument beziehungsweise nach Nützlichkeit für einen Leser und die Zusammenfassung aller Kommentare für ein Dokument.

Innerhalb dieser drei Schwerpunkte wurden verschiedene Forschungsarbeiten, Ansätze und eigene Beiträge eingeordnet. Die Filterung nach relevanten, lesenswerten Kommentaren wird demnach oft zur Erkennung und Vermeidung von Missbrauch in Form von Spam eingesetzt. Die Zusammenfassung von Kommentaren mit ähnlichem Inhalt spielt eine wichtige Rolle bei der Erfassung von Meinungen, Stimmungen und Trends in der Nutzerschaft und wird oft in Online-Shops genutzt. Auch die Kommentarsortierung nach verschiedenen Relevanzkriterien wird dort verwendet, um die besten Beiträge besonders hervorzuheben.

Im Gegensatz dazu existieren nur wenige Arbeiten, die sich mit Webkommentaren als zusätzliche Quelle für das Retrieval der kommentierten Dokumente beschäftigen. Nur zwei Forschungsarbeiten analysieren den quantitativen Einfluss von Kommentaren auf das Keyword-Retrieval, zwei weitere Arbeiten beschäftigen sich mit einer qualitativen Verbesserung des Retrievals. Mishne and Glance [2006] betrachten Kommentare in Blogs. Die Autoren zeigen darin, dass die Ergebnismenge bei einem Keyword-Retrieval um 15% gesteigert werden kann, wenn Kommentartexte zusätzlich zum Blogtext indiziert werden. Die Ergebnisse stützen sich dabei allerdings auf nur 40 Retrieval-Anfragen. Die Qualität beziehungsweise Relevanz der Retrieval-Ergebnisse wird nicht genauer analysiert. Aus der Tatsache, dass mithilfe von Kommentaren mehr Dokumente gefunden werden, kann nicht geschlossen werden, dass diese zusätzlichen Dokumente auch relevant für die jeweilige Anfrage sind. In [Yee et al., 2009] werden Kommentare auf dem Videoportal YouTube untersucht und mit Titel, Kurzbeschreibung und Schlagwörtern als Informationsquelle für das Keyword-Retrieval verglichen. Retrieval-Anfragen werden dafür zufällig aus den Schlagwörtern des Korpus erzeugt. Die Autoren weisen aber darauf hin, dass diese synthetischen Anfragen nicht mit einem realen Anfrageverhalten übereinstimmen. Die Autoren verbessern die Retrieval-Genauigkeit in ihren Experimenten um bis zu 15% unter Verwendung von Kommentaren. Der Grad der Verbesserung ist dabei relativ anzusehen, da auch hier keine Aussagen über die tatsächliche Relevanz der Ergebnisse zu den Anfragen gemacht werden.

Potthast [2009] untersucht die Fähigkeit von Kommentaren das kommentierte Dokument zu beschreiben. Dazu werden Texte und Kommentare des

Nachrichtenportals Slashdot untersucht. Unter Nutzung verschiedener Retrieval-Modelle wie dem Vektorraummodell werden Kommentare und Nachrichten miteinander verglichen. Der Autor zeigt, dass schon 10 Kommentare eine deutliche Ähnlichkeit zum kommentierten Text aufweisen und 100 bis 500 Kommentare den Nachrichtentext im Ranking ersetzen können, ohne dass die Ähnlichkeit aus der Duplizierung der Nachrichten in den Kommentaren entsteht. Der Autor zeigt somit, dass Kommentare deskriptiv für Dokumente in Textform sind und schlägt vor, dieses Ergebnis auch auf Dokumente in anderen Formen, wie Bilder und Videos zu übertragen.

Weiterführend zeigt der Autor Potthast et al. [2010], dass über Webkommentare ein Cross-Media-Retrieval möglich ist: Wenn sich die Kommentartexte zweier Mediendokumente unterschiedlicher Art inhaltlich stark ähneln, so besitzen auch die dazugehörigen Dokumente eine hohe Ähnlichkeit. Dazu wurde die Kosinus-Ähnlichkeit im Vektorraummodell für Paare von Slashdot- und YouTube-Kommentartexten berechnet. Eine manuelle Analyse der Mediendokumente von Paaren mit hoher Kommentartextähnlichkeit zeigt, dass sich diese Mediendokumente ebenfalls in hohem Maße ähneln. Kommentare auf Mediendokumenten, die nicht textueller Natur sind, eignen sich demnach ebenfalls, um den semantischen Inhalt des Dokuments zu beschreiben.

Die vorliegende Arbeit schließt an diese Arbeiten an und untersucht beide Aspekte, den quantitativen und qualitativen Beitrag von Kommentaren für das annotationsbasierte Keyword-Retrieval von Multimedia-Dokumenten. Dazu werden Kommentare auf Mediendokumenten in Form von Text, Bild, Ton und Video verschiedener Internetportale untersucht.

2.5 Evaluierung von Retrieval-Systemen

Um die Qualität eines IR-Systems zu messen bedarf es eines Evaluierungskorpus. Ein solches Korpus besteht aus einer Dokumentsammlung und einer Menge von Anfragen, wobei alle relevanten Dokumente in der richtigen Sortierung bekannt sind beziehungsweise von menschlichen Experten zugewiesen wurden. Die Relevanz wird häufig binär, als relevant und nicht relevant oder in wenigen Abstufungen davon angegeben. Über verschiedene Evaluierungsma-

ße kann dann ein Retrieval-System getestet werden. Die Trefferquote (Recall) und die Genauigkeit (Precision) sind die wichtigsten Maße für die Qualität der Ergebnismenge unabhängig von deren Reihenfolge:

$$Recall = \frac{|rel \bigcap f|}{|rel|} \quad \text{und} \quad Precision = \frac{|rel \bigcap f|}{|f|}$$

Dabei wird die Menge aller relevanten Dokumente rel der Dokumentkollektion zu einer Anfrage mit der Menge der insgesamt gefundenen Dokumente f verrechnet. Precision und Recall werden für eine Anzahl von Fragen gemessen und beispielsweise arithmetisch gemittelt. Sie sind dabei nicht unabhängig voneinander. Ein System, das für jede Suchanfrage alle Dokumente der Kollektion zurückliefert, hätte einen idealen Recall, aber eine sehr schlechte Precision. Aus diesem Grund wird im sogenannten F-Maß (F-Measure) das harmonische Mittel aus beiden Werten gebildet.

Die Qualität des Rankings in Bezug zur idealen Relevanzsortierung kann mit verschiedenen Korrelationsmaßen, wie beispielsweise Kendalls Tau (τ), bestimmt werden:

$$\tau = \frac{n_c - n_d}{\sqrt{n_c + n_d + extra_1} \cdot \sqrt{n_c + n_d + extra_2}}$$

Kendalls Tau bestimmt den Grad der Unabhängigkeit zweier Variablen. Die Variablen sind dabei die Ränge der Dokumente in zwei Sortierungen. Für jedes Dokument, das in beiden Rangfolgen vorkommt, wird ein Wertepaar aus den beiden Rängen ermittelt. Für alle möglichen Paarungen dieser Wertepaare wird ermittelt, wie viele in ihrer Anordnung der relativen Größe übereinstimmen (n_c), wie viele andersherum angeordnet sind (n_d) sowie die Anzahl der Paare mit gleichem Wert im Rang aus der ersten Liste ($extra_1$) und aus der zweiten Liste ($extra_2$). Paare, die sich in beiden Werten gleichen, werden nicht betrachtet.

Ein weiteres Gütemaß für Ergebnislisten ist der Normalized-Discounted-

Cumulative-Gain ($NDCG$):

$$NDCG = \frac{DCG}{IDCG} \quad \text{mit} \quad DCG = rel_1 + \sum_{i=2}^{N} \frac{rel_i}{log_2(i)}$$

Im Discounted-Cumulative-Gain (DCG) wird der Nutzen jedes Rangs der Ergebnisliste summiert. Dabei werden zwei Aspekte berücksichtigt: Der Nutzen eines Dokuments ist höher, je höher seine Relevanz in Bezug auf die Anfrage ist. Ein Dokument ist von höherem Nutzen, wenn es einen höheren Rang hat als Dokumente mit geringerer Relevanz. Dazu wird der Relevanzwert jedes Dokuments mit dem Logarithmus seines Rangs reduziert und über alle Dokumente summiert. Für den $NDCG$ wird über den DCG einer idealen Relevanzsortierung $IDCG$ normiert. Durch den $NDCG$ kann der Nutzen von Ergebnislisten eines Retrieval-Systems bestimmt und mit Ergebnislisten mit jeweils gleicher Anfrage eines anderen Systems verglichen werden. In der Praxis ist es allerdings kaum möglich einen exakten numerischen Wert für die Relevanz eines Dokuments zu bestimmen. Daher wird manuell häufig nur bestimmt, ob ein Dokument relevant, verwandt oder irrelevant ist und entsprechend die Werte 2, 1 oder 0 zugewiesen.

Für verschiedene Aufgabenstellungen im Information-Retrieval existieren entsprechende Relevanzkorpora. Vor allem die Text-Retrieval-Konferenz TREC bietet im Rahmen von wissenschaftlichen Workshops solche Korpora zum Entwickeln und Vergleichen von Retrieval-Systemen an. Beispiele dafür sind Web Track, Blog Track, Video Track und TeraByte Track.[2] In keinem dieser Korpora sind Webkommentare enthalten. Die Evaluierung und der Vergleich von Retrieval-Systemen, die Kommentare als Informationsquelle nutzen, zu herkömmlichen Systemen, ist somit mit den vorgestellten Mitteln nicht möglich. Precision und Recall können nicht bestimmt werden, wenn keine Kollektion mit bekannten relevanten Dokumenten und Anfragen vorhanden ist. Die Erstellung eines solchen Korpus ist sehr aufwendig, da die Relevanz von gefundenen Dokumenten zur Anfrage zuverlässig nur manuell bestimmt werden kann. Daher werden in den Experimenten die Eigenschaften und das Verhalten von

[2] `http://trec.nist.gov/tracks.html`

Webkommentaren und herkömmlichen Annotationen verglichen und die Ergebnislisten von Retrieval-Systemen, die diese jeweils als Quelle benutzen, verglichen. Eine ideale Ergebnisliste existiert dabei nicht. Im nächsten Kapitel wird ein Korpus aus Multimedia-Dokumenten und deren Annotationen und Kommentaren aus sechs verschiedenen Internetportalen vorgestellt.

Kapitel 3

Ein Kommentarkorpus zur Analyse

Um Webkommentare zu analysieren und mit Annotationen im Keyword-Retrieval zu vergleichen, entstand im Rahmen der Arbeit eine Dokumentsammlung (Korpus) verschiedener Portale, die Medieninhalte präsentieren und Kommentare besitzen. Der erstellte Korpus wird in diesem Kapitel präsentiert:

3.1 Ausgewählte Multimediaportale Kommentare und Annotationen von sechs verschiedenen Multimedia-Portalen wurden für den Korpus zusammengetragen. Die Auswahlkriterien werden in diesem Abschnitt erläutert und die sechs Portale kurz vorgestellt.

3.2 Erstellung des Korpus Das Vorgehen beim Download von Mediendokumenten und Kommentaren der vorgestellten Portale wird erläutert und die Struktur der indizierten Daten beschrieben.

3.3 Vergleich von Annotier- und Kommentierverhalten In diesem Abschnitt werden Unterschiede und Gemeinsamkeiten beim Kommentieren und Annotieren auf den einzelnen Portalen analysiert und zwischen den Portalen verglichen.

3.1 Ausgewählte Multimediaportale

Um eine umfassende Aussage über den Beitrag von Webkommentaren zum Multimedia-Retrieval treffen zu können wurden Kommentare zu allen Medientypen, die im Internet genutzt werden, analysiert. Für die Medientypen Musik, Kurzvideo, Film, Bild und Text wurde daher je ein Internetportal ausgewählt. Um möglichst viele Mediendokumente und Kommentare untersuchen zu können, wurden jeweils sehr populäre Portale ausgewählt. Ein weiteres Kriterium für die Auswahl stellte die technische Zugänglichkeit dar, denn viele Portale schützen sich vor einem automatischen Download durch die Begrenzung der Webseitenaufrufe pro Zeit. Weitere Aspekte bei der Ermittlung und dem Download der Webseiten werden in Abschnitt 3.2 erläutert. Annotationen, Kommentare und andere Daten der Mediendokumente der folgenden sechs Portale wurden im Korpus zusammengetragen:

Last.fm `http://www.last.fm`
Last.fm ist ein Internetradio und ein soziales Netzwerk, auf dem die Nutzer ihre Musikvorlieben mit anderen Nutzern teilen können. Es präsentiert seinen Nutzern Musik auf der Basis ihrer Hörgewohnheiten und der anderer Nutzer mit ähnlichem Musikgeschmack. Das Portal wurde 2002 gegründet und gehört seit 2007 zum US-amerikanischen Medienkonzern CBS Corporation.

YouTube `http://www.youtube.com`
YouTube ist ein Internet-Videoportal, auf dem Nutzer kostenfrei Videos präsentieren und ansehen können. Es wurde 2005 gegründet und im Jahr darauf von Google übernommen. Die Videosammlung besteht sowohl aus nutzergenerierten Filmen als auch aus kurzen Film- und Fernsehausschnitten sowie Musikvideos. Genaue Angaben über die Zahl der angebotenen Videos gibt es nicht, aber YouTube gilt als Marktführer im Bereich der Videoportale. Laut eigenen Angaben vom Mai 2010[1] hat YouTube über 2 Milliarden Seitenaufrufe pro Tag.

[1] `http://youtube-global.blogspot.com/2010/05/at-five-years-two-billion-views-per-day.html`, Stand: 20. 02. 2011

Internet Movie Database - IMDb `http://www.imdb.com`
IMDb entstand 1990 aus der Newsgroup rec.arts.movies, einer Diskussionsplattform filmbegeisterter Internetnutzer. Das Portal gehört seit 1998 dem Internet-Versandhaus Amazon. Die Nutzung und Mitwirkung ist größtenteils kostenlos; Filmeinträge und Filmkommentare können von allen registrierten Nutzern erstellt werden. IMDb listet über eine Million Filmeinträge, stellt die jeweiligen Filme aber nicht zum Ansehen zur Verfügung. Es werden lediglich Informationen und Kommentare angeboten.

Picasa `http://www.picasa.com`
Picasa ist eine Bildarchivierungs- und Verwaltungssoftware, die von Google kostenlos angeboten wird. Mit dem Dienst Picasa-Webalben können Fotos im Internet publiziert werden. Die Fotos können mit umfangreichen Metadaten wie Kameraeinstellungen und Ortsangaben versehen und in persönlichen Fotoalben organisiert werden. Andere Nutzer haben die Möglichkeit diese Fotos zu kommentieren.

The Huffington Post - HuffPost `http://www.huffingtonpost.com`
HuffPost ist eine Internet-Nachrichtenplattform, die als Blog kommentierte Internetlinks zu verschiedenen Nachrichtenquellen präsentiert. In die Blogtexte können außerdem Bilder und Videos eingebunden werden. HuffPost wurde 2005 gegründet und im Februar 2011 an den US-amerikanischen Online-Dienst AOL verkauft.

Blogger `http://www.blogger.com`
Blogger ist ein Hosting-Anbieter für kostenlose Blogs. Blogger wurde 1999 gegründet und gehört seit 2003 zu Google. Den Blogautoren wird ein einfaches Framework zur Verfügung gestellt, um Texte auch ohne HTML-Kenntnisse zu verfassen. Einen zentralen Index für alle existierenden Blogs gibt es nicht. Die populärsten Blogs werden jedoch pro Monat von Bloggern gewählt und zentral präsentiert.

3.2 Erstellung des Korpus

Für jedes Portal wurden alle Informationen zu den Mediendokumenten mit den dazugehörigen Kommentaren heruntergeladen. Die Mediendokumente selbst wurden nicht gespeichert, da der Ressourcenaufwand gerade bei Video- und Musikdateien sehr groß ist und eine direkte Analyse des Inhalts nicht relevant für das Forschungsvorhaben ist. Da keines der Portale, ausgenommen IMDb, ein vollständiges, zugängliches Register bereitstellte, wurden die Portale gecrawlt, das heißt automatisch durchsucht und Webseiten mit Mediendokumenten heruntergeladen. Für das Herunterladen der Daten bei IMDb, HuffPost und Last.fm wurde das freie Web-Crawling-Framework Scrapy[2] in der Version 0.10.3 verwendet. Ausgehend von einer Startseite wurden dabei alle Links verfolgt, ähnlich der Traversierung eines gerichteten Graphen. Alle Seiten, die Mediendokumente und Kommentare enthalten wurden gespeichert. Bei den drei Portalen, die zu Google gehören (YouTube, Picasa und Blogger) wurde die von Google selbst angebotene Programmierschnittstelle[3] (API) genutzt. Über die Programmierschnittstelle wurden generierte Suchanfragen gestellt und die Ergebnisseiten mit Kommentaren gespeichert. Die generierten Suchanfragen bestanden aus Einzelbuchstaben und Zahlen sowie einer Auswahl der Top-Suchanfragen von Google[4].

Beide Verfahren unterlagen allerdings Beschränkungen: Beim Crawlen sind nur die Webseiten erreichbar, die Ziel einer Kette von Links ausgehend von der Startseite sind. Nicht bei allen Webseiten ist das aber der Fall. Die Nutzung der Suchfunktion des API wurde zusätzlich durch Google selbst beschränkt. Pro Suche waren maximal 1000 Ergebnisseiten und für diese je maximal 1000 Kommentare abrufbar. Durch diese Beschränkungen war es nicht möglich einen vollständigen Korpus aller Mediendokumente und Kommentare eines Portals zu erstellen. Es ist bei keinem der Portale genau bekannt, wie viele Mediendokumente insgesamt vorhanden sind, wodurch keine Aussage über die Größe des Anteils der gecrawlten Mediendokumente an der Gesamtmenge getroffen werden kann. Die Korpora stellen somit eine zufällige Teilmenge aller Dokumente

[2] `http://www.scrapy.org`
[3] `http://code.google.com/intl/de-DE/apis/gdata/docs/directory.html`
[4] `http://www.google.com/insights/search`

der Portale für den Zeitraum August bis Oktober 2010 dar.

Um den Speicheraufwand einzugrenzen wurden nur Mediendokumente beachtet, die mindestens einen Kommentar besitzen. Aus den heruntergeladenen Internetseiten der Portale wurden alle Metadaten und Annotationen sowie alle Kommentare der Mediendokumente extrahiert und im XML-Format gespeichert. Tabelle 3.1 gibt eine Übersicht über alle Annotationen, Metadaten und Daten, die beim Download erhoben wurden. Für alle sechs Portale existieren die Annotationen Titel, Zusammenfassung und Schlagwörter. Für die Portale Blogger und HuffPost entspricht die Zusammenfassung dem eigentlichen Mediendokument, dem Text. Allerdings können darin auch Links, Bilder und Videos enthalten sein, die nicht heruntergeladen wurden. Bei IMDb existieren zwei Kategorien für Schlagwörter. Beide wurden separat gespeichert, für die weiteren Analysen aber zusammengefasst. Die Metadaten der Portale sind im Gegensatz zu den Annotationen nicht einheitlich und teilweise auch medienspezifisch, wie beispielsweise die Länge eines Videos. In den weiteren Analysen und Experimenten werden sie außer Acht gelassen. Tabelle 3.2 listet alle vorhandenen Daten auf, die für die Kommentare gespeichert wurden. Ein Kommentartext existiert auf allen Portale, ein separater Kommentartitel hingegen nicht. Der Titel wurde daher für die weiteren Analysen nicht verwendet. Metadaten der Kommentare existieren nur für drei der Portale und werden im Folgenden ebenfalls nicht betrachtet.

Die Dokumente sowie die Kommentare sind über Identifikationsnummern (ID) eindeutig identifizierbar. Die Kommentare besitzen zusätzlich zu ihrer eigenen ID die ID des Mediendokuments. Dadurch ist eine Referenzierung jederzeit möglich. Für die weiteren Analysen und Experimente wurden Dokumente und Kommentare für jedes Portal in jeweils einem Lucene-Index über ihre IDs indiziert. Für die Kommentare wurde die Referenz-ID zusätzlich indiziert. Dadurch ist es möglich alle Kommentare für ein bestimmtes Dokument effizient anzufragen.

	Last.fm	**YouTube**	**IMDb**	**Picasa**	**HuffPost**	**Blogger**
Annotationen						
Titel	+	+	+	+	+	+
Zusammenfassung	+	+	+	+	+	+
Schlagwörter	+	+	Plot Genre	+	+	+
Metadaten						
Autor		+		+	+	+
Bewertungen		Min/Max/Mittel Stimmen Favoriten	Mittel Stimmen			
Verwandte Objekte	+	+			+	
Dokumentaufrufe	Anzahl Hörer Anzahl gespielt	+				
Standort	+	+				
Sonstiges		Videolänge Credits	Jahr	Bild-URL		Kurzfassung Kategorie
Korpusdaten						
ID	+	+	+	+	+	+
Publikationsdatum		+		+	+	+
Änderungsdatum				+	+	+
Download-Datum	+	+	+	+	+	+
URL der Seite	+	+	+	+	+	+

Tabelle 3.1: Übersicht über gespeicherte Daten der Mediendokumente für die analysierten Portale.

	Last.fm	**YouTube**	**IMDb**	**Picasa**	**HuffPost**	**Blogger**
Inhalt						
Titel		+	+			+
Text	+	+	+	+	+	+
Metadaten						
Autor	+	+	+	+	+	+
Sonstiges			Herkunft Autor Nützlichkeit		Fans	Kurzfassung Autoren-Level
Korpusdaten						
ID	+	+	+	+	+	+
ID des Dokuments	+	+	+	+	+	+
Publikationsdatum	+	+	+	+	+	+
Download-Datum	+	+	+	+	+	+
URL der Seite	+	+	+	+	+	+
URL des Dokuments	+	+	+	+	+	+

Tabelle 3.2: Übersicht über gespeicherte Daten der Kommentare für die analysierten Portale.

3.3 Vergleich von Annotier- und Kommentierverhalten

Um eine allgemeine Aussage über Kommentare und Annotationen für Mediendokumente treffen zu können, ist es notwendig, diese für verschiedene Portale zu untersuchen. Die Fragestellung lautet: Wie stark unterscheidet sich das Kommentier- beziehungsweise Annotierverhalten der Internetnutzer zwischen den verschiedenen Medienarten und Portalen? Um diese Frage zu beantworten, wurde der Korpus statistisch untersucht. Tabelle 3.3 gibt eine Übersicht über die absoluten Mengen und durchschnittlichen Verteilungen von Dokumenten und deren Annotationen und Kommentaren.

	Last.fm	**YouTube**	**IMDb**	**Picasa**	**HuffPost**	**Blogger**
Allgemein						
Dokumente	7.280	335.073	24.396	4.268.502	66.392	104.941
Kommentare	282.487	79.672.287	1.110.489	7.381.391	2.851.964	1.026.828
Annotationen						
Dokumente mit Titel	100,00 %	100 %	100,00 %	100,00 %	99,99 %	92,79 %
Wörter pro Titel	2,18	6,64	2,85	2,40	8,30	4,44
Standardabweichung	3,02	3,15	1,54	1,29	3,35	3,01
Dokumente mit Zusammenfassung	99,95 %	95,52 %	88,62 %	40,01 %	99,50 %	95,31 %
Wörter pro Zusammenfassung	259,62	95,77	19,36	3,28	677,37	290,51
Standardabweichung	146,86	173,63	9,61	9,23	623,83	488,76
Dokumente mit Schlagwörtern	100 %	100 %	100,0 %	6,82 %	100 %	42,66 %
Schlagwörter pro Dokument	4,96	15,21	6,69	0,44	9,24	1,05
Standardabweichung	0,32	13,52	1,93	3,31	6,92	1,74
Wörter gesamt pro Dokument	266,77	117,63	28,90	6,12	694,91	296,00
Standardabweichung	146,61	177,51	10,34	10,17	625,46	529,49
Kommentare						
Dokumente mit Kommentaren	100 %	100 %	100 %	100 %	100 %	100 %
Kommentare pro Dokument	38,80	237,78	45,52	1,73	42,96	9,78
Standardabweichung	33,11	329,55	124,04	2,41	289,62	22,45
Wörter pro Kommentar	11,89	17,17	210,11	9,12	25,91	77,27
Standardabweichung	19,01	19,10	158,69	12,20	30,93	257,74
Wörter pro Dokument	461,20	4.082,60	9.564,24	15,78	1.113,11	756,06
Standardabweichung	487,66	6.087,53	27.272,37	43,24	2.502,40	2.447,50
Verhältnis der Textmengen						
Annotationen zu Kommentaren	1 : 1,97	1 : 106,84	1 : 302,16	1 : 4,69	1 : 3,65	1 : 9,43
Standardabweichung	1 : 2,71	1 : 233,49	1 : 788,02	1 : 14,52	1 : 31,00	1 : 200,96

Tabelle 3.3: Übersicht über im Korpus enthaltene Annotationen und Kommentare mit deren durchschnittlicher Wortanzahl und Textmengenverhältnissen

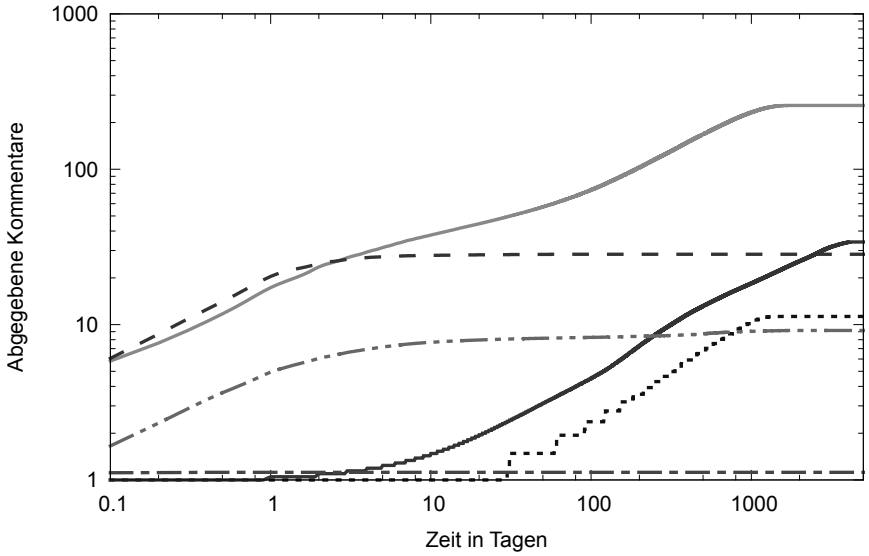
Die sechs betrachteten Portale unterscheiden sich stark in ihrer Größe beziehungsweise in der Anzahl der präsentierten Mediendokumente. Die Anzahl der im Korpus vorhandenen Dokumente, wie in Tabelle 3.3 zu sehen, gibt eine Übersicht der relativen Größenverhältnisse, da absolute Dokument- und Kommentaranzahl für die Portale nicht bekannt sind. Last.fm ist das Portal mit den wenigsten kommentierten Dokumenten, Picasa mit über vier Millionen das mit den meisten. Auch das Annotationsverhalten unterscheidet sich auf den betrachteten Portalen stark. In der Tabelle ist zu erkennen, dass die drei Annotationsformen nicht immer verwendet werden. Auf Picasa werden Zusammenfassungen nur bei etwa 40 %, Schlagwörter lediglich bei 7 % der Dokumente verfasst. Die Ausführlichkeit, mit der Dokumente annotiert werden, unterscheidet sich also zwischen den Portalen. Dies ist auch in der Länge der einzelnen Annotationsformen erkennbar. Während die Dokumenttitel mit 2 bis 8 Wörtern ähnlich lang sind und auch deren Standardabweichung recht gering ist, unterscheiden sie sich in der Länge der Zusammenfassungen hingegen stark. HuffPost hat mit durchschnittlich 677 Wörtern die längsten Zusammenfassungen, was in diesem Fall daran liegt, dass hier die Zusammenfassung das eigentliche Mediendokument (der Nachrichtentext) ist. Bei Blogger und Last.fm sind die Zusammenfassungen mit etwa 250 Wörtern ebenfalls recht lang. Bei IMDb und Picasa werden dagegen mit durchschnittlich 19 beziehungsweise 3 Wörtern Zusammenfassungen mit lediglich einem bis wenigen Sätzen verfasst. Die Standardabweichungen der Wortlängen der Zusammenfassungen sind verglichen mit den jeweiligen Durchschnittswerten sehr groß. Dies liegt darin begründet, dass die Verteilung der Längen einer zipfschen Verteilung und keiner Normalverteilung ähneln. Die Anzahl der Schlagwörter pro Dokument liegt zwischen 15 bei YouTube und 1 bei Blogger. Bei Picasa werden wiederum mit durchschnittlich $0,44$ Wörtern die wenigsten Schlagwörter vergeben.

Zusammenfassend kann man feststellen, dass sich die Ausführlichkeit und der damit verbundene Aufwand des Annotierens auf den Portalen stark unterscheidet. Die durchschnittliche Gesamtwortmenge, von 700 bei HuffPost und etwa 6 bei Picasa zeigt dies sehr deutlich. Das kann einerseits am jeweiligen Portal andererseits auch am jeweiligen Medium liegen. Ein Portal wie

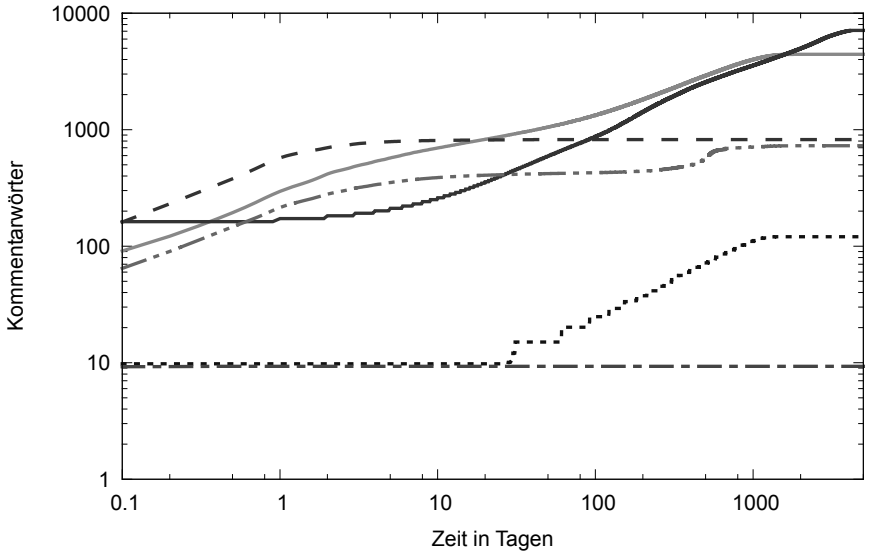
Picasa, wo sehr viele Dokumente präsentiert werden, könnte die Motivation und Sorgfalt beim Annotieren negativ beeinflussen. Die Wahrscheinlichkeit für einen Beitrag eines Autors auf einem Portal wie Picasa gesehen zu werden ist bei der großen Menge präsentierter Dokumente sehr gering, die Sorgfalt und Ausführlichkeit des Autors beim Annotieren des Beitrags auch. Der Aufwand der Erstellung des präsentierten Mediums selbst spielt sicherlich ebenfalls eine wichtige Rolle. Fotos beispielsweise können schnell und in großer Menge erzeugt werden. Recherchierte Nachrichtentexte oder Videos dagegen sind aufwändiger zu erstellen, die Zeit, die benötigt wird Annotationen dafür zu verfassen ist relativ zur Erstellung vergleichsweise gering.

Kommentare unterscheiden sich ebenfalls in Länge und Anzahl zwischen den Portalen. Auf YouTube werden mit durchschnittlich 237 die meisten Kommentare pro Dokument abgegeben. Die Kommentare sind aber mit 16 Wörtern sehr kurz. Im Unterschied dazu sind Kommentare auf IMDb mit etwa 200 Wörtern bedeutend länger aber weniger zahlreich. Dort werden im Durchschnitt nur etwa 45 Kommentare für ein Dokument abgegeben. Trotzdem ergibt sich eine Gesamttextmenge von etwa 10.000 Wörtern für ein Dokument. Picasa hat unter allen Portalen die wenigsten Kommentare pro Dokument und zugleich die kürzesten. Pro Dokument existiert nur eine Textmenge von etwa 15 Wörtern in allen Kommentaren. Eine Aussage über die Motivation des Kommentierens lässt sich daraus schwer treffen. Die Popularität und damit die Besucherzahlen eines Portals sowie die Menge der präsentierten Mediendokumente haben hier sicher einen entscheidenden Einfluss. Die benötigte Zeit für den Konsum des jeweiligen Mediums kann ebenfalls eine wichtige Rolle spielen. Ein Video oder gar einen Kinofilm in voller Länge zu sehen dauert bedeutend länger als sich ein Foto anzuschauen. Die Zeit, die ein Nutzer dem Medium widmet, könnte die Stärke seiner Meinung und die Motivation diese mitzuteilen beeinflussen.

Ein weiterer Unterschied im Kommentarverhalten auf den betrachteten Portalen liegt im zeitlichen Verlauf der Kommentarabgabe. Für diese Analyse wurde der Kommentarzuwachs und der Zuwachs an Wörtern ausgehend vom Datum der Veröffentlichung des Mediendokuments gemessen. Da für IMDb und Last.fm kein Veröffentlichungsdatum des Mediendokuments zu Verfügung



(a) Durchschnittlicher Zuwachs an Kommentaren



(b) Durchschnittlicher Zuwachs der Textmenge aller Kommentare

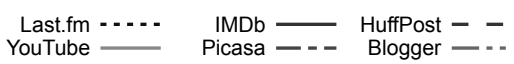


Abbildung 3.1: Zeitlicher Verlauf der Kommentarabgabe

steht, wird als Annäherung das Datum des ersten Kommentars angenommen. In Abbildung 3.1(a) und 3.1(b) ist der zeitliche Verlauf der Kommentarabgabe für alle Portale dargestellt. Bei Picasa ist durch die geringe Kommentarmenge pro Dokument kein Zuwachs ersichtlich. Bei HuffPost und YouTube werden innerhalb des ersten Tages etwa gleich viele Kommentare abgegeben. Bei HuffPost kommen danach aber kaum neue Kommentare dazu, dagegen werden bei YouTube kontinuierlich neue Kommentare abgegeben. Zwischen 100 und 1000 Tagen verstärkt sich der Zuwachs sogar noch etwas. Blogger und HuffPost haben den gleichen Kurvenverlauf, allerdings werden bei Blogger weniger Kommentare in der selben Zeit abgegeben. Auch IMDb, Last.fm und YouTube haben einen ähnlichen Verlauf. Wobei bei IMDb und Last.fm der Kommentarzuwachs erst viel später einsetzt und auch insgesamt weniger Kommentare abgegeben werden.

Aus diesen Messungen kann geschlussfolgert werden, dass die einzelnen Portale unterschiedliche Dynamik besitzen. Bei HuffPost und Blogger werden sehr schnell nach Dokumentveröffentlichung Kommentare abgegeben, aber schon nach wenigen Tagen scheinen Dokumente ihre Aktualität verloren zu haben. Bei IMDb und Last.fm hingegen ist die Aktualität des Dokuments für Nutzer scheinbar weniger wichtig. Dokumente werden erst lange nach ihrer Veröffentlichung wahrgenommen und kommentiert, Kommentare werden dagegen aber auch noch nach drei Jahren für ein Dokument abgegeben. Bei YouTube ist festzustellen, dass sehr schnell nach Veröffentlichung des Mediendokuments das Kommentieren einsetzt und über drei Jahre hinweg annähernd gleich bleibt.

Für das Keyword-Retrieval mit Kommentaren als Quelle ist eher die Textmenge von Bedeutung, um eine genaue Gewichtung der Indexterme zu erhalten. Im Hinblick darauf ist festzustellen, dass der Zuwachs an Kommentarwörtern dem Anstieg der Kommentarmenge bei jedem Portal gleicht. Später abgegebene Kommentare unterscheiden sich in ihrer Wortanzahl nicht wesentlich von früher abgegebenen. Durch die unterschiedlichen durchschnittlichen Kommentarlängen gleichen sich aber die Größenverhältnisse für YouTube, IMDb, HuffPost und Blogger an. Nach einem Tag sind auf diesen vier Portalen etwa 200 bis 400 Wörter im gesamten Kommentartext vorhanden, nach 10 Tagen etwa 300 bis 800 Wörter und nach 100 Tagen 400 bis 1000. Bei Huff-

Post und Blogger stagniert der Zuwachs danach. Es werden nur noch wenige Kommentare abgegeben und die Textmenge vergrößert sich nicht mehr. Bei YouTube und IMDb steigt die Anzahl der Kommentare und damit die Textmenge weiter. Nach 1000 Tagen sind über 40.000 Wörter pro Mediendokument als Kommentartext vorhanden.

Auf allen sechs Portalen besitzen Kommentare mehr Wörter als Annotationen. Es ergeben sich allerdings unterschiedliche durchschnittliche Verhältnisse von Kommentar- zu Annotationstextmenge: bei IMDb gibt es etwa 300 mal mehr Wörter in den Kommentaren als in den Annotationen, bei YouTube ergibt sich ein Verhältnis von 106 zu 1. Das jeweilige Verhältnis der anderen Portale ist etwa zwei Zehnerpotenzen kleiner. Blogger hat ungefähr 10 mal mehr Kommentarwörter als Annotationswörter, Picasa und HuffPost etwa 4 mal mehr und Last.fm 2.

Betrachtet man die Annotationen und Kommentare als mögliche Informationsquellen für das Keyword-Retrieval, so sind die vorgestellten Portale unterschiedlich gut dafür geeignet. Jedes Dokument benötigt eine möglichst große beschreibende Textmenge, um daraus möglichst viele, gut gewichtete Indexterme zu erhalten. Für Picasa stehen im Durchschnitt pro Dokument nur 6 Annotationswörter und 15 Kommentarwörter zur Verfügung. Ohne genaue Messungen kann spekuliert werden, dass hier keine guten Retrieval-Ergebnisse zu erwarten sind. Dokumente auf Last.fm, HuffPost und Blogger besitzen relativ viele Wörter in den Annotationen, haben aber zwei bis zehn mal mehr Kommentarwörter. Für das Keyword-Retrieval sind diese Portale durch beide Textmengen potentiell sehr gut geeignet. Bei YouTube und IMDb gibt es jeweils nur recht wenige Wörter in den Annotationen, aber um so mehr in den Kommentaren. Auf diesen Portalen sollte der Einfluss der Kommentare als Informationsquelle für das Retrieval sehr viel größer sein beziehungsweise sinnvolle Ergebnisse überhaupt erst ermöglichen. Kommentare auf IMDb könnten gegenüber YouTube-Kommentaren den Vorteil haben, dass sie sehr viel länger sind und damit potentiell weniger Wort- und Satzwiederholungen haben.

Kapitel 4

Retrieval-Experimente mit Webkommentaren

Für die Evaluierung eines Retrieval-Systems ist ein Testkorpus mit bekannter Relevanz von Suchanfragen zu allen Dokumenten notwendig (siehe Kapitel 2.5). Alle bekannten Korpora mit bekannten Ergebnisdokumenten besitzen keine Webkommentare. Das selbst erstellte Korpus mit Kommentaren besitzt hingegen keine bekannten Ergebnisdokumente. Dadurch ist keine Einschätzung der Retrieval-Qualität über Precision und Recall sowie über NDCG möglich. In den folgenden Experimenten werden daher verschiedene Retrieval-Eigenschaften und Retrieval-Ergebnisse von Kommentaren untersucht und mit denen der herkömmlichen Annotationen (Titel, Beschreibung und Schlagwörter) verglichen.

Bei den durchgeführten Vergleichen der Ergebnismengen unterschiedlicher Retrieval-Methoden ist zu beachten, dass für keine der Mengen eine Aussage über die Relevanz getroffen werden kann. Abbildung 4.1 zeigt dieses Problem. Bei der Evaluierung von Retrieval-Systemen wird die Schnittmenge zwischen der Ergebnismenge $\mathbf{M}_1$ und der Menge aller relevanten Dokumente $\mathbf{M}_{rel}$ ermittelt und mit der Größe von $\mathbf{M}_1$ beziehungsweise $\mathbf{M}_{rel}$ über Precision und Recall in Beziehung gesetzt. Diese können dann wiederum mit Precision und Recall einer zweiten Ergebnismenge $\mathbf{M}_2$ verglichen werden. Da $\mathbf{M}_{rel}$ für die Experimente nicht zur Verfügung steht und somit Precision und Recall nicht

bestimmbar sind, werden stattdessen die Mengen $\mathbf{M}_1$ und $\mathbf{M}_2$ direkt miteinander verglichen. Die Größe der Schnittmenge zwischen $\mathbf{M}_1$ und $\mathbf{M}_2$ gibt so beispielsweise Auskunft über die Menge der gemeinsam gefundenen Dokumente. Die Überdeckung der einen Menge durch die zweite kann ebenfalls aufschlussreich sein. Zu beachten ist dabei allerdings immer, dass der Anteil von relevanten Dokumenten, welcher eigentlich von Interesse ist, weder in $\mathbf{M}_1$ und $\mathbf{M}_2$, noch in der Schnittmenge bekannt sind. Es kann somit nur die mengenmäßige Veränderung einer Retrieval-Methode (Retrieval mit Annotationen) zu einem anderen (Retrieval mit Kommentaren) eingeschätzt werden.

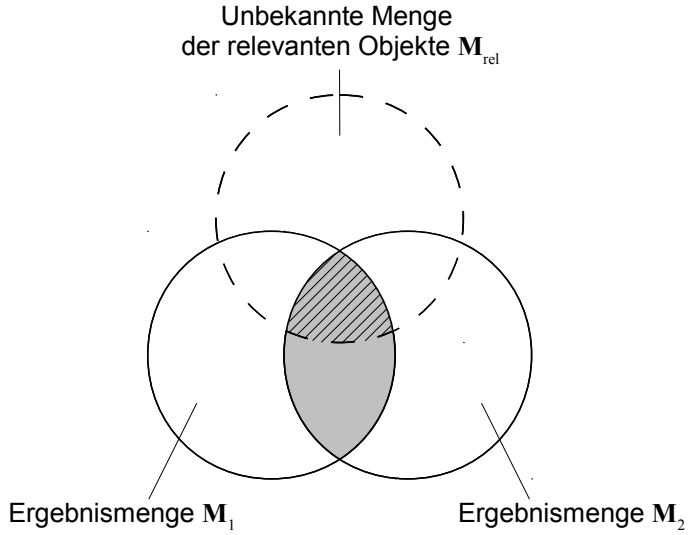


Abbildung 4.1: Vergleich zweier Ergebnismengen einer Suchanfrage, ohne die Menge der relevanten Dokumente zu kennen.

Im Folgenden werden fünf Experimente präsentiert, die das Retrieval mit Kommentaren mit dem Retrieval mit Annotationen vergleichen. Dabei werden der Zweck, die genutzten Maße und der Versuchsaufbau eines jeden Experiments erläutert und die Ergebnisse anschließend ausgewertet. Das Kapitel ist wie folgt gegliedert:

4.1 Technische Aspekte Die in den Experimenten genutzten Verfahren, wie Wortdekomposition, Index- und Retrieval-Verfahren, werden in diesem

Abschnitt vorgestellt.

4.2 Wortschatzvergleich In diesem Experiment wird die Textähnlichkeit und die Schnittmenge des Wortschatzes von Kommentaren und Annotationen bestimmt, um eine Aussage treffen zu können, ob Kommentare Annotationen im Retrieval ersetzen können.

4.3 Diskriminierungseigenschaft Hierin wird die Eigenschaft von Annotations- und Kommentartexten untersucht, Dokumente der Kollektion voneinander unterscheiden zu können.

4.4 Vergleich von Retrieval-Ergebnissen Dieser Versuch vergleicht die Ergebnislisten einer großen Anzahl von realen Suchanfragen des Retrievals mit Annotationen und Kommentaren der sechs Internetportale des Kommentarkorpus.

4.5 Vergleich verschiedener Relevanzsortierungen Da exakte nach Relevanz sortierte Ergebnislisten nicht zur Verfügung stehen, werden in diesem Versuch die Trefferlisten von Websuchmaschinen zum Vergleich mit dem Kommentar-Retrieval herangezogen. Verschiedene Korrelationsmaße geben Aufschluss über die Ähnlichkeit der Sortierungen.

4.6 Manuelle Relevanzbestimmung Im letzten vorgestellten Experiment erfolgt eine manuelle Einschätzung der Relevanz von Suchtreffern. Dazu wurden Treffer ausgewählt, die nur über das Retrieval mit Kommentaren gefunden wurden.

4.1 Technische Aspekte

In den Experimenten werden mehrfach Wortverteilungen und Wortmengen untersucht. Die Wörter wurden dazu durch eine vollständige Wortdekomposition aus den entsprechenden Texten extrahiert. Dafür wurden die Texte zunächst über eine einfache Heuristik gesäubert. Sie wurden dabei in Kleinbuchstaben gewandelt und alle Sonderzeichen, Zahlen und einzeln stehenden Buchstaben wurden durch die Anwendung regulärer Ausdrücke entfernt. Zur Extraktion von Wörtern aus dem fortlaufenden Text wurde die Java-Bibliothek ICU4J des ICU-Projekts[1] verwendet, die aufgrund von Leerzeichen und verschiedenen Heuristiken Einzelwörter erkennt. Zur Wortstammreduktion wurde der Porter-Stemmer-Algorithmus genutzt. Zur Stoppwortentfernung kam eine Liste der häufigsten Stoppwörter zum Einsatz. Daher ist diese vollständige Wortdekomposition sprachabhängig, denn Wortstammreduktion und Stoppwortentfernung sind für die einzelnen Sprachen unterschiedlich. Für die Experimente wird daher die Annahme getroffen, dass alle Texte der Portale in englischer Sprache verfasst sind, da jeweils auch nur die englischen Seiten der Portale gecrawlt wurden. Trotzdem können anderssprachige Texte gerade in den Kommentaren vorkommen. Diese Tatsache wird allerdings nicht näher betrachtet, da sich eine automatische Spracherkennung auf sehr kurzen Texten nicht zuverlässig durchführen lässt.

In allen Experimenten kommt die freie Suchmaschine Lucene[2] der Apache Software Foundation in der Version 3.0.2 zum Einsatz. Die Indexfunktion von Lucene wurde, wie in Kapitel 3.2 beschrieben, zur Speicherung und Organisation des Kommentarkorpus verwendet. Dazu wurden die IDs der Korpusdokumente indiziert. Des Weiteren wurde die Suchfunktion für alle Retrieval-Experimente genutzt. Dazu wurden alle Annotationsformen und alle Kommentare eines Dokuments jeweils addiert und indiziert. Die Wortdekomposition wurde dabei von Lucene automatisch durchgeführt. Dabei kommt ebenfalls der Porter-Stemmer-Algorithmus und eine Stoppwortliste der englischen Sprache zum Einsatz. Für die Suche nutzt Lucene das in Kapitel 2.3 vorge-

[1] `http://site.icu-project.org/`
[2] `http://lucene.apache.org`

stellte Vektorraummodell. Die Relevanzbestimmung erfolgt durch die Berechnung des Skalarprodukts der mit *tf·idf* gewichteten Termvektoren. Die Bestimmung der Termgewichte erfolgt ebenfalls durch Lucene, Veränderungen wurden nicht vorgenommen. Auch Anpassungen weiterer, in Lucene zur Verfügung stehender Parameter wurden nicht vorgenommen, da für das Kommentar- und Annotations-Retrieval keine Prämissen aufgestellt werden konnten und Lucene eine ausgewogene und robuste Standardeinstellung bereitstellt. Die Suche mit Lucene auf indizierten Kommentaren wird im folgenden Text als Kommentar-Retrieval, die Suche auf Annotationen als Annotations-Retrieval bezeichnet.

4.2 Wortschatzvergleich

Im diesem Experiment wird der Wortschatz von Kommentaren und Annotationen verglichen. Dadurch soll eine Einschätzung möglich werden, in welchem Maße sich der Inhalt von Annotationen in den Kommentaren widerspiegelt und ob Kommentare die Annotationen im Keyword-Retrieval als Informationsquelle ersetzen können.

Maße

Als Vergleichsmaße werden dazu die Kosinus-Textähnlichkeit und die Überdeckung der Wortmengen genutzt. Die Kosinus-Textähnlichkeit ist eines der gebräuchlichsten Maße im Information-Retrieval, um die inhaltliche Ähnlichkeit zweier Texte ermitteln zu können. Dabei wird das Skalarpodukt der Vektoren der *tf·idf*-Termhäufigkeit zweier Texte im Vektorraummodell berechnet (siehe 2.3). Zum Vergleich der Überschneidung der Ergebnismengen wird häufig der Jaccard-Koeffizient verwendt, der das Verhältnis der Schnittmenge zur Vereinigungsmenge berechnet. Da Kommentare aber das 2- bis 300-fache an Wörtern im Vergleich zu Annotationen besitzen, ist der Jaccard-Koeffizient hier weniger gut geeignet. Die Größe der Vereinigungsmenge würde zu stark von der Menge der Kommentarwörter dominiert werden und der Koeffizient dementsprechend klein sein. Aus diesem Grund wird die Überdeckung der in den Annotationen verwendeten Wörter durch die Kommentarwörter über den gerichteten Jaccard-Koeffizient berechnet. Die Schnittmenge aus Annotations- und Kommentarwörtern wird hierbei nicht wie im eigentlichen Jaccard-Koeffizienten

durch die Vereinigungsmenge, sondern nur durch die Menge der Annotationswörter geteilt. Es wird für jedes Dokument der Anteil der Wörter der Annotationen bestimmt, die auch in den Kommentaren vorkommen. Die Häufigkeit der Wörter wird dabei nicht betrachtet.

Aufbau

Jedes Portal wurde in diesem Experiment einzeln betrachtet. Es wurde eine vollständige Dekomposition von Annotations- und Kommentartexten jedes Dokuments durchgeführt und Wortüberdeckung sowie Kosinus-Ähnlichkeit berechnet. Für alle Dokumente eines Portals wurde der arithmetische Mittelwert bestimmt. Zusätzlich wurden Wortüberdeckung und Kosinus-Ähnlichkeit in Abhängigkeit von der Anzahl der Kommentartextwörter berechnet.

Textähnlichkeit

Abbildung 4.2 zeigt die durchschnittliche Kosinus-Textähnlichkeit von Annotationen und Kommentaren in Abhängigkeit von der Textlänge der Kommentare. Die durchschnittliche Textähnlichkeit steigt mit wachsender Kommentarlänge auf allen Portalen von 0,0 bis etwa 0,2. Nur auf den Portalen Blogger und HuffPost, die Text als Mediendokument besitzen, steigt die Textähnlichkeit von etwa 0,1 bis auf 0,3 beziehungsweise 0,4 an. Der Anstieg aller Kurven schwächt sich jeweils im hinteren Bereich ab und kehrt sich teilweise sogar um. Mit steigender Anzahl von Kommentarwörtern haben Annotationen und Kommentare demnach mehr Wörter gemeinsam. Ab einem bestimmten Bereich übersteigt die Wortanzahl in den Kommentaren die der Annotationen. Die Textähnlichkeit nimmt daher wie erwartet ab dem Bereich der durchschnittlichen Annotationslänge des jeweiligen Portals ab. Durch das große Ungleichgewicht zwischen Textmengen von Kommentaren und Annotationen verliert die Kosinus-Textähnlichkeit ab diesem Bereich ihre Aussagekraft in Bezug darauf, wie stark sich der Inhalt der Annotationen auch in den Kommentaren widerspiegelt.

Wortschatzüberdeckung

Der Anteil der Annotationswörter, die auch in Kommentaren vorkommen, ist sehr stark von der Länge des Kommentartextes und damit von der Anzahl der Kommentare abhängig. Dies ist in Abbildung 4.3 sehr deutlich zu erkennen.

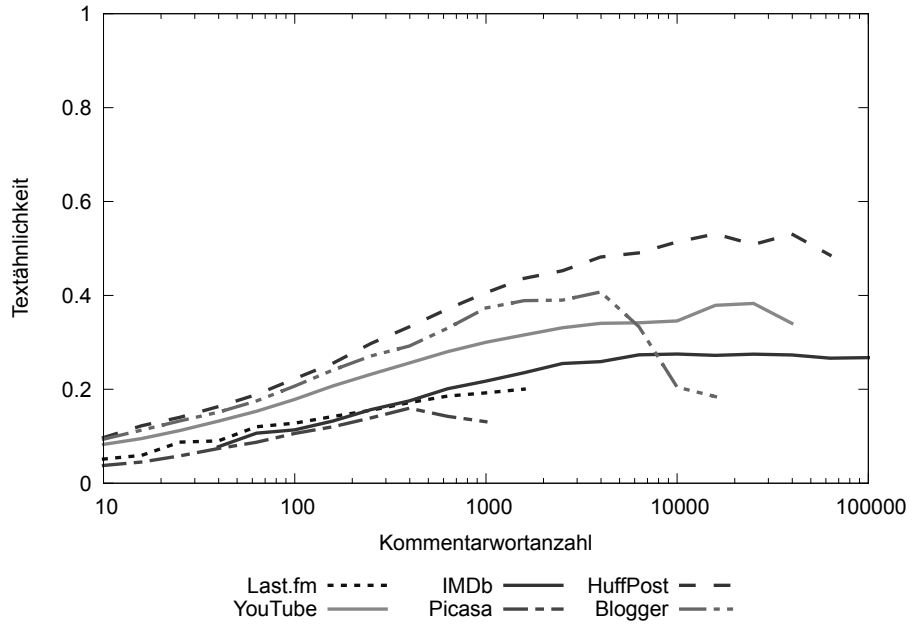
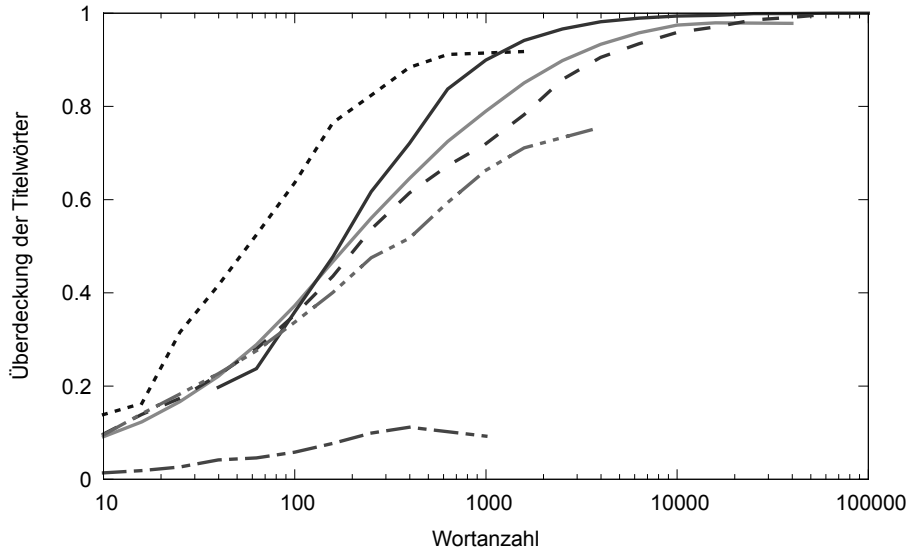


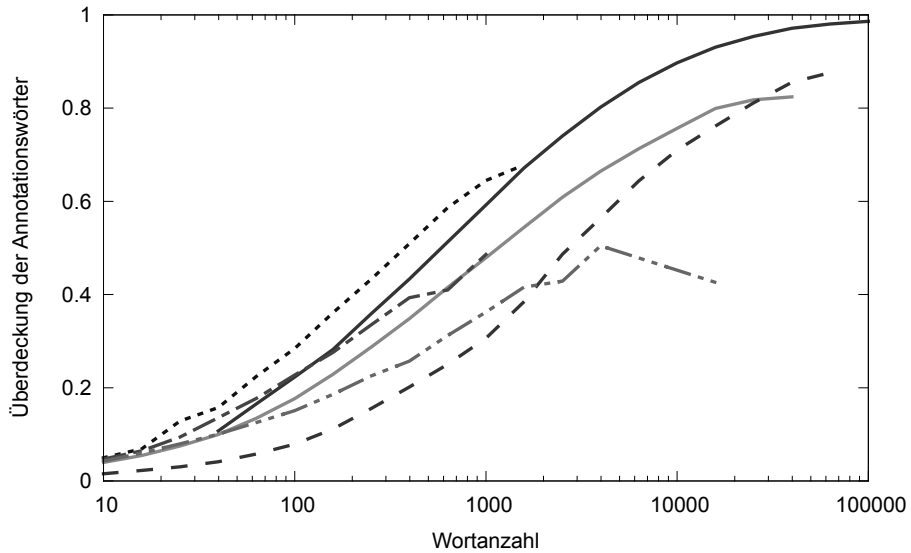
Abbildung 4.2: Durchschnittliche Kosinus-Textähnlichkeit von Annotationen und Kommentaren

Mit steigender Wortzahl nimmt auch die Überdeckung zu und erreicht bei drei der sechs Portale einen Wert von über 80 %. Diese hohen Werte werden allerdings erst ab einer Textmenge von mehreren tausend Kommentarwörtern erreicht. Die Überdeckung der Wörter der Dokumenttitel steigt eher und auch schneller an. Hier wird auch eine vollständige Überdeckung von einigen Portalen erreicht. Die Kurvenverläufe aller Portale ähneln sich stark. Einzig Picasa sticht bei der Überdeckung der Titelwörter heraus. Die Überdeckung steigt dort kaum an und erreicht kaum mehr als 10 %. Dieses Verhalten ist damit zu erklären, dass präsentierte Fotos bei Picasa sehr häufig keine aussagekräftigen Titel besitzen. Oft werden als Dokumenttitel die von der Kamera vergebenen Dateinamen, wie „IMG_4332.jpg“, verwendet. Solche Titel tauchen in den Kommentaren dieser Dokumente nicht wieder auf.

Da sich die Textmenge von Kommentaren zwischen den Portalen unterscheidet und nur auf wenigen Portalen Mengen von mehreren tausend Wörtern erreicht werden, ist die durchschnittliche Überdeckung für die betrachteten Portale sehr unterschiedlich, wie in Tabelle 4.1 zu erkennen ist. Bei IMDb



(a) Durchschnittliche Wortschatzüberdeckung der Titelwörter durch Kommentarwörter



(b) Durchschnittliche Wortschatzüberdeckung der Annotationswörter durch Kommentarwörter

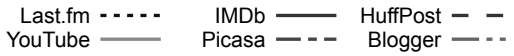


Abbildung 4.3: Durchschnittliche Wortschatzüberdeckung der Titel und aller Annotationen durch Kommentare

	Last.fm	YouTube	IMDb	Picasa	HuffPost	Blogger
Kosinus-Textähnlichkeit	0,14	0,28	0,23	0,03	0,29	0,22
Überdeckung Titel	0,70	0,65	0,88	0,01	0,50	0,37
Überdeckung Annotationen	0,40	0,43	0,68	0,03	0,18	0,18

Tabelle 4.1: Wortschatzvergleich von Annotationen und Kommentaren

wird mit 68 % der höchste Wert erreicht. Annotationen sind dort sehr kurz, Kommentare hingegen sehr lang. Picasa besitzt die geringste Überdeckung. Gerade 3 % der Annotationswörter kommen in den Kommentaren vor. Dies ist nicht verwunderlich bei einer durchschnittlichen Annotationslänge von nur 6 Wörtern und durchschnittlich 16 Kommentarwörtern.

Die Ergebnisse dieses Versuchs können wie folgt zusammen gefasst werden: Kommentare könnten Annotationen in Bezug auf das Auffinden der Mediendokumente ersetzen. Nahezu alle Wörter der Annotationen werden mit zunehmender Kommentarmenge dort ebenfalls benutzt. Die Menge der Kommentarwörter ist hier allerdings ein entscheidender Einflussfaktor. Auf verschiedenen Portalen wird eine entsprechende Kommentarwortzahl nur selten erreicht. Die Suchergebnisse hängen somit stark vom Kommentierverhalten der Internetnutzer ab, was in der Praxis sehr ungünstig wäre.

Die Frage, ob sich Kommentare und Annotationen inhaltlich ähneln, kann nur bedingt beantwortet werden. Die Messung der Überdeckung zeigt, dass Kommentare die Wörter der Annotationen ebenfalls nutzen. Ein inhaltlicher Zusammenhang der beiden Texte ist demnach vorhanden. Dieser ist allerdings wiederum von der Kommentaranzahl abhängig. Die Messung der Kosinus-Textähnlichkeit hat sich als wenig aussagekräftig erwiesen. Die durchschnittliche Ähnlichkeit ist sehr niedrig und eine inhaltliche Ähnlichkeit daraus kaum ableitbar. Dies liegt vor allem am großen Ungleichgewicht der Textlängen von Kommentaren und Annotationen. Auch ist davon auszugehen, dass kurze Annotationen nur wenig zur inhaltlichen Beschreibung des eigentlichen Mediendokuments beitragen. Für eine konkretere Aussage müssten komplexere Vergleichsverfahren als die Standardmethode der Kosinus-Textähnlichkeit genutzt werden.

4.3 Diskriminierungseigenschaft

In diesem Experiment soll ermittelt werden, wie gut Kommentare für das Keyword-Retrieval geeignet sind. Dazu wird die Diskriminierungseigenschaft der in Kommentaren genutzten Wörter betrachtet. Für das Retrieval ist es sehr wichtig, dass ein Dokument möglichst spezifische Wörter enthält, die dieses Dokument von den anderen Dokumenten des Korpus gut diskriminieren.

Maße

Zum Vergleich der Diskriminierungseigenschaft von Kommentaren und Annotationen wird die Kullback-Leibler (KL) Divergenz [Kullback, 1959] verwendet. Die KL-Divergenz gibt die Unwahrscheinlichkeit einer Wahrscheinlichkeitsverteilung zu einer zweiten des selben Ereignishorizonts an:

$$D(P \parallel Q) = KL(P, Q) = \sum_{x \in d} P(x) \cdot \log_2 \frac{P(x)}{Q(x)}$$

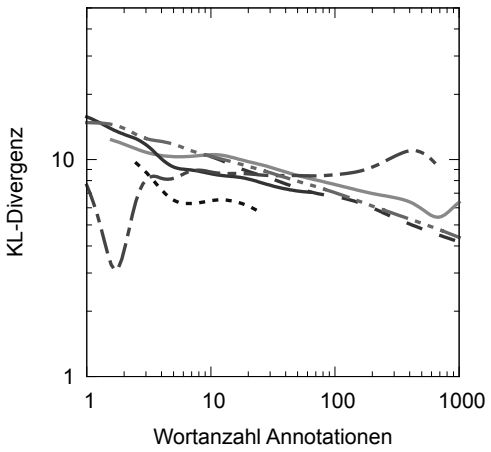
In diesem Experiment wird die Auftrittswahrscheinlichkeit $P(x)$ des Terms im Dokument durch die Termhäufigkeit $tf_d(x)$ im Dokument repräsentiert. Die Auftrittswahrscheinlichkeit $Q(x)$ wird über die Termhäufigkeit $tf_D(x)$ in der gesamten Kollektion ermittelt. Je unwahrscheinlicher das Auftreten eines Terms in einem Dokument gegenüber dem Auftreten in der Kollektion ist, desto mehr Information trägt er und desto größer ist die Diskriminierungskraft gegenüber anderen Dokumenten.

Aufbau

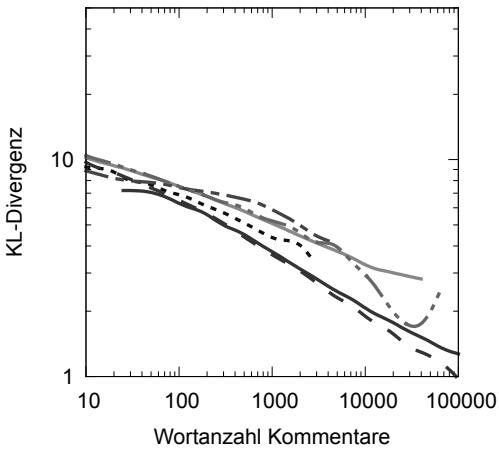
Für das Experiment wurde jedes Portal in zwei Kollektionen geteilt: die eine enthält nur Annotationen, die andere nur Kommentartexte. Für jede Kollektion wurde ein Sprachmodell aus den enthaltenen Termen und deren Auftrittswahrscheinlichkeit gebildet. Die Terme wurden durch die komplette Dekomposition aus den jeweiligen Texten ermittelt. Für jedes Dokument wurde die durchschnittliche KL-Divergenz aller Annotations- und Kommentarwörter ermittelt und über alle Dokumente arithmetisch gemittelt.

Ergebnisse

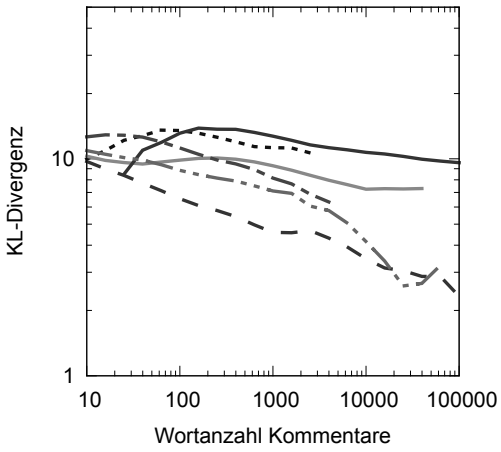
Es wurde beobachtet, dass die durchschnittliche KL-Divergenz mit der Wort-



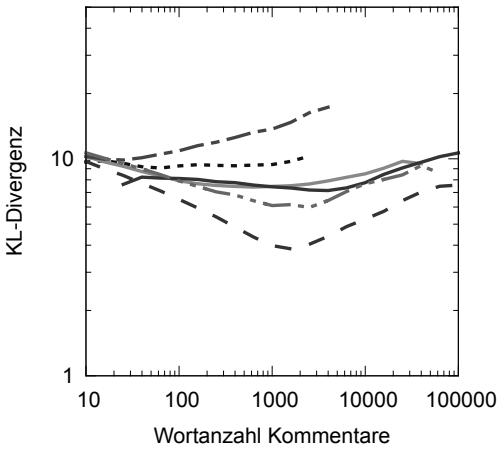
(a) KL-Divergenz der Annotationswörter



(b) KL-Divergenz der Kommentarwörter



(c) KL-Divergenz der n Kommentarwörter mit höchstem $tf{\cdot}idf$-Gewicht (n = Anzahl Annotationswörter des jeweiligen Objektes)



(d) KL-Divergenz der n Kommentarwörter mit mittlerem $tf{\cdot}idf$-Gewicht (n = Anzahl Annotationswörter des jeweiligen Dokuments)

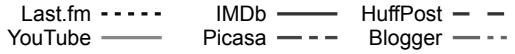


Abbildung 4.4: Durchschnittliche KL-Divergenz von Annotationen und Kommentaren pro Dokument

anzahl negativ korreliert. In Abbildung 4.4(a) und 4.4(b) ist die durchschnittliche KL-Divergenz für Annotationen und Kommentare in Abhängigkeit von der jeweiligen Textmenge dargestellt. Es ist festzustellen, dass die KL-Divergenz mit steigender Wortanzahl sinkt. Besonders bei Kommentaren ist das deutlich zu erkennen, da diese sehr viel länger als Annotationen sein können. Dieses Verhalten ist mit der Eigenschaft der KL-Divergenz zu erklären. Je länger ein Text ist, desto mehr nähert sich die Wortverteilung dem Sprachmodell an, wodurch die KL-Divergenz sinkt. Es ist zu beobachten, dass die Portale mit weniger stark ausgeprägtem Ungleichgewicht von Annotations- zu Kommentartextmenge eine vergleichbare KL-Divergenz besitzen. Dies trifft auf Last.fm, HuffPost und Blogger zu. YouTube und IMDb mit einem sehr starken Ungleichgewicht besitzen eine viel geringere durchschnittliche KL-Divergenz in den Kommentaren, da diese bedeutend länger als die Annotationen sind. Picasa bildet dabei eine Ausnahme. Dort ist die KL-Divergenz trotz sehr weniger Wörter in den Annotationen nur gering. Die Annotationen besitzen somit nur wenig diskriminative Wörter.

Eine negative Korrelation von KL-Divergenz und Wortanzahl bedeutet allerdings nicht, dass die Zahl stark diskriminativer Wörter in längeren Texten abnimmt. Eher ist davon auszugehen, dass der Anteil von wenig diskriminativen Wörtern, wie Stoppwörtern, zunimmt. Da der Kommentartext eines Dokuments meist sehr viel länger als der Annotationstext ist, haben Kommentare auch erwartungsgemäß eine geringere durchschnittliche KL-Divergenz. Aus diesem Grund werden in einem nächsten Schritt die Kommentarwörter jedes Dokuments reduziert, sodass die KL-Divergenz auf der gleichen Anzahl an Annotations- und Kommentarwörtern des Dokuments berechnet werden kann. Zur Reduktion wurden die Kommentarwörter nach ihrer Wichtigkeit für das jeweilige Dokument sortiert und ausgewählt. Die Wichtigkeit wurde über den $tf{\cdot}idf$-Wert des Wortes ermittelt. Zur Bestimmung der KL-Divergenz wurden nur noch die Wörter mit den höchsten $tf{\cdot}idf$-Werten verwendet. Da besonders selten vorkommende Wörter, wie auch Rechtschreibfehler, durch diese Auswahl besonders bei kurzen Texten ein zu hohes Gewicht zugewiesen bekommen, wurden in einem zweiten Versuch Wörter mit mittlerem $tf{\cdot}idf$-Wert ausgewählt.

Die Abbildungen 4.4(c) und 4.4(d) stellen die Verteilung der KL-Divergenz mit reduzierter Wortzahl dar. Bei der Auswahl der Top-*tf·idf*-Terme ist zu erkennen, dass die KL-Divergenz weniger stark absinkt beziehungsweise für vier der sechs Portale kurzzeitig sogar ansteigt. Die Verteilung der KL-Divergenz bei der Auswahl der mittleren *tfidf*-Terme zeigt einen gegensätzlichen Verlauf. Sie fällt auf den meisten Portalen anfangs deutlich, wenn auch weniger stark als bei der Betrachtung aller Terme. Nach Erreichen des Minimums steigt sie wieder deutlich an. Das Minimum wird dabei auf den einzelnen Portalen ungefähr bei der jeweils durchschnittlichen Wortzahl des Kommentartextes eines Dokuments erreicht. Für dieses Phänomen konnte keine Erklärung gefunden werden.

Abschließend kann festgestellt werden, dass die durchschnittliche KL-Divergenz durch die Reduktion der Kommentarwörter auf den Portalen teilweise deutlich steigt. Tabelle 4.2 stellt die verschiedenen ermittelten Werte gegenüber. Auf den Portalen Last.fm, Picasa, HuffPost und Blogger liegt sie deutlich über dem Wert der KL-Divergenz der Annotationen. Bei YouTube und IMDb erreicht sie etwa den Annotationswert.

	Last.fm	**YouTube**	**IMDb**	**Picasa**	**HuffPost**	**Blogger**
Annotationen	6,45	8,62	8,07	5,99	5,19	7,08
Kommentare	6,25	4,82	3,31	9,33	5,63	7,56
Top TFIDF-Terme	12,07	7,96	11,93	11,90	6,13	8,85
Mittlere TFIDF-Terme	9,40	7,27	7,71	10,39	5,88	8,14

Tabelle 4.2: Mittlere KL-Divergenz aus Annotationswörtern und Kommentarwörtern pro Mediendokument

Aus diesem Versuch kann die Erkenntnis gezogen werden, dass Kommentare und Annotationen vergleichbare diskriminative Eigenschaften besitzen. Es konnte darüber hinaus gezeigt werden, dass mit steigender Kommentartextmenge eines Mediendokuments auch die Menge an Wörtern steigt, die dieses Dokument von anderen Dokumenten einer Kollektion unterscheidet. Kommentare könnten für das Retrieval, gerade durch ihre große Textmenge einen Mehrwert zum Auffinden von Multimedia-Dokumenten erzielen. Auch wenn

anzunehmen ist, dass bei Zunahme der Textmenge, die Zahl von für das Retrieval ungeeigneten, häufig vorkommenden Wörtern stark zunimmt, so steigt auch die Zahl der Wörter, die zum Auffinden des kommentierten Dokuments besonders wichtig sein können.

4.4 Vergleich von Retrieval-Ergebnissen

In den beiden vorangegangen Experimenten wurden Eigenschaften von Annotationen und Kommentaren der Korpusdokumente unabhängig von der Suche analysiert. Im folgenden Experiment werden die Ergebnistreffer auf eine große Anzahl konkreter Suchanfragen untersucht. Der Schwerpunkt liegt hier in der Veränderung der Ergebnismenge, wenn Kommentare statt Annotationen als Informationsquelle genutzt werden beziehungsweise auf beiden gemeinsam gesucht wird. Da pro Dokument viel mehr Wörter in den Kommentaren als in den Annotationen vorhanden sind, sollte sich dies auch in den jeweiligen Ergebnismengen widerspiegeln. Die in der Analyse des Wortschatzes gemessene Überdeckung von Annotations- und Kommentartext sollte sich ebenfalls in der Überdeckung der Ergebnismengen feststellen lassen. Zusätzlich zum Mengenvergleich wird die Sortierung der Ergebnisse untersucht. Dabei soll ermittelt werden, wie stark sich die Relevanzsortierungen beider Methoden ähneln.

Maße

Zur Analyse der Suchergebnisse wurden die Treffermengen von Annotations- und Kommentar-Retrieval sowie deren Schnittmengen erfasst. Die Überdeckung von Annotationstreffern durch die Kommentartreffer wurde als gerichteter Jaccard-Koeffizient gemessen. Aus den Größen der Ergebnismengen und deren Schnittmenge ergibt sich ein Verhältnis für die Steigerung der Treffermenge des Retrievals, wenn Annotationen und Kommentare gemeinsam indiziert werden. Zur Analyse der Ergebnissortierungen wurde die Überdeckung der ersten zehn Suchergebnisse unabhängig von ihrem Rang ermittelt. Zusätzlich wurde der Rangkorrelationskoeffizient Kendalls Tau der über beide Verfahren gefundenen Dokumente der gesamten Sortierung und ebenfalls der ersten zehn Ergebnisse berechnet.

Aufbau

Für dieses Experiment wurde für jedes Portal ein Lucene-Indexe aus den Annotationen und ein zweiter aus den Kommentaren der jeweiligen Dokumente erstellt. Jeweils beiden Indexen wurden Anfragen gestellt und die Ergebnismengen über die beschriebenen Maße verglichen. Die Anfragen stammen aus einer Liste von 20 Millionen Suchanfragen, die 2006 von über 650.000 Internetnutzern über einen Zeitraum von drei Monaten an die Internetsuchmaschine von AOL gestellt wurden[3]. Als Parameter für dieses Experiment wurde die Anzahl der Suchwörter betrachtet. Die Verknüpfung der Anfragewörter mit den booleschen Operatoren UND und ODER wurde gesondert betrachtet, da sich die Mengen der Suchtreffer dadurch erheblich unterscheiden. Mit dem Operator UND müssen alle verknüpften Suchwörter in den Ergebnistreffern enthalten sein, wodurch die Menge der Dokumente erheblich kleiner wird als beim Operator ODER, wo nur mindestens eines der Suchwörter enthalten sein muss.

Ergebnisse

Tabelle 4.3 gibt eine Übersicht über die Ergebnismengen und Messwerte für alle Portale mit Suchanfragen der Länge eins. Es ist festzustellen, dass sich die durchschnittliche Menge der gefundenen Dokumente stark zwischen den Portalen unterscheidet. Haupteinflussfaktoren sind die verschiedene Anzahl von Dokumenten der Korpora sowie die Textmengen der jeweiligen Annotationen und Kommentare. Auf YouTube werden mit Abstand die meisten Dokumente beim Kommentar-Retrieval gefunden, was durch die Größe des Korpus und der großen Kommentartextmenge pro Dokument zu erklären ist. Bei Last.fm werden hingegen die wenigsten Dokumente gefunden, das Korpus ist hier mit lediglich 7.000 Dokumenten auch am kleinsten.

Um die Ergebnismengen unabhängig von der Korpusgröße zu bewerten, wird im Folgenden die Steigerung der Gesamtmenge im Verhältnis zur Menge der mit Annotationen gefundenen Dokumente und die Überdeckung der Menge des Annotations-Retrievals durch das Kommentar-Retrieval unter Berücksichtigung der Schnittmenge aus beiden betrachtet. Hierbei fallen besonders die Messwerte des Portals IMDb auf. Dort vergrößert sich die Ergebnismenge um

[3] `http://www.gregsadetsky.com/aol-data/`

	Last.fm	YouTube	IMDb	Picasa	HuffPost	Blogger
Gefundene Objekte						
über Annotationen	3,4	609,3	26,5	282,6	723,8	534,1
über Kommentare	71,0	5.321,3	1.119,6	679,5	727,2	646,8
Verbesserung	76,8	17,6	115,6	3,1	2,4	3,7
Überdeckung						
insgesamt	0,57	0,28	0,67	0,10	0,09	0,13
@10	0,10	0,11	0,14	0,03	0,10	0,10
Rangkorrelation τ						
insgesamt	0,03	0,28	0,14	0,26	0,24	0,37
@10	-0,05	0,12	0,10	0,22	0,13	0,22

Tabelle 4.3: Vergleich der Suchergebnisse für Anfragen der Länge eins auf allen Portalen.

das 115-fache, wenn mit Annotationen und Kommentaren gemeinsam gesucht wird. Etwa zwei Drittel der mit Annotationen gefundenen Dokumente werden dabei auch mit Kommentaren gefunden. Die starke Steigerung ist dabei durch das extrem ungleiche Textmengenverhältnis von Annotationen zu Kommentaren zu erklären. Dokumente bei IMDb besitzen im Durchschnitt etwa 300 mal mehr Wörter als Annotationen. Annotationen sind dazu auch noch sehr kurz. Die Wahrscheinlichkeit, dass ein Suchwort in den Kommentaren vorkommt ist damit viel höher als das Vorkommen in den Annotationen. Die hohe Überdeckung der Annotationsergebnisse durch die Kommentarergebnisse ist bei IMDb der gemessenen Wortschatzüberdeckung aus Abschnitt 4.2 sehr ähnlich.

Für die Portale Picasa, HuffPost und Blogger ist eine Beziehung von Mengensteigerung und Textmengenverhältnis sowie Überdeckung und Wortschatzüberdeckung ebenfalls erkennbar. YouTube hat zwar eine deutlich stärkere Steigerung der Ergebnismenge, diese ist aber im Verhältnis zum Textmengenunterschied von Annotationen und Kommentaren eher gering. YouTube besitzt pro Dokument zwar sehr viele, aber auch sehr kurze Einzelkommentare pro Dokument. Dadurch ergibt sich eine große Wortmenge, die aber sehr viele Textwiederholungen aufweist. Dies könnte erklären, warum die Steigerung hier weniger stark ausfällt. Bei Last.fm wurde ebenfalls eine hohe Mengensteigerung und Überdeckung gemessen, obwohl das Textmengenverhältnis nur etwa 2 zu 1 beträgt. Hier ist dies eher durch starke statistische Schwankungen

zu begründen. Das Korpus von Last.fm ist sehr klein, woraus sich sehr kleine Ergebnismengen zur Auswertung ergeben. Dadurch schwanken die Messwerte zu stark und der arithmetische Mittelwert ist wenig aussagekräftig. Auch die anderen Messwerte zeigen einen starken Unterschied zu den anderen Portalen. Aufgrund der geringen Korpusgröße wird Last.fm in diesem Versuch nicht weiter betrachtet.

In der Relevanzsortierung der Ergebnisdokumente unterscheidet sich das Retrieval auf Annotationen und auf Kommentaren sehr stark auf allen Portalen. Der Rangkorrelationskoeffizient τ ist dabei auf IMDb mit $0,14$ am geringsten. Die Sortierung der mit den beiden Methoden gefundenen Dokumente besitzt eine sehr geringe positive Korrelation. Die Trefferreihenfolgen unterscheiden sich demnach stark voneinander. YouTube, Picasa und HuffPost haben eine ähnliche Rangkorrelation von etwa $0,25$. Bei Blogger liegt die Rangkorrelation bei $0,36$ und ist somit doppelt so hoch wie bei IMDb. Trotzdem kann bei diesem Wert kaum von einer ähnlichen Sortierung gesprochen werden. Die Rangkorrelation der obersten 10 gemeinsam gefundenen Dokumente ist auf allen Portalen geringer als die Rangkorrelation aller gemeinsamen Dokumente. Dies verstärkt die Aussage, dass die Ergebnissortierung des Annotations-Retrievals der des Kommentar-Retrievals kaum ähnelt. Die Verfahren bewerten demnach die Relevanz eines Dokuments zur Suchanfrage völlig unterschiedlich.

Der Relevanzwert ergibt sich bei beiden Verfahren über die Wichtigkeit des Suchwortes für das Dokument. Im Vektorraummodell wird dieser Wert über *tfidf*, also die Termhäufigkeit und die inverse Dokumentfrequenz bestimmt. Bei der Suche mit einem einzelnen Wort spielt die inverse Dokumentfrequenz für die Sortierung keine Rolle, da sie nicht dokumentabhängig und somit für alle gefundenen Dokumente gleich ist. Die unterschiedliche Bewertung der Relevanz eines Dokuments durch die beiden Verfahren resultiert demnach ausschließlich aus der unterschiedlichen Termhäufigkeit des Suchwortes in Annotationen und Kommentaren des Dokuments. Annotationen besitzen, wie in Kapitel 3.3 analysiert, auf den meisten Portalen nur sehr wenige Wörter. Das Suchwort, lässt man Stoppwörter außer Acht, kann in solch kurzen Texten nur ein bis wenige Male vorkommen. Die Termhäufigkeit *tf* verrechnet aber die Auftrittshäufigkeit des Wortes mit der Gesamtzahl der Wörter. Dementsprechend hat diese

Gesamtzahl bei sehr kurzen Texten einen großen Einfluss auf den tf-Wert und damit auf den Relevanzwert des Dokuments. Demzufolge kann geschlussfolgert werden, dass die Relevanzberechnung und somit auch die Relevanzsortierung von kurzen Annotationstexten sehr ungenau ausfällt. Die Wortmengen in Kommentaren sind auf fast allen Portalen wesentlich größer. Dadurch kann die Berechnung der Wichtigkeit eines Suchwortes in diesen Texten viel differenzierter und genauer ausfallen. Der Unterschied der Textlängen ist somit eine mögliche Erklärung für unterschiedliche Sortierungen gleicher Dokumente durch die verschiedenen Methoden. Bei den Kommentartexten ist allerdings zu beachten, dass diese aus der Summe aller Einzelkommentare des jeweiligen Dokuments bestehen. Welchen Einfluss dies auf die Relevanzsortierung mit Kommentaren hat, konnte in diesem Experiment nicht ermittelt werden.

Anzahl der Anfragewörter	**1**	**2**	**3**	**4**	**5**
Gefundene Objekte					
über Annotationen	609,3	6.572,7	13.101,9	17.690,8	22.186,2
über Kommentare	5.321,3	38.270,0	69.322,2	88.328,4	104.153,3
Verbesserung	17,6	9,4	8,0	7,4	6,9
Überdeckung					
insgesamt	0,28	0,36	0,42	0,45	0,48
@10	0,11	0,09	0,09	0,09	0,08
Rangkorrelation τ					
insgesamt	0,28	0,28	0,25	0,23	0,22
@10	0,12	0,09	0,09	0,11	0,12

Tabelle 4.4: Vergleich der Suchergebnisse für ODER-verknüpfte Anfragen verschiedener Länge auf dem Portal YouTube.

In diesem Versuch wurde ebenfalls die Abhängigkeit der Messwerte von der Anzahl der Suchwörter gemessen. Die Suchwörter der AOL-Suchwort-Liste wurden dafür mit ODER und mit UND verknüpft. Die Tabellen 4.4 und 4.5 zeigen die Messergebnisse für das Portal YouTube. Dabei ist klar zu erkennen, dass die Ergebnismengen für beide Methoden bei der Verknüpfung mit ODER mit steigender Suchwortanzahl stark zu und bei der UND-Verknüpfung stark abnehmen. Dies entspricht der Erwartung. Als richtiger Suchtreffer für die ODER-Verknüpfung zählt jedes Dokument, das mindestens eines der Such-

Anzahl der Anfragewörter	**1**	**2**	**3**	**4**	**5**
Gefundene Objekte					
über Annotationen	609,3	153,3	51,20	17,5	4,9
über Kommentaren	5.321,3	4.190,1	2.206,9	1.159,4	788,12
Verbesserung	17,6	72,5	229,6	456,3	634,1
Überdeckung					
insgesamt	0,28	0,20	0,15	0,13	0,13
@10	0,11	0,08	0,04	0,03	0,02
Rangkorrelation τ					
insgesamt	0,28	0,30	0,29	0,29	0,25
@10	0,12	0,22	0,24	0,23	0,16

Tabelle 4.5: Vergleich der Suchergebnisse für UND-verknüpfte Anfragen verschiedener Länge auf dem Portal YouTube.

wörter besitzt. Dadurch kommen viel mehr Dokumente des Korpus als Treffer in Frage. Die Bedingung der UND-Verknüpfung ist dagegen viel härter: Jedes Suchwort muss im Text vorkommen, wodurch sehr viel weniger Dokumente als Treffer gelten.

Interessant ist hierbei, dass auf YouTube die Anzahl der Treffer aus Annotationen und Kommentaren unterschiedlich stark steigt beziehungsweise sinkt. Das Anwachsen der Treffermenge bei steigender Suchwortanzahl mit ODER-Verknüpfung ist beim Annotations-Retrieval stärker ausgeprägt als beim Retrieval mit Kommentaren, die Abnahme der Treffermenge bei UND-Verknüpfung allerdings ebenfalls. Die ist auch bei den anderen Portalen, ausgenommen Picasa, zu beobachten. Dieses Phänomen ist auch in der gemessenen Steigerung der Treffermenge beim gemeinsamen Retrieval mit Annotationen und Kommentaren auf allen Portalen, bis auf Picasa, zu beobachten. Der Grund für dieses Verhalten konnte nicht eindeutig bestimmt werden.

Die von beiden Methoden gefundenen Treffer und damit die Überdeckung von Annotationstreffern durch die Kommentartreffer steigt ebenfalls bei ODER-Verknüpfung auf allen Portalen an. Bei IMDb werden bei einer Suchwortanzahl von 5 im Durchschnitt sogar über 80 % der mit Annotationen gefundenen Dokumente auch mit Kommentaren gefunden. Bei Verknüpfung der Suchwörter mit UND ist ein gegenläufiges Verhalten zu beobachten, die Überdeckung sinkt. Dieses Verhalten ist zu erwarten und mit der Bedingung von UND- beziehungsweise ODER-Verknüpfung zu erklären.

Weder die Anzahl der Suchwörter noch deren Verknüpfung hat scheinbar Einfluss auf die Rangkorrelation der Relevanzsortierungen. Die Messwerte für Kendalls Tau der gesamten Sortierung und Kendalls Tau der ersten zehn Treffer verändern sich kaum auf den betrachteten Portalen.

Aus diesem Experiment wird die Erkenntnis gezogen, dass Kommentare die Treffermenge des Keyword-Retrievals auf Multimedia-Dokumenten immens steigern können. Haupteinflussfaktor auf die Steigerung ist das Verhältnis von Kommentar- zu Annotationslänge. Je größer dieses Verhältnis ist, desto mehr Treffer sind über Kommentare zu finden. Des Weiteren kann resümiert werden, dass sich die Relevanzsortierung der beiden Verfahren stark unterscheidet.

4.5 Vergleich verschiedener Relevanzsortierungen

Die Sortierung der Suchergebnisse ist für das Information-Retrieval von großer Bedeutung. Aus diesem Grund wird die Relevanzsortierung von Kommentaren in diesem Versuch genauer untersucht. Weder für die Sortierung des Annotations-Retrievals noch des Kommentar-Retrievals ist ohne ein Korpus mit bekannter Relevanz von Suchanfragen zu allen Dokumenten eine Aussage zur Qualität zu treffen. Daher werden die Sortierungen zweier Websuchmaschinen, als in der Praxis genutzte Methoden zum Vergleich herangezogen. Die beiden Websuchmaschinen sind dabei als Black-Box anzusehen, da weder deren Dokumentkollektionen noch die genutzten Relevanzfunktionen bekannt sind. Zusätzlich werden verschiedene synthetische Sortierungen durchgeführt und mit allen anderen verglichen. Eine zufällige Sortierung wurde ebenfalls durchgeführt.

Da die Webanfrage einer großen Menge von Suchanfragen sehr zeitintensiv ist, wurde nur eines der sechs Portale für das Experiment ausgewählt. Das Portal YouTube besitzt die größte Menge an Mediendokumenten und hat dabei, wie in Kapitel 3.3 analysiert wurde, relativ viele Annotationswörter sowie eine Vielzahl an Kommentaren. Als Websuchmaschinen wurden Bing[4] von Microsoft und die YouTube-eigene Suche ausgewählt, da für beide ein API zur

[4] `\http:\\bing.com`

Verfügung steht. Es wurden jeweils die 100 besten Suchergebnisse der Suchmaschinen betrachtet. Das Experiment wurde auf 30 Rechnern über einen Zeitraum von zwei Wochen mit über einer Million Anfragen durchgeführt. Für etwa 400.000 Suchanfragen wurden Ergebnisse zurückgeliefert und ausgewertet.

Maße

Als Vergleichsmaße wurden der Korrelationskoeffizient Kendalls Tau und die Überdeckung jeweils einer Sortierung durch eine andere für jede Suchanfrage gemessen und über alle Anfragen arithmetisch gemittelt. Kendalls Tau wurde für die Gesamtmenge der von beiden gefundenen Treffer jeweils zweier Sortierungen sowie für die besten 10 gemeinsamen Treffer erfasst.

Aufbau

An beide Suchmaschinen wurden Anfragen der AOL-Anfragenliste gestellt und die Ergebnisse als Listen von Dokument-IDs gespeichert. Im Kommentarkorpus nicht vorhandene Dokumente wurden von YouTube nachgeladen und in das Korpus eingefügt. Auf dem Korpus wurde danach die selbe Anfrage auf Annotationen und Kommentaren ausgeführt. Für das Experiment wurden nur Ergebnistreffer betrachtet, die auch von Bing oder YouTube gefunden wurden. Als synthetische Sortierungen wurden jeweils die durchschnittliche Nutzerbewertung, die Anzahl der Aufrufe, die Menge der Kommentarwörter sowie eine Zufallsfunktion verwendet. Alle Ergebnislisten wurden für jede Anfrage miteinander über die beschriebenen Maße verglichen. Für den Versuch wurde die Anzahl der Suchwörter als Parameter betrachtet und alle Maße in Abhängigkeit dazu ebenfalls bestimmt. Alle Suchwörter wurden mit ODER verknüpft, da dies wie im vorigen Experiment ermittelt, die Treffermenge erheblich steigert.

Ergebnisse

In Tabelle 4.6 sind die Ergebnisse des Vergleichs der Kommentar-, der Annotations- und der Zufallssortierung zu den beiden Suchmaschinen sowie des Vergleichs der Suchmaschinen untereinander dargestellt. Bei der Überdeckung der Webtreffer durch die Suchtreffer des Kommentar-Retrievals ist festzustellen, dass diese im Durchschnitt höher ist als die Überdeckung der Annotationstreffer im vorigen Versuch. Zu Bing liegt die Überdeckung ab 2 Suchwörtern schon

Suchmaschine	**Bing**					**YouTube**				
Anzahl der Anfragewörter	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Kommentarsortierung										
Überdeckung	0,31	0,72	0,76	0,73	0,72	0,23	0,39	0,44	0,64	0,66
Rangkorrelation τ	-0,02	0,06	0,10	0,10	0,10	0,01	-0,06	-0,04	-0,05	-0,05
Rangkorrelation τ@10	-0,01	0,08	0,09	0,09	0,09	0,01	-0,04	-0,02	-0,06	-0,06
Annotationssortierung										
Überdeckung	0,86	0,94	0,93	0,92	0,90	0,88	0,97	0,97	0,98	0,96
Rangkorrelation τ	0,20	0,10	0,11	0,10	0,10	0,20	-0,02	0,00	-0,04	-0,03
Rangkorrelation τ@10	0,13	0,11	0,12	0,12	0,12	0,12	0,02	0,04	-0,01	-0,01
Sortierung YouTube										
Überdeckung	0,27	0,04	0,01	0,03	0,03	1,00	1,00	1,00	1,00	1,00
Rangkorrelation τ	0,21	0,17	0,16	0,22	0,22	1,00	1,00	1,00	1,00	1,00
Rangkorrelation τ@10	0,16	0,16	0,16	0,22	0,22	1,00	1,00	1,00	1,00	1,00
Zufällige Sortierung										
Überdeckung	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Rangkorrelation τ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Rangkorrelation τ@10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Tabelle 4.6: Vergleich der der Ergebnislisten des Kommentar-Retrievals mit den Suchmaschinen Bing und YouTube für unterschiedliche Anfragelängen.

bei über 70 %. Dies könnte ein Hinweis darauf sein, dass bei Bing auch Kommentartexte zur Suche verwendet werden. Zu beachten ist dabei allerdings, dass die Trefferlisten der Suchmaschinen nur maximal 100 Dokumente enthalten. Demnach ist ebenfalls möglich, dass Treffer, die auch mit Kommentaren gefunden werden, in den Webtreffern aus anderen Gründen in die Top-100 sortiert werden, die tatsächliche Überdeckung aber niedriger ist. Ein weiterer Unterschied im Vergleich der Webtreffer und der Annotationsergebnisse des vorigen Versuchs ist, dass die Rangkorrelation hier noch einmal deutlich niedriger ist. Sie liegt im Vergleich zu Bing zwischen annähernd 0 und $0,1$ und mit YouTube bei 0 und $-0,06$. Es ist also festzustellen, dass es keine Übereinstimmung der Sortierungen der Websuchen und des Kommentar-Retrievals gibt. Die Bewertung der Relevanz von Dokumenten unterscheidet sich demnach deutlich zwischen den Suchmaschinen und dem Kommentar-Retrieval. Grund dafür können eine verschiedene Datenbasis und unterschiedliche zugrunde liegende Modelle beziehungsweise Relevanzfunktionen sein.

Die Vermutung, dass die beiden Suchmaschinen unterschiedliche Modelle nutzen, wird durch den Vergleich der Annotations-Retrievalergebnisse und der Webtreffer erhärtet. Hierin ist eine sehr hohe Trefferüberdeckung zu be-

obachten. Nahezu alle Dokumente, die von den Websuchmaschinen gefunden werden, sind auch mit dem Annotations-Retrieval auffindbar. Die Suchwörter kommen bei diesen Dokumenten demnach immer in den Annotationen vor. Dokumente, wo dies nicht der Fall ist, liefern die Suchmaschinen kaum. Daraus kann die Schlussfolgerung gezogen werden, dass der durchsuchte Index bei allen drei Methoden gleich ist. Dass der Wert der Überdeckung von Webtreffern durch Annotationstreffer trotz gleicher Datenbasis nicht genau 1 beträgt, könnte möglicherweise in unterschiedlicher Wortdekomposition beziehungsweise Wortstammreduktion der Dokumenttexte und Suchwörter durch die verschiedenen Methoden begründet liegen. Websuchmaschinen korrigieren auch häufig automatisch offensichtliche Rechtschreibfehler in den Suchwörtern, dies wurde im Retrieval auf den Annotationstexten nicht durchgeführt. Trotz hoher Überdeckung ist die gemessene Rangkorrelation zwischen Websuche und Annotationen sehr niedrig. Des Weiteren ist ein starker Sprung zwischen einem und zwei Suchwörtern festzustellen, der vermutlich in unterschiedlicher Gewichtung der Suchwörter bei ODER-Verknüpfung begründet liegt.

Unter der Annahme, dass auch die Suchmaschinen auf Basis von Annotationen Dokumente in ihrer Kollektion suchen, muss geschlussfolgert werden, dass sich die Relevanzfunktionen der verschiedenen Methoden sehr stark unterscheiden. Die Relevanzbestimmung erfolgt demnach bei den Websuchmaschinen nicht durch die Bestimmung der Kosinus-Ähnlichkeit von Termhäufigkeitsvektoren wie im genutzten Modell beim Annotations-Retrieval. Die Bewertung der Relevanz der Suchtreffer unterscheidet sich im Übrigen auch zwischen den beiden Websuchen stark. Auch die Überdeckung der Webtreffer ist sehr niedrig. Da die verglichenen Trefferlisten der Suchmaschinen nur maximal 100 Dokumente enthalten, ist die gemessene Überdeckung die der Top-100 Treffer beider Listen. Die Rangkorrelation der Trefferlisten ist ebenfalls sehr niedrig und beträgt im Durchschnitt nur $0,2$. Sie ist etwa gleich der Korrelation zwischen Kommentartreffern und Annotationstreffern des vorigen Versuchs.

Weitere Beobachtungen, die bei der Auswertung diese Versuchs gemacht wurden, werden im Folgenden kurz aufgezeigt:

Beim Vergleich der Rangkorrelation der Kommentartreffer und der Sortierung dieser Treffer nach Kommentartextmenge und Kommentaranzahl, konnte

ein hoher negativer Wert von etwa $-0,5$ bei der Suchwortmenge 1 beobachtet werden. Das bedeutet, dass in der Relevanzsortierung des Kommentar-Retrievals Dokumente mit geringerer Kommentartextmenge beziehungsweise wenigen Kommentaren bevorzugt werden. Dieses Verhalten ist wahrscheinlich ebenfalls mit den im vorigen Versuch beschriebenen Eigenschaften der $tf{\cdot}idf$-Gewichtung auf sehr kurzen Texten zu erklären und weist darauf hin, dass diese Art der Gewichtung als weniger geeignet für das Kommentar-Retrieval anzusehen ist.

Des Weiteren wurde eine relativ hohe Korrelation von $0,35$ zwischen Bingtreffern und der Sotrierung nach Nutzerbewertung sowie der Sortierung nach Kommentaranzahl gemessen. Demnach haben die Nutzerbewertung und auch die Kommentaranzahl einen Einfluss auf die Sortierung durch Bing. Die Beeinflussung könnte direkt sein, indem diese Daten in der Relevanzberechnung verwendet werden oder auch indirekt, beispielsweise durch die Auswertung von Hyperlinks. Ein YouTube-Video, das von YouTube-Nutzern sehr gut bewertet beziehungsweise sehr häufig kommentiert wurde, wird sehr wahrscheinlich auch häufiger von anderen Internetseiten referenziert und verlinkt.

Als Vergleichsbasis und zur Evaluierung der Messmethoden wurden alle Treffersortierungen mit einer zufälligen Sortierung verglichen. Alle Ergebnisse entsprachen dabei den Erwartungen. So wurde stets eine durchschnittliche Rangkorrelation von genau 0 gemessen. Die Verteilung der Korrelationswerte entsprach genau der zu erwartenden Standard-Normalverteilung.

Weiterhin wurde in diesem Experiment der Einfluss der Kommentartextmenge auf die Relevanzsortierung untersucht. Dabei konnte aber keine signifikante Veränderung der Rangkorrelation im Vergleich zu den verschiedenen Treffersortierungen festgestellt werden.

Zusammenfassend wird aus diesem Versuch die Erkenntnis gezogen, dass sich die Treffersortierung des durchgeführten Retrievals mit Kommentaren sehr deutlich von den Relevanzsortierungen der beiden untersuchten Websuchmaschinen unterscheidet. Da die Suchmaschinen als Black-Box anzusehen sind, kann über die Gründe nur spekuliert werden. Die Annahme, dass beide Annotationen als Hauptinformationsquelle verwenden, konnte mit verschiedenen Vergleichsergebnissen bekräftigt werden. Die Treffersortierung des Retrievals

auf Annotationen mit einem einfachen Vektorraummodell unterscheidet sich ebenfalls stark von den beiden Websortierungen. Websuchmaschinen nutzen meist komplexere Modelle, die zusätzliche Metadaten in die Relevanzberechnung einfließen lassen. Auf deren Relevanzsortierung haben diese Daten offensichtlich einen großen Einfluss. Dennoch muss festgestellt werden, dass auf beiden in der Praxis oft verwendeten Suchmaschinen, das Problem der Relevanzbewertung für Multimedia-Dokumente ungelöst bleibt, da sich beide verglichenen Sortierungen deutlich voneinander unterscheiden. Somit kann keine Erkenntnis in Bezug auf die Qualität des Kommentar-Retrievals durch einen Vergleich mit diesen Suchmaschinen gewonnen werden.

4.6 Manuelle Relevanzbestimmung

Ziel dieses Experiments ist es, die Qualität des Kommentar-Retrievals einzuschätzen. Es soll ermittelt werden, ob die nur durch das Kommentar-Retrieval gefundenen Suchtreffer relevant für die jeweilige Anfrage sind und auch in eine sinnvolle Reihenfolge sortiert werden. Da diese Aufgabenstellung nur mit einem erheblichen Aufwand gelöst werden kann, wurden für diesen Versuch nur jeweils die besten drei Suchtreffer des vorhergehenden Experiments untersucht, die nur vom Kommentar-Retrieval, nicht aber von den Websuchmaschinen und dem Annotations-Retrieval gefunden wurden. Eine Versuchsperson wies jedem der drei Trefferdokumente des Videoportals YouTube einen dreistufigen Relevanzwert zu, nachdem sie sich diese Videos angeschaut hat. Die drei Relevanzabstufungen waren: relevant, verwandt und irrelevant. Ein Treffervideo wurde als relevant für eine Anfrage gewertet, wenn es inhaltlich den Kern der Suchanfrage traf beziehungsweise die gesuchten Begriffe einen wichtigen Aspekt des Inhalts des Videos darstellten. Als verwandter Treffer galt ein Video, das nur thematisch zur Anfrage passte. Die Suchbegriffe mussten dabei nur als Nebenaspekte in den untersuchten Videos vorkommen. Irrelevante Suchtreffer durften keinen inhaltlichen Bezug zur Anfrage besitzen.

Ergebnisse

Bei der Durchführung des Versuchs stellte sich heraus, dass es für die Versuchspersonen, besonders bei Anfragen mit einem Wort, schwierig war, das Informationsbedürfnis einzuschätzen, um die Relevanz von Videos zur Anfra-

ge zu bestimmen. Diese Anfragen waren häufig zu allgemein, wie beispielsweise „shortcut“, „hardware“ oder „nicknames“. Für viele Anfragen war auch fraglich, ob sie über das Medium Video überhaupt zu beantworten sind wie zum Beispiel: „text applications“, „picture upload“ oder „postal codes“. Diese Anfragen erzielten bei der Suche zwar Treffer, da die Suchwörter auch in Kommentaren vorkamen, die Treffervideos waren aber meist irrelevant. Aus diesem Grund wurde der Versuch einerseits auf Anfragen mit zwei und drei Suchwörtern beschränkt und andererseits den Versuchspersonen zusätzlich die Aufgabe gestellt, nur Anfragen auszuwerten, die sie selbst als per Video beantwortbar einschätzten.

Nach Auswertung der besten drei Suchtreffer von insgesamt 281 Anfragen wurden durch drei Testpersonen 28 % der Videos als irrelevant, 27 % als verwandt und 45 % als relevant für die jeweilige Anfrage bewertet. 7 % der Suchanfragen lieferten keine relevanten oder verwandten Treffer. 21 % der Anfrageergebnisse enthielten einen, 26 % zwei und 46 % drei relevante oder verwandte Treffer. Besonders viele relevante Treffer wurden mit Eigennamen, wie Personen- und Produktnamen sowie medizinischen Begriffen gefunden. Zusammengesetzte Bezeichnungen, die Ortsnamen enthalten, wie beispielsweise „Minnesota State University“ ergaben meist keine oder nur verwandte Treffer.

Innerhalb der relevanten Suchtreffer konnten drei Klassen identifiziert werden, die begründen, warum diese Treffer nur mit Kommentaren gefunden wurden:

Ungenaue Annotationen Die Annotationen dieser Videos sind offensichtlich zu kurz und haben kaum beschreibenden Charakter.

Synonyme Die Wörter der Suchanfrage deckten sich inhaltlich vollständig mit den Annotationen, stellten aber Synonyme der darin verwendeten Wörter dar.

Nebenaspekte im Video Der Autor des Videos erachtete diese Aspekte in den Annotationen nicht als erwähnenswert. Innerhalb der Kommentare wurden diese aber dennoch erwähnt.

Für irrelevante Suchtreffer konnten ebenso drei Kategorien ermittelt werden,

warum diese Treffer nicht relevant für die Suchanfrage sind, obwohl die Suchwörter in den Kommentaren enthalten sind:

Einzelne Suchwörter dominieren Eines der Suchwörter kommt in den Kommentaren sehr häufig vor und bezieht sich inhaltlich auf das Video. Die Zusammensetzung der gesamten Anfrage beschreibt aber ein anderes Informationsbedürfnis.

Irrelevante Zusatzinformationen Die Suchwörter werden in den Kommentaren benutzt, um Aspekte zu erläutern, die über den Inhalt des Videos hinausgehen und durch das Video selbst nicht beantwortet werden.

Spam Das Vorhandensein der Suchwörter in den Kommentaren hat keinerlei Relevanz für das Video. Diese Kommentare stellen häufig einen Missbrauch der Kommentierfunktion dar.

Die Suchergebnisse können durchaus als überraschend bezeichnet werden. Weniger als ein Drittel der Top-3-Treffer wurden als irrelevant angesehen und bei weniger als einem Drittel der Anfragen waren unter diesen Treffern weniger als 2 relevante oder verwandte Videos. Bei der Bewertung gilt es zu beachten, dass bei all diesen Treffern die Suchwörter ausschließlich in den Kommentaren enthalten waren und nicht in den Annotationen genannt wurden. Ebenso ist für die Suchanfragen, die keine relevanten Treffer in den Top-3-Ergebnissen enthielten, unbekannt, ob für diese Anfragen überhaupt relevante Dokumente im gesamten Korpus enthalten waren. Das Keyword-Retrieval mit Webkommentaren liefert demnach relevante Suchtreffer, die ohne Kommentare nicht gefunden werden. Für eine genauere Analyse und Messung der Retrieval-Qualität auf Kommentaren müssten allerdings mehr Suchanfragen ausgewertet und auch deren gesamte Sortierung der Ergebnislisten nach Relevanz bewertet werden. Es wurden verschiedene Kategorien für relevante und irrelevante Treffer identifiziert. Diese Kategorien müssten ebenfalls genauer untersucht werden, um die Relevanzberechnung des Kommentar-Retrievals zu verbessern.

Kapitel 5

Zusammenfassung und Ausblick

Im folgenden Kapitel werden die Ergebnisse der durchgeführten Experimente kurz zusammengefasst und in Bezug auf den Nutzen von Webkommentaren für das keyword-basierte Multimedia-Retrieval bewertet. Weiterführend werden aus den gewonnenen Erkenntnissen weitere Analysen vorgeschlagen und verschiedene Möglichkeiten aufgezeigt, ein Kommentar-Retrieval in der Praxis zu ermöglichen.

5.1 Einschätzung des Nutzens von Webkommentaren

Bei der Auswahl der Experimente wird deutlich, dass die Aussagekraft der Messungen im Hinblick auf die Retrieval-Performance stark mit dem Aufwand der Versuchsdurchführung korreliert. Abbildung 5.1 zeigt diesen Trade-Off von Qualität und Quantität der Messungen. Die Analyse des Wortschatzes und der Diskriminierungseigenschaft konnte für alle Dokumente der sechs Portale des Experimentkorpus durchgeführt werden. Sie geben aber nur wenig Aufschluss über den Beitrag von Kommentaren auf das Retrieval. Die manuelle Relevanzmessung hingegen konnte nur bei wenigen Dokumenten und Suchanfragen eines Portals durchgeführt werden, da hier individuell jedes einzelne Dokument beurteilt wurde. Dafür ist die Aussagekraft dieses Experiments sehr groß. Würde dieses für sehr viele Dokumente und Anfragen durchgeführt, wäre

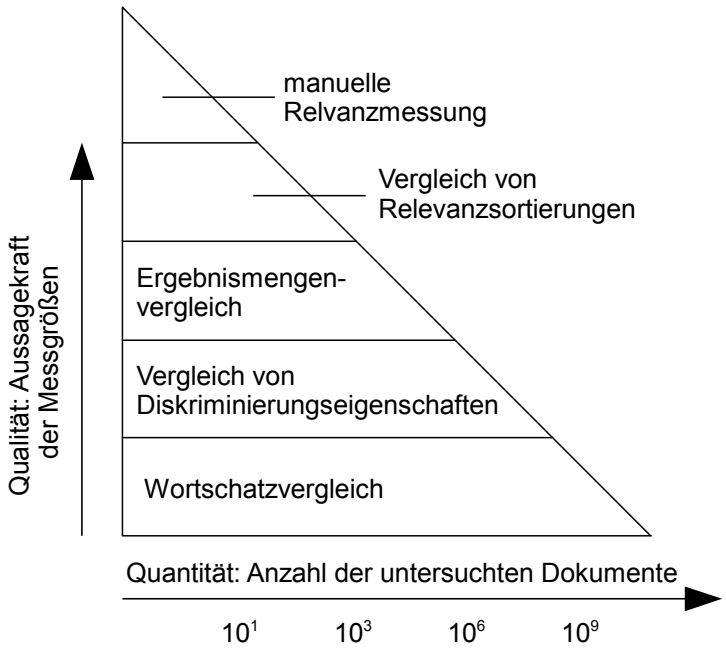


Abbildung 5.1: Aufbau der Experimente. Die Aussagekraft korreliert mit dem Aufwand der Experimente.

der Einfluss direkt bestimmbar, da diese Daten einem Evaluierungskorpus entsprechen würden. Die Ergebnisse der durchgeführten Experimente ermöglichen aber dennoch durch ihre Auswahl, eine breitere Aussage über die Eigenschaften von Kommentaren und deren Einfluss auf das Keyword-Retrieval zu treffen.

Der Wortschatzvergleich zeigte, dass Kommentare theoretisch in der Lage sind, Annotationen in Bezug auf das Finden von Mediendokumenten, welches nur durch das Vorhandensein der Suchwörter in den Dokumenten bestimmt wird, zu ersetzen. Praktisch ist dies aber nur selten möglich, da es maßgeblich von der Kommentartextmenge abhängig ist, ob alle Annotationswörter in den Kommentaren vorkommen. Der Vergleich der inhaltlichen Ähnlichkeit von Kommentaren und Annotationen erbrachte hingegen keine eindeutige Aussage. Als Hauptproblem der Messungen erwies sich die meist geringe Wortmenge der Annotationen. Hier müssten andere Messmethoden gefunden werden, die

eine genauere Aussage ermöglichen. In Bezug auf die Diskriminierungseigenschaft sind Kommentarwörter und Annotationswörter im Gesamtdurchschnitt gleichwertig. Es konnte im Versuch aber gezeigt werden, dass Kommentare durch ihre sehr viel größere Wortmenge, diskriminierendere Wörter enthalten und somit besser für das Keyword-Retrieval geeignet sind.

Ein direkter Retrieval-Vergleich von Annotationen und Kommentaren mit einer großen Anzahl von realen Suchanfragen zeigte, dass Kommentare die Menge der Suchtreffer erheblich steigern konnten. Den größten Einfluss auf diese Steigerung hatte wiederum die Menge an Kommentartext beziehungsweise das Verhältnis von Kommentar- zu Annotationstext. In der Größe der Schnittmenge von Annotations- und Kommentartreffern spiegelte sich die Überdeckung der jeweiligen Texte der Wortschatzanalyse wider. Im Vergleich der Relevanzsortierungen zeigte sich, dass das Kommentar-Retrieval die Relevanz von Dokumenten zu Suchanfragen gegenüber dem Annotations-Retrieval und auch gegenüber zweier Websuchmaschinen sehr unterschiedlich bewertet. Bei der Bewertung dieses Ergebnisses ist allerdings zu beachten, dass auch die Websuchmaschinen nachweislich sehr unterschiedliche Relevanzbewertungen verwenden. Für keine der betrachteten Sortierungen ist bekannt, welche der Realität am besten entspricht. Eine Aussage über die Qualität des Kommentar-Retrievals ist durch den Vergleich mit den Websuchmaschinen nicht möglich.

Die manuelle Bewertung einer geringen Menge von Suchtreffern, die ausschließlich mit Kommentaren gefunden wurden, zeigte, dass mehr als zwei Drittel dieser Ergebnisse relevant in Bezug auf ihre Anfrage waren. Für eine genauere Einschätzung der Relevanz von Kommentarergebnissen muss dies allerdings mit einer größeren Menge von Suchanfragen und für die gesamte Treffermenge jeder Suchanfrage durchgeführt werden. Ebenso müssten andere Retrievalergebnisse, wie beispielsweise von Websuchmaschinen, bewertet werden, um eine Vergleichsbasis zu haben. Die Betrachtung der Relevanz von Kommentartreffern zeigte gleichzeitig die Stärke von Kommentaren gegenüber Annotationen im Retrieval. Besonders Dokumente mit sehr kurzen Annotationen werden erst durch deren Kommentare überhaupt auffindbar. Ebenfalls kann der Einfluss der Subjektivität und Sorgfalt des Autors beim Annotieren eines Mediendokuments eingeschränkt werden. Da Kommentare von vielen

verschiedenen Autoren verfasst werden, kommen darin verschiedene Beschreibungen, aber auch verschiedene Aspekte des Mediendokuments vor, die erst das Auffinden des Dokuments ermöglichen.

Zusammenfassend kann der Nutzen von Kommentaren für das Keyword-Retrieval von Mediendokumenten wie folgt bewertet werden: Kommentare besitzen das Potential das Retrieval enorm zu verbessern. Sie besitzen ähnliche Retrieval-Eigenschaften wie Annotationen. Sie können die Schwäche von Annotationen, die hauptsächlich in der geringen Textlänge liegen, ausgleichen und so das Auffinden von Mediendokumenten deutlich verbessern. Die größte Stärke der Kommentare ist deren Textlänge beziehungsweise Anzahl. Diese ist allerdings von Portal zu Portal und auch von Medium zu Medium unterschiedlich ausgeprägt. Ebenso ist sie zeitabhängig. Neu erstellte Mediendokumente würden erst gefunden werden, wenn genügend viele Kommentare vorhanden sind. Kommentare sollten deshalb das existierende Annotations-Retrieval ergänzen, da sie eine vorhandene, aber ungenutzte Ressource an Information darstellen.

Die Qualität der Relevanzbestimmung und die daraus resultierende Relevanzsortierung des Kommentar-Retrievals kann aus den Ergebnissen der durchgeführten Experimente nicht aussagekräftig bewertet werden. Dies ist hauptsächlich dem Mangel einer Vergleichsmöglichkeit geschuldet. In der Praxis genutzte Retrieval-Methoden stellen keine Vergleichsbasis dar, da diese die Relevanz zur Suchanfrage nachweislich unterschiedlich bewerten.

5.2 Weiterführende Analysen und verbesserte Relevanzberechnungen

Um den Nutzen und speziell die Qualität des Kommentar-Retrievals genauer bestimmen zu können, wäre es möglich die vorgestellten Experimente zu vertiefen und mit anderen Maßen und Messmethoden zu ergänzen. Ohne eine gute Vergleichsbasis werden diese Analysen aber nur bedingt aussagekräftigere Ergebnisse erzielen. Der nächste logische Schritt ist somit eine Vergleichsbasis in Form eines Korpus mit bekannter Relevanz von Suchanfragen und Medi-

endokumenten zu entwickeln. Bei der manuellen Erstellung eines Relevanzkorpus könnten Crowd-Sourcing-Methoden wie beispielsweise Amazon Mechanical Turk[1] eingesetzt werden. Dort könnte die Bewertung von Mediendokumenten zu Suchanfragen als Aufgabe gestellt werden, die gegen einen entsprechende Entlohnung von einer Vielzahl von Internetnutzern bearbeitet wird.

Eine andere Möglichkeit für den Aufbau eines Relevanzkorpus ist die Anreicherung eines bekannten Relevanzkorpus, der keine Kommentare besitzt, mit Kommentaren eines Korpus, der keine Relevanzdaten besitzt. Dokumente des Relevanzkorpus, die Dokumenten des Kommentarkorpus sehr stark ähneln, werden deren Kommentare übertragen. Zur Bestimmung der Ähnlichkeit könnten beispielsweise Verfahren der Bildanalyse genutzt werden. Die Übertragung zwischen Dokumenten unterschiedlicher Medienarten wäre ebenfalls, wie von Potthast et al. [2010] gezeigt wurde, möglich.

Mit einem solchen Relevanzkorpus ist ebenfalls die Erforschung und Entwicklung spezieller Relevanzfunktionen für das Kommentar-Retrieval möglich. Der Fakt, dass der Kommentartext eines Mediendokuments aus einer Vielzahl von Einzelkommentaren besteht, müsste dabei besonders berücksichtigt werden und wurde in den Experimenten der vorliegenden Arbeit außer Acht gelassen. Die Termgewichtung über Termhäufigkeit und inverse Dokumenthäufigkeit im Vektorraummodell kann beispielsweise durch die Auftrittshäufigkeit des Terms in den Einzelkommentaren des Dokuments sinnvoll ergänzt werden. Des Weiteren ist denkbar, die Texte der Einzelkommentare nicht, wie in den vorgestellten Experimenten zu addieren, sondern zu überlagern oder zusammenzufassen. Die Filterung von ungeeigneten Kommentaren und Spam sollte ebenfalls für ein praxistaugliches Retrieval-System eine Rolle spielen.

Für ein besseres Multimedia-Retrieval sollten Kommentare und Annotationen gemeinsam genutzt werden. Hierfür müssen ebenfalls neue Relevanzfunktionen gefunden werden, die beide Textformen unter Berücksichtigung ihrer speziellen Eigenschaften nutzen. Learning-to-Rank-Methoden sind dafür sehr gut geeignet. Die Computerrepräsentationen der Dokumente bestehen dort aus Vektoren verschiedener Eigenschaften. Die Gewichtung dieser Vektoren wird durch eine große Zahl von Trainingsbeispielen in Form von Suchanfragen und

[1] `http://aws.amazon.com/mturk/`

dazu relevanten Dokumenten gelernt.

Für alle vorgeschlagenen Verbesserungen ist allerdings ein Relevanzkorpus unerlässlich, ohne den Kommentare weiterhin als Informationsquelle für das Information-Retrieval ungenutzt bleiben. Der Aufwand der Erstellung muss angesichts des gezeigten Potentials von Kommentaren unbedingt bewältigt werden.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*. ACM Press books. ACM Press.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine* 1. *Computer networks and ISDN systems*, 30(1-7):107–117.

Kullback, S. (1959). *Information theory and statistics*. Wiley.

Mishne, G. and Glance, N. (2006). Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*. Citeseer.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Potthast, M. (2009). Measuring the descriptiveness of web comments. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 724–725. ACM.

Potthast, M., Stein, B., and Becker, S. (2010). Towards comment-based cross-media retrieval. In *Proceedings of the 19th international conference on World wide web*, pages 1169–1170. ACM.

Potthast, M., Stein, B., Loose, F., and Becker, S. (2011). Information Retrieval in the Commentsphere. Accepted Paper for: ACM Transactions on Intelligent Systems and Technology (TIST).

Salton, G. and McGill, M. (1983). *Introduction to modern information retrieval*, volume 1. McGraw-Hill New York.

Stock, W. (2006). *Information Retrieval: Informationen suchen und finden*. Einführung in die Informationswissenschaft. Oldenbourg.

Yee, W., Yates, A., Liu, S., and Frieder, O. (2009). Are Web User Comments Useful for Search? *Proc. LSDS-IR*, pages 1613–0073.