UNIVERSITÄT KOBLENZ · LANDAU

Universität Koblenz-Landau
WeST - Institute for
Web Science and Technologies

Entwicklung und Evaluation eines Verfahrens zur Diversifizierung von strukturierten Inhalten in Sozialen Medien

Diplomarbeit

vorgelegt von

Andreas Ens
201210299

Betreuer Dr. Dr. Sergej Sizov
Erstprüfer Prof. Dr. Steffen Staab

Koblenz, den 27.02.2012

Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

	Ja	Nein
Mit der Einstellung der Arbeit in die Bibliothek bin ich einverstanden.	□	□
Der Veröffentlichung dieser Arbeit im Internet stimme ich zu.	□	□

(Ort, Datum) (Unterschrift)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abstract

Das Kernthema dieser Arbeit ist die Untersuchung von *Diversity* (Diversität) im Kontext sozialer Plattformen. Konkret weist die Arbeit jedoch zwei Schwerpunkte auf: Zum einen wird ein entropie-basiertes Verfahren für die Diversifizierung von Suchergebnissen auf strukturierten Daten vorgestellt. Das Verfahren verwendet die bedingte Entropie, um aus einer gegebenen Menge strukturierter Daten eine diverse Menge von Ressourcen zu bestimmen. Zum anderen wird ein neuartiges Verfahren für die Evaluierung von Diversitätsmaßen vorgestellt, das anhand von Tag-Zuweisungen zu den Ressourcen einer Menge die Diversität dieser Menge messen kann. In einem Experiment, das auf realen Daten der *Bibsonomy* Plattform durchgeführt wird, werden die Ergebnisse des entropie-basierten Verfahrens mit denen eines Baseline-Diversitätsverfahrens verglichen. Hierbei kommt das in dieser Arbeit entwickelte Evaluationsverfahren zum Einsatz, um die Diversität der Ergebnisse beider Verfahren zu messen.

1 Einleitung

Trotz des inzwischen fast schon angestaubten Begriffes Web 2.0, erfreuen sich die hinter dem Begriff stehende Ideologie, die Techniken und insbesondere die daraus resultierenden Produkte wachsender Beliebtheit, mit stetig wachsenden Nutzerzahlen und immer größer werdenden Inhalten, auch allgemein hin als *Content* bezeichnet. Eine ganze Gattung von Web 2.0 Produkten stellen so genannte *Social Resource Sharing Systeme* dar. Im Grunde handelt es sich hierbei um einen Oberbegriff für Plattformen im Web, in denen die Nutzer Ressourcen aller Art hochladen und damit anderen Nutzern zur Verfügung stellen können. Neben der simplen Speicherung und Veröffentlichung von Ressourcen, können Nutzer diese mit beliebigen Schlagwörtern, den so genannten *Tags*, beschriften und damit die Suche nach bestimmten Ressourcen vereinfachen.

Klassifizieren lassen sich Social Resource Sharing Systems am besten anhand der Ressourcen, die in ihnen veröffentlicht werden können. Das allseits bekannte *YouTube* beispielsweise ist eine Plattform für Videos aller Art, während in *Flickr* Fotografien bereitgestellt werden können. Eine Untergruppe der Social Resource Sharing Systems bilden so genannten *Social Bookmark Plattformen*. Unter den bekannteren Vertretern im deutschen Raum sind *Bibsonomy*, *Delicious* und *Mister-Wong*, allerdings existieren noch zahlreiche mehr. Sie ermöglichen das Speichern und Verwalten von Lesezeichen und in manchen Fällen von Publikationen, so beispielsweise in Bibsonomy.

Derartige Web 2.0 Plattformen sind seit Jahren ein Triebmittel für die Webforschung. Ein kleiner Teilaspekt der Webforschung stellt die Diversität (*Diversity*) von Suchergebnissen dar. Um zu verstehen, was sich hinter dem Begriff Diversität verbirgt, eignet sich ein in zahlreichen wissenschaftlichen Publikationen verwendetes Beispiel. Man stelle sich einen Nutzer eines Web Automobil Portals vor. Das Portal erlaubt es seinen Kunden nach Automobilen zu suchen und dabei bestimmte Attribute zu definieren - Attribute

wie beispielsweise *Modell*, *Hersteller*, *Farbe*, *Alter*, *Laufleistung*, *Motorisierung* und viele mehr. Sucht unser Nutzer nach einem Ford Mustang und definiert darüber hinaus keine weiteren Attribute, ist es wahrscheinlich, dass er zwar viele Treffer bekommt, jedoch auf der ersten Seite - je Seite werden 10 Treffer angezeigt - ausschließlich rote Mustangs neueren Baujahres aufgeführt werden. Hat unser fiktiver Nutzer jedoch eine Aversion gegen rote Mustangs, bleibt ihm nichts anderes übrig als weiter zu blättern oder seine Suche zu konkretisieren. Konkretisiert er beim zweiten Versuch seine Anfrage zu stark, d.h. er gibt zu viele Parameter vor, erhält unser Nutzer sehr wahrscheinlich eine kleine Menge homogener Ergebnisse und ihm entwischt u.U. ein Schnäppchen oder eine Rarität. Um das beschriebene Problem zu umgehen, würde es sich anbieten, dem Nutzer bei seiner ersten Suche, insbesondere auf der ersten Seite, unterschiedliche Ford Mustangs zu präsentieren. Unterschiedlich in dem Sinne, dass neben roten Mustangs auch schwarze, blaue und rosa Modelle unterschiedlicher Baujahre, unterschiedlicher Motorisierung sowie unterschiedlicher Preisklassen vertreten sind.

Dieses Beispiel lässt sich leicht auf Social Bookmark Systems und konkret auf Publikationen erweitern. Auch hier möchten Nutzer, wenn sie zu einem bestimmten Thema wissenschaftliche Publikationen suchen, keine Duplikate oder sehr ähnliche Publikationen als Ergebnis erhalten. Viel mehr wünschen sie sich unterschiedliche Blickwinkel und Lösungsansätze für die betrachtete Problemstellung und die damit verbundene Anfrage an das System.

1.1 Motivation

Sowohl die Entwickler bzw. die Betreiber von Suchmaschinen im Web als auch die der Social Bookmark Plattformen stehen im Wesentlichen vor den selben Problemstellungen und Herausforderungen: Sie müssen mit immer größer werden Datenmengen zurechtkommen und dabei stets die Nutzerzufriedenheit im Auge behalten. Selbst eine junge und im Wesentlichen auf die Forschergemeinde fokussierte Plattform, wie beispielsweise

Bibsonomy, kommt bereits auf weit über 350.000 Lesezeichen und 650.000 Publikationen (vgl. [BHJ+10]), der Datensatz von Delicious ist um ein Vielfaches größer. Der Aspekt der Nutzerzufriedenheit kann viele Aufgaben umfassen: Ergonomische Eigenschaften der Oberfläche, Reaktionszeiten für Anfragen und vieles mehr. In dieser Diplomarbeit wird jedoch ausschließlich die Diversität von Suchergebnissen betrachtet.

Die meisten Nutzer, so *Drouso* und *Pitoura* [DPP10], sind eher an Informationsbrocken interessiert, die jedoch alle Teile ihres Informationswunsches abdecken, statt lediglich zu einem Teil des Informationswunsches viele und u.U. sehr ähnliche Ergebnisse zu bekommen. Um homogene Ergebnislisten zu vermeiden, wurden in den letzten Jahren Anstrengungen unternommen, um Verfahren und Algorithmen zu entwickeln, die diversifizierte Ergebnislisten generieren. Im Kapitel *Related Work* werden einige Ansätze näher erläutert. Ein wenig erforschter Bereich ist die Diversifizierung von Suchergebnissen auf *strukturierten Daten*. Hierbei handelt es sich um Daten, die in der Regel in relationalen Datenbanken vorliegen und eine gleichartige Struktur aufweisen.

Neben der Entwicklung neuer Diversitätsverfahren für die unterschiedlichsten Anwendungsdomänen stellt sich in der Forschung oftmals die Frage, wie man derartige Verfahren evaluieren kann. Bei einer genauen Recherche von Veröffentlichungen, die sich der beschriebenen Thematik widmen, wird deutlich, dass alle Evaluierungsansätze nicht die Diversität von Suchergebnissen erfassen, sondern lediglich die Nutzerzufriedenheit, die eine Folge diversifizierter Suchergebnisse ist.

1.2 Zielsetzung

Die Zielsetzung dieser Diplomarbeit ist im Grunde zweigeteilt: Das erste Teilziel besteht darin, ein Diversitätsmaß zu definieren und ein Verfahren zu implementieren, das auf strukturierten Daten und in der Domäne der Web 2.0 Plattform Bibsonomy, im Rahmen

eines ausgewählten Anwendungsszenarios für eine Anfrage diverse Suchergebnisse generieren kann. Die Art der Suchergebnisse ist beschränkt auf Publikationen und aus einer kompletten Ergebnisliste soll das Verfahren eine maximal diverse Submenge von Publikationen bestimmen. Inwieweit das Verfahren und damit die zugrunde liegende Metrik tatsächlich dazu im Stande ist, diverse Submengen von Publikationen zu bestimmen, wird anhand eines ausgewählten Experiments auf dem Bibsonomy Datensatz demonstriert.

Das zweite und wesentliche Teilziel der Arbeit ist die Definition sowie die erste Erprobung eines neuartigen Verfahrens für die Evaluierung von Diversitätsmaßen. Die wesentliche Aufgabe hierbei ist die Entwicklung eines Verfahrens, das anhand von Tag-Zuweisungen die Diversität einer Menge von Ressourcen messen kann. Hierfür muss jedoch gezeigt werden, dass sich Tags für das Gruppieren von Ressourcen anhand inhaltlicher Merkmale eignen. Neben dieser Grundbedingung muss eine Metrik definiert werden, welche die durch die Tags ausgedrückte Diversität einer Menge von Ressourcen in numerischer Form quantifizieren kann. Die angesprochene Erprobung des Evaluationsverfahrens wird anhand eines Experiments durchgeführt. Bei diesem Experiment werden die Ergebnisse eines einfachen Diversitätsverfahrens mit denen des in dieser Arbeit definierten Diversitätsverfahrens verglichen (evaluiert). Das Experiment soll die grundlegende Eignung des neuartigen Evaluationsverfahrens für die Evaluierung von Diversitätsmaßen zeigen.

1.3 Gliederung

Gegliedert ist die vorliegende Diplomarbeit wie folgt: Nach diesem einleitenden ersten Kapitel, in dem zum einen die Motivation für die Untersuchung von Diversität auf strukturierten Daten und zum anderen die Ziele der Arbeit genannt wurden, werden im zweiten Kapitel thematisch verwandte Arbeiten genannt und einige interessante Ansätze kurz erläutert. Das dritte Kapitel widmet sich den technischen Grundlagen, die in den nachfolgenden Definitionen verwendet werden. Der Grund für das Einfügen eines derartigen Kapitels ist im Wesentlichen damit begründet, dass einige der nachfolgenden Definition

ein relativ hohes Maß an Komplexität aufweisen und für deren Verständnis eine gewisse Grundlage geschaffen werden soll. In Kapitel 4 wird das entropie-basierte Diversitätsmaß definiert und dessen Implementierung, in Form eines lauffähigen Frameworks, präsentiert. Kapitel 5 bildet den Kern der vorliegenden Arbeit und beschäftigt sich mit dem Evaluationsverfahren für Diversitätsmaße. Hier wird zum einen diskutiert, inwieweit sich Tags für das inhaltliche Gruppieren von Ressourcen eignen und zum anderen die Metrik definiert, mit der die Diversität einer Menge von Ressourcen gemessen werden kann. In Kapitel 6 wird schließlich anhand eines ausgewählten Experiments gezeigt, dass sich das Evaluationsverfahren grundlegend für die Evaluierung von Diversitätsmaßen eignet. Kapitel 7 liefert eine Zusammenfassung der wichtigsten Ergebnisse der Arbeit und einen Ausblick auf weiterführende Aktivitäten.

2 Related Work

In den folgenden beiden Unterkapiteln werden verwandte Arbeiten besprochen. Zum einen werden interessante Ansätze für Diversitätsmaße vorgestellt und kurz erläutert. Zum anderen sollen Evaluationsverfahren für Diversitätsmaße vorgestellt werden.

2.1 Diversitätsmaße und Methoden

Die Diversifizierung von Suchergebnissen ist beispielsweise im Bereich des *Information Retrieval* (IR) oder der Web-Suche kein wirklich neuer Gegenstand der Forschung. Eine der ersten Arbeiten zu diesem Thema veröffentlichten 1998 *Carbonell* und *Goldstein* [CGG98]. Inzwischen ist Diversity ein weit gefasster Begriff, in dem unterschiedlichste Interpretationen von dem aggregiert werden, was man unter diversen Suchergebnissen versteht, d.h. wie man Diversität definiert oder diese evaluiert. Eine sehr gelungene Übersicht zum Thema Diversity wurde 2010 von *Drosou* und *Pitoura* [DPP10] veröffentlicht. Sie teilen die gängigen Diversity Definitionen in drei Klassen ein:

- Content-based Diversity Definitions
- Novelty-based Diversity Definitions
- Coverage-based Diversity Definitions

Content-based Diversity Definitionen basieren auf Variationen des $p - dispersion$ Problems. *Drosou* und *Pitoura* beschreiben das Problem als die Auswahl von p aus n Punkten (in einem Raum beliebiger Dimension) auf die Weise, dass die Distanz für jedes Paar von ausgewählten Punkten maximal ist. In anderen Worten kann man diesen Sachverhalt so beschreiben, dass für eine Menge von Items eine Submenge ausgewählt werden soll, die nur aus solchen Items besteht, die paarweise möglichst unähnlich zueinander sind.

Die Novelty-based Diversity Definitionen verbinden Novität mit Diversität. Dabei wird argumentiert, dass ein Item genau dann als divers bezüglich der bereits bekannten Items angesehen wird, wenn es neue und damit in den bekannten Items noch nicht enthaltene Informationen enthält.

Die Klasse der Coverage-based Diversity Definitionen aggregiert die Definitionen, die zu einem durch die Anfrage definierten Topic möglichst viele Subtopics bestimmen und die den Subtopics zugrunde liegenden Items den Nutzern zurück liefern. Im Grunde handelt es sich hier um clustering-basierte Ansätze: Die Ergebnisse für eine Anfrage werden in zusammenhängende Cluster unterteilt und aus jedem Cluster wird ein Repräsentant entnommen. Entropie-basierte Ansätze, wie der in dieser Arbeit, fallen im Grunde in diese Klasse von Diversity Definitionen.

Drosou und *Pitoura* diskutieren zahlreiche Publikationen und unterteilen diese auf Basis der drei genannten Klassen. Auffällig ist jedoch, dass nur in wenigen Publikationen Diversität auf strukturierten Daten untersucht wird. Im Wesentlichen handelt es sich hierbei um die Publikationen von *Vee* und Kollegen [VSAYAY09], *Liu* und Kollegen [LSC09] sowie *Stefanidis* und Kollegen [SDPP10].

Vee und Kollegen [VSAYAY09] untersuchen, wie diverse Suchergebnisse im Kontext eines Web-Shops effizient berechnet werden können. Items werden hierbei als Mengen von Features definiert und die jeweiligen Feature-Werte werden paarweise miteinander verglichen. Auch in dieser Diplomarbeit werden Items zwar als Feature-Mengen definiert, die Bildung einer diversen Submenge von Suchergebnissen geht jedoch über das paarweise Vergleichen von Feature-Werten hinaus.

Liu und Kollegen [LSC09] untersuchen Diversität nicht im eigentlichen Sinne. Viel mehr untersuchen sie das Problem, wie man für eine gegebene Menge von Suchergebnissen und eine Menge von Features, die die Suchergebnisse beschreiben, diejenigen Features auswählt, die die Unterschiede der Ergebnisse am besten widerspiegeln. Basierend auf den Features, anhand derer man die Daten am besten voneinander unterscheiden kann, können diverse Ergebnislisten generiert werden.

Stefanidis und Kollegen betrachten in [SDPP10] Diversität im Zusammenhang mit der Personalisierung von schlüsselwort-basiertem Suchen in relationalen Datenbanken. Genau genommen werden Diversität und die Überdeckung möglichst vieler Nutzerpräferenzen als ein Gegengewicht für die oftmals mit der Personalisierung einhergehenden Überspezialisierung von Suchergebnissen untersucht. Die Diversität von Tupelmengen wird auf Basis einer Distanz-Metrik (ähnlich zur *Jaccard* Ähnlichkeit/Distanz) berechnet. Hierfür wird zunächst die paarweise Distanz von Tupeln bestimmt und die Distanz der Tupelmengen ergibt sich aus dem Mittel der paarweisen Distanzen. Dabei gilt, je größer die Distanz, um so unterschiedlicher, also diverser, sind die betrachteten Tupelmengen.

Publikationen zur Diversity auf nicht-strukturierten Daten gibt es sehr viele. Da *Drosou* und *Pitoura* [DPP10] bereits eine sehr umfassende Übersicht bieten, sollen im Folgenden nur die besonders interessanten Ansätze vorgestellt werden.

Die Publikation von *Carbonell* und *Goldstein* [CGG98] reiht sich in die Klasse der Novalty-based Diversity Definitionen ein. Bei dieser Arbeit handelt es sich um eine der ersten Veröffentlichungen zum Thema Diversity und vermutlich um die erste Veröffentlichung überhaupt, die Novität mit Diversität verbindet. Der Kontext der Arbeit liegt im Bereich des *Text-Retrieval* und des *Summarization*. Sie definieren in der Arbeit das so genannte *Maximal Marginal Relevance* Kriterium (MMR). Das Kriterium zielt auf die Reduzierung von Redundanz in Suchergebnissen ab, bei gleichzeitiger Wahrung der Relevanz der Ergebnisse.

Ein ebenfalls sehr interessantes Anwendungsszenario für Diversity diskutieren *Smyth* und *McClave* [SMM01]. Sie untersuchen Diversität im Zusammenhang mit fall-basierten *Recommender Systemen* (CBRs). Fall-basierte Recommender Systeme schlagen Lösungen (Cases) für neue Probleme vor, in dem sie Lösungen für vorangegangene Probleme heranziehen. Oder anders gesagt: Für ein neues Problem wird nach Lösungen für bereits bekannte Probleme gesucht, die dem neuen Problem ähnlich sind. Dabei kann es vorkommen, dass sich alle gefundenen Lösungsansätze gleichen und alternative Lösungsansätze nicht weiter betrachtet werden. Eine Diversifizierung der gefundenen Lösungsansätze schafft hier Abhilfe. *Smyth* und *McClave* definieren eine Qualitätsmetrik, in der Diversität und Ähnlichkeit kombiniert werden. Diese Metrik muss maximiert werden, wenn ein

Case in die Ergebnismenge von Cases aufgenommen werden soll.

In der letzten Publikation, die hier im Zusammenhang mit Diversitätsmaßen erwähnt werden soll, beschäftigen sich *Gollapudi* und *Agrawal* [AGH+09] mit *Disambiguation* – einem mit Diversity verwandten Begriff. Im Prinzip geht es darum, dass Suchanfragen, beispielsweise in Web Suchmaschinen, in der Regel mehrdeutig sind. Sucht ein Nutzer beispielsweise nach *Jaguar*, weiß die Suchmaschine nicht, ob der Nutzer Informationen zu einem Sportwagen oder einer Raubkatze wünscht. Die meisten Nutzer suchen vermutlich nach dem Sportwagen. Damit jedoch die Nutzer, die Informationen zu der Raubkatze suchen, ebenfalls zufrieden gestellt werden, sollten zumindest einige wenige Ergebnisse auf der ersten Seite der Ergebnisliste die Raubkatze als Inhalt haben. Ähnlich wie *Clarke* und *Kollegen* [CKC+08] nutzen *Gollapudi* und *Agrawal* das Konzept der Information Nuggets, bezeichnen diese jedoch als *Kategorien*, mit dem Ziel, bevorzugt solche Dokumente in die Top-k Ergebnisliste aufzunehmen, die neue Informationen enthalten. Sowohl die Nutzeranfragen als auch die Dokumente sind modelliert als Mengen von Kategorien. Dokumente, die sehr eindeutig eine bestimmte Kategorie definieren, kommen in die Ergebnisliste, was wiederum bewirkt, dass andere Dokumente aus derselben Kategorie an Relevanz verlieren und damit die Wahrscheinlichkeit sinkt, dass sie in die Top-k Ergebnisliste aufgenommen werden.

2.2 Evaluationsstrategien

Seit geraumer Zeit werden für die Evaluierung von Retrieval-Systemen Verfahren und Metriken entwickelt, die die Effektivität der Systeme quantifizieren können und damit vergleichbar machen. Erst seit einigen wenigen Jahren jedoch wird für die Bestimmung der Effektivität auch die Diversität der Ressourcen betrachtet. Drei bekannte Evaluationsverfahren werden im Folgenden vorgestellt.

Für die Evaluierung von Suchergebnissen, mit dem Ziel die Effektivität von Retrieval Systemen zu messen, schlagen *Clarke und Kollegen* [CKC+08] eine Metrik vor, die sie $\alpha - NDCG$ genannt haben. Die Metrik basiert auf einem einfachen Nutzer-Modell, das davon ausgeht, dass in einer gerankten Ergebnisliste die relevantesten Dokumente in den oberen Rängen stehen sollen. Sie betrachten sowohl die Anfragen als auch die Dokumente als Mengen

von so genannten *Information Nuggets*. In anderen Veröffentlichungen entsprechen die Information Nuggets dem Konzept der Subtopics. Die Zugehörigkeit von Nuggets zu Dokumenten wird ermittelt durch binäre Nutzerentscheidungen, abgeschwächt durch einen Fehlerkoeffizienten α. Die Anzahl der Nuggets, die ein Dokument aufweist, bestimmt die graduelle Relevanz des Dokuments in der Ergebnisliste. Der *gain* eines Dokuments wird basierend auf seinem Rang in der Ergebnisliste sowie dem Vorhandensein redundanter Informationen bezüglich der Dokumente auf den vorangegangenen Rängen bestimmt.

Metzler und Chapelle [CMZ+09] entwickelten eine Evaluationsmetrik, die im Wesentlichen auf demselben Nutzer-Modell aufbaut wie die Metrik von *Clarke und Kollegen* [CKC+08]. D.h. die Relevanz, oder besser gesagt der *gain*, eines Dokuments auf Rang i hängt von den Dokumenten auf den vorangegangenen Rängen ab (*cascade-model*). Die Metrik nutzt graduelle Relevanz-Zuweisungen zu den Dokumenten und generiert für jeden Rang i in der Ergebnisliste einen Wahrscheinlichkeitswert, der die Erwartung dafür angibt, das ein Nutzer durch das Dokument auf Rang i zufrieden gestellt wird. Die Abwertung der Relevanz eines Dokuments erfolgt auf Basis der Anzahl der Dokumente, die der Nutzer zuvor betrachtet hat. Die Betrachtung der Diversität in der Ergebnisliste ist nur ein Randthema der Arbeit und der vorgeschlagene Ansatz ähnelt ebenfalls stark dem von Clarke und Kollegen. Es wird angenommen, dass einer Query eine Verteilung von Subtopics zugrunde liegt. Die Bewertung der Dokumente auf den einzelnen Rängen wird degradiert, wenn diese Subtopics aufweisen, die bereits von den Dokumenten auf den vorangegangenen Rängen abgedeckt werden.

Clarke und Kollegen [CKC+08] sowie *Metzler und Chapelle* [CMZ+09] gehen von der Annahme aus, dass alle Subtopics gleich relevant sind für die Query. Das bedeutet aber, dass unpopuläre Subtopics genauso behandelt werden, oder besser gesagt denselben Einfluss auf den gain eines Dokuments haben, wie sehr populäre Subtopics. Um diesen Nachteil auszugleichen entwickelten *Sakai und Kollegen* [SCS10] eine Metrik, die zum einen solche Ergebnislisten bevorzugt, deren Dokumente möglichst viele Subtopics abdecken. Zum anderen werden Dokumente besser bewertet, die hoch relevant sind für populäre Subtopics, als solche, die nur eine geringe Relevanz für nicht-populäre Subtopics aufweisen. Für beide Forderungen wird jeweils eine eigene klassische Metrik verwendet und in Form einer linearen Kombination zusammengebracht.

Für die erste Eigenschaft wird die altbekannte Metrik *S-recall* verwendet, die jedoch im Kontext der Arbeit als *I-recall* bezeichnet wird. Es handelt sich hierbei um eine einfache Metrik, die auf binären Relevanzbewertungen angewendet wird. *I-recall* ist im Wesentlichen die normalisierte Vereinigung der Mengen von Subtopics, die den Dokumenten der Ergebnisliste zugrunde liegen. Damit werden solche Ergebnislisten belohnt, die möglichst alle Subtopics der Query abdecken. Für die zweite Eigenschaft eignen sich prinzipiell alle möglichen Metriken, so lange sie auf graduellen Relevanzbewertungen angewendet werden können. *Sakai und Kollegen* schlagen NDCG (*Normalized Discounted Cummulative Gain*) vor, das bereits die Grundlage für die Metrik von *Clarke und Kollegen* [CKC+08] bildet. Es handelt sich hierbei um ein positions-basiertes Effektivitätsmaß, das relevante Dokumente auf hohen Rängen belohnt und auf niedrigen abstraft.

Die Entwickler der beiden erstgenannten Verfahren definieren eine Reihe von fragwürdigen Annahmen, wie beispielsweise die Unabhängigkeit oder die identische Relevanz der Subtopics. Diese Annahmen werden oftmals angezweifelt, so beispielsweise von den Entwicklern des zuletzt präsentierten Evaluationsverfahrens, was letztendlich zu einer einfachen Metrik führte, die zwar das Nutzerverhalten adäquat modelliert, dabei jedoch zweifelhafte Annahmen möglichst vermeidet.

Auffällig ist bei den drei präsentierten Verfahren, dass sie zum einen nicht die Diversität an sich betrachten und zum anderen einen speziell evaluierten Datensatz benötigen. In beiden Punkten unterscheidet sich das Evaluationsverfahren, das in dieser Arbeit definiert wird, deutlich von den drei präsentierten Verfahren. Alle Ansätze definieren die Diversität als einen Discount-Faktor, der Einfluss auf die Relevanz der Dokumente auf den einzelnen Rängen der Ergebnisliste hat. Die Relevanz einer Ergebnisliste für einen Nutzer hängt damit in gewisser Weise davon ab, wie redundant die Dokumente der Liste sind. Im Gegensatz dazu misst das Evaluationsverfahren aus dieser Arbeit die Diversität einer Menge von Ressourcen. Darüber hinaus benötigt das Evaluatinsverfahren keinen speziell evaluierten Datensatz. Das bedeutet, dass (graduelle) Relevanz-Bewertungen der Subtopics oder der Dokumente nicht notwendig sind. Einfache Tag-Zuweisungen zu einer Menge von Ressourcen reichen vollkommen aus. Derartige Datensätze sind bereits im großen Umfang vorhanden und müssen daher nicht erst mühevoll und kostspielig erzeugt werden.

3 Technische Grundlagen

In diesem Kapitel sollen die wesentlichen Grundlagen vorgestellt werden, die für das Verständnis dieser Arbeit erforderlich sind. Im Zentrum der Erläuterungen steht das Konzept der Entropie, das sowohl für das Diversitätsmaß als auch für das Evaluationsverfahren die Grundlage bildet. Für das Diversitätsmaß wird zusätzlich eine Sonderform der Entropie, die so genannte *bedingte Entropie*, vorgestellt. Für das Evaluationsverfahren ist eine Betrachtung des *Information Gain Ratio* notwendig.

3.1 Entropie

Der Begriff der Entropie hat seine Wurzeln in der Physik und konkret in der Thermodynamik und der statistischen Mechanik. Erst *Claude Shannon* brachte 1948 mit seiner Veröffentlichung *A Mathematical Theory of Communication* [Sha01] die Entropie aus dem Bereich der Physik in den der Informationstheorie und damit letzten Endes auch in die Informatik.

Oftmals wird die Entropie interpretiert als ein Maß für die Unordnung eines Systems, wobei diese diese Anschauung nicht immer zutreffend ist. In der Informationstheorie gilt die Entropie als ein Maß für den mittleren Informationsgehalt einer Nachricht, wobei der Begriff Information im statistischen und nicht im semantischen Sinne zu deuten ist. Insbesondere gilt jedoch, dass die Entropie einer Zufallsvariablen ein Maß für die Ungewissheit ist, die mit der Zufallsvariablen assoziiert ist. Ist die Ungewissheit darüber, welchen Wert die Zufallsvariable annehmen wird, groß, dann ist auch die Entropie der Zufallsvariablen

groß. Ist die Zufallsvariable gleichverteilt, dann kann die Zufallsvariable alle Werte mit gleicher Wahrscheinlichkeit annehmen und die Ungewissheit darüber, welchen Wert sie annimmt, ist maximal und damit auch die Entropie der Zufallsvariablen. Daher kann man auch sagen, dass die Entropie ein Maß für den Grad der Gleichverteilung der zugrunde liegenden Daten ist. In diesem Sinne ist die Entropie H einer diskreten Zufallsvariablen X mit einem endlichen Alphabet $V_X = \{x_1, ..., x_n\}$ mit den Wahrscheinlichkeiten $p_X(x_i)$ wie folgt definiert:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot log_2 p(x_i) \tag{3.1}$$

Auffällig ist die Verwendung des Logarithmus Dualis. Die Basis des Logarithmus gibt die Einheit der Informationen an. Da in der Informationstheorie in der Regel mit Bitstreams gearbeitet wird, hat sich der Logarithmus zur Basis 2 etabliert - prinzipiell ist jedoch jeder andere Wert für die Basis geeignet.

Im Folgenden sollen einige wichtige Anmerkungen und Eigenschaften der Entropie genannt werden:

1. Die Entropie ist nicht negativ. Dafür sorgt das negative Vorzeichen sowie die Tatsache, dass Wahrscheinlichkeiten zwischen 0 und 1 liegen und der Logarithmus zur Basis 2 für alle Werte aus dem Bereich $]0, \ldots, 1]$ negative Werte liefert.
2. Per Definition gilt: $log_2(0) = 0$. Damit wird zum einen sichergestellt, dass im Falle eines sicheren Ereignisses, d.h. $p(x_i) = 1$, die Entropie gleich 0 ist. Zum anderen wird dadurch verhindert, dass $log_2(0)$ ein undefiniertes Ergebnis annimmt.
3. Ist die zugrunde liegende Zufallsvariable gleichverteilt, dann ist die Entropie maximal.

3.2 Bedingte Entropie

Die bedingte Entropie ist ein Maß für die Ungewissheit über den Wert einer Zufallsvariablen X, der verbleibt, nachdem der Wert einer anderen Zufallsvariablen Y bekannt wird. In der Fachliteratur existieren mehrere leicht voneinander abweichende Definitionen der bedingten Entropie. Die Definition in dieser Arbeit lehnt sich an die Definition [1] an.

Sei eine diskrete Zufallsvariable X mit dem Wertevorrat $V_X = \{x_1, ..., x_n\}$ gegeben. Für jedes $i \in [1, ..., n]$ ist $P(X = x_i) \geq 0$ und es gilt $\sum_{i=1}^{n} P(X = x_i) = 1$. Sei weiterhin ein Ereignis A mit $P(A) > 0$ gegeben. Dann wird $H(X|A)$ definiert als die Entropie der bedingten Verteilung $P_{X|A}$:

$$H(X|A) = -\sum_{i=1}^{n} p_{X|A}(x_i) \cdot log_2 p_{X|A}(x_i) \tag{3.2}$$

Sei nun zusätzlich zur Zufallsvariablen X eine weitere Zufallsvariable Y gegeben. Analog zu X sei $V_Y = \{y_1, ..., y_m\}$ der Wertevorrat, den Y annehmen kann und es gelte für jedes $j \in [1, ..., m]$, dass die Wahrscheinlichkeit $P(Y = y_j) \geq 0$ und die Summe aller Einzelwahrscheinlichkeiten gleich eins ist. Für ein y_j ist die bedingte Entropie $H(X|Y = y_j)$ wie folgt definiert:

$$H(X|Y = y_j) = -\sum_{i=1}^{n} p_{X|Y}(x_i, y_j) \cdot log_2 p_{X|Y}(x_i, y_j) \tag{3.3}$$

Damit lässt sich die bedingte Entropie $H(X|Y)$ wie folgt definieren:

$$H(X|Y) = -\sum_{j=1}^{m} H(X|Y = y_j) \cdot p_Y(y_j) \tag{3.4}$$

1 *http://www.graphics.ethz.ch/teaching/former/infotheory0607/.../slides4_4.pdf*

Definition 3.4 betrachtet lediglich zwei Zufallsvariablen. Für das Diversitätsmaß, das in dieser Arbeit definiert wird, ist jedoch die Betrachtung der bedingten Entropie für mehr als zwei Zufallsvariablen notwendig. Aus diesem Grund wird im Folgenden die Adaption von 3.4 auf drei Zufallsvariablen vorgenommen – eine Betrachtung von mehr als drei Zufallsvariablen erfolgt analog. Sei daher zusätzlich zu den Zufallsvariablen X und Y eine dritte Zufallsvariable Z gegeben. Diese habe den Wertevorrat $V_Z = \{z_1, ..., z_l\}$ und sei darüber hinaus analog definiert wie X und Y. Dann ist die bedingte Entropie von X unter der Bedingung, dass Y und Z bekannt sind, wie folgt definiert:

$$H(X|YZ) = -\sum_{k=1}^{l} H(X|Y, Z = z_k) \cdot p_Z(z_k) \tag{3.5}$$

Dabei ist die Entropie $H(X|Y, Z)$ analog definiert zur Definition 3.3 . Im Folgenden sollen einige Eigenschaften der bedingten Entropie erwähnt werden:

1. Die bedingte Entropie $H(X|Y)$ der beiden Zufallsvariablen X und Y kann Werte zwischen 0 und $H(X)$ annehmen.
2. Die Entropie hat den Wert 0, wenn aus dem bekannten Y der Wert für X funktional bestimmt werden kann: $H(X|Y) \Leftrightarrow X = f(Y)$.
3. Die Entropie nimmt den Wert $H(X)$ an, falls die beiden Zufallsvariablen im stochastischen Sinne unabhängig voneinander sind.

3.3 Information Gain Ratio

Die Grundlage des Evaluationsverfahrens ist der C4,5 Algorithmus oder genauer gesagt die Metrik, die dieser verwendet, nämlich das *Normalized Information Gain* oder auch *Information Gain Ratio*. C4,5 ist im Bereich des maschinellen Lernens oder des *Data Mining* ein seit Langem bekanntes und häufig verwendetes Verfahren. Auf Basis eines klassifizierten

Trainingsdatensatzes baut der Algorithmus einen Entscheidungsbaum auf und erlaubt damit die Klassifizierung neuer Daten. Der Trainingsdatensatz besteht aus einer Menge von Attributen samt deren Ausprägungen. Eines der Attribute des Datensatzes muss ein kategorisches Attribut sein. Kategorische Attribute können beispielsweise Werte der Art $\{true, false\}$ oder $\{in, out\}$ annehmen.

Im Entscheidungsbaum entsprechen die Knoten den nicht-kategorischen Attributen des Trainingsdatensatzes und Kanten stellen mögliche Werte dar, die die Attribute annehmen können. Einträge des Datensatzes werden im Baum folglich als Pfade vom Wurzelknoten zu den Blättern dargestellt. Die Blätter spezifizieren wiederum die erwarteten Werte des kategorischen Attributes für die Einträge. Für jedes Attribut prüft der Algorithmus, wie sinnvoll dieses die Trainingsdaten auf bestimmte Klassen verteilt. Als Kriterium für die Qualität der Aufteilung dient das bereits erwähnte Normalized Information Gain, dessen Kern die Entropie darstellt. Im Folgenden soll der Normalized Information Gain, auch bezeichnet als Information Gain Ratio, genauer beschrieben werden.

Wie bereits erwähnt, ist der Kern des Information Gain Ratio die Entropie. Da die Entropie bereits weiter vorne (siehe Definition 3.1) erläutert wird, soll an dieser Stelle lediglich erwähnt werden, dass man die Entropie für eine gegebene Wahrscheinlichkeitsverteilung $P = (p_1, p_2, ..., p_n)$ auch als die Information interpretieren kann, die P enthält. Im Kontext des C4,5 Algorithmus wird die Entropie oftmals als $info$ bezeichnet und wie folgt interpretiert: Ist ein Trainingsdatensatz T, bestehend aus mehreren nicht-kategorischen Attributen sowie einem kategorischen Attribut, gegeben und ist der Datensatz auf Basis des kategorischen Attributes in disjunkte Klassen $C_1, C_2, ..., C_k$ unterteilt, dann ist die benötigte Information für die Identifikation der Klasse eines Elements aus T gerade $info$. Dabei wird $info$ auf der Wahrscheinlichkeitsverteilung der Klassen angewendet – diese kann wie folgt definiert werden:

$$P = (\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, ..., \frac{|C_k|}{|T|}) \tag{3.6}$$

Hat beispielsweise das kategorische Attribut lediglich die beiden Werte $\{true, false\}$, so

sieht P wie folgt aus:

$$P = (\frac{|\text{Einträge} = true|}{|\text{Einträge} = true \vee false|}, \frac{|\text{Einträge} = false|}{|\text{Einträge} = true \vee false|} \tag{3.7}$$

Bis jetzt wurde die Aufteilung des Datensatzes lediglich auf Basis des kategorischen Attributes betrachtet. Für die Definition des Information Gain muss der Datensatz jedoch zunächst auf Basis eines nicht-kategorischen Attributes X in die Mengen $T_1, T_2, ..., T_n$ partitioniert werden. In diesem Fall lässt sich die Information, die man benötigt, um die Klasse eines Elements aus T zu bestimmen, wie folgt definieren:

$$infoX(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \cdot info(T_i) \tag{3.8}$$

Der Information Gain (oder einfach nur *gain*) ist definiert als die Differenz der beiden soeben definierten Informationen. Das ist zum einen die Information zur Bestimmung der Klasse eines Elements aus dem Trainingsdatensatz und zum anderen die Information, die notwendig ist, um die Klasse eines Elements aus dem Trainingsdatensatz zu identifizieren, nachdem der Wert von X bestimmt wurde (wobei X ein nicht-kategorisches Attribut ist). Der Information Gain ist demnach wie folgt definiert:

$$gain(X) = info(T) - infoX(T) \tag{3.9}$$

Dabei gilt, dass die Attribute X, die die Einträge eindeutiger klassifizieren, einen höheren gain-Wert erhalten als solche, die für eine Klassifizierung viel zu diffus sind.

In Definition 3.9 wird der Information Gain in seiner natürlichen Form definiert. In dieser Form neigt er jedoch dazu, solche Attribute zu favorisieren, die viele Werte annehmen können. Die Ursache für diese Eigenschaft liegt in der Berechnung von $infoX$ und generell

in der Entropie. Der ungünstigste Fall, der eintreten kann, ist der, dass ein Attribut X für jeden Eintrag genau einen einmaligen Wert annehmen kann. D.h. X hat genauso viele Ausprägungen wie es Einträge im Testdatensatz gibt. In diesem Fall wäre aber $infoX = 0$ und $gain$ maximal (bzw. gleich $info(T)$) und das, obwohl das betrachtete Attribut X für eine Klassifizierung der Daten nicht geeignet ist.

Um die Favorisierung von Attributen mit vielen unterschiedlichen Werten vorzubeugen, wird der Information Gain normalisiert. Hierfür wird jedoch die Information benötigt, die sich durch die Aufteilung von T auf Basis des betrachteten Attributes X ergibt. In zahlreichen Publikationen (z.B. [Mor02]) wird diese Information als $splitinfo$ bezeichnet und sie ist wie folgt definiert:

$$splitinfo(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \cdot log_2(\frac{|T_i|}{|T|}) \tag{3.10}$$

Mit den Definitionen 3.9 und 3.10 kann nun die Definition des Information Gain Ratio ($gainratio$) erfolgen:

$$gainratio(X) = \frac{gain(X)}{splitinfo(X)} \tag{3.11}$$

4 Diversitätsmaß

Dieses Kapitel bildet das erste Teilthema der vorliegenden Diplomarbeit. In den folgenden Unterkapiteln werden zunächst die Grundlagen für das Diversitätsverfahren behandelt. Anschließend wird die Implementierung des Verfahrens, in Form eines Java-Frameworks, präsentiert. Den Schluss bildet ein Motivationsbeispiel, in dem die grundlegende Funktionsweise des Verfahrens demonstriert wird.

4.1 Ausgangssituation

In den einführenden Kapiteln wird gesagt, dass eine der Zielsetzung dieser Arbeit darin besteht, ein entopie-basiertes Diversitätsmaß auf strukturierten Daten zu entwickeln und dessen Eignung für den besagten Zweck, anhand eines Experiments und unter Verwendung eines neuartigen Evaluationsverfahrens, zu untersuchen. Strukturierte Daten stellen damit den Ausgangspunkt für die Definition des Diversitätsmaßes in dieser Arbeit dar. In dieser Arbeit wird Beispielhaft für die große Vielfalt der unterschiedlichen strukturierten Datensätze, der Datensatz von Bibsonomy verwendet.

Bei Bibsonomy handelt es sich um eine Plattform für die Verwaltung von Publikationen und Lesezeichen. Die Nutzer dieser Plattform können Publikationen oder Lesezeichen mit zusätzlichen Informationen, den so genannten Metainformationen, versehen und speichern. Im Falle der Publikationen können der Autor (oder die Autoren), das Erscheinungsjahr, der Herausgeber bzw. die veröffentlichende Institution, die Art der Publikation,

die Edition und einige weitere Informationen durch die Nutzer definiert werden. Besonders erwähnt werden sollen Tags, mit denen auch unterschiedliche Nutzer dieselbe Publikation annotieren können. Bei Tags handelt es sich um Schlagwörter, die oftmals einen beschreibenden Charakter haben und den Nutzern dabei helfen sollen, entweder bestimmte Publikationen schneller wiederfinden zu können oder eine Reihe ähnlicher Publikationen durch die Vergabe gleicher (oder ähnlicher) Tags thematisch zu gruppieren. D.h. Tags eignen sich für die Navigation und Organisation von Ressourcen.

Bibsonomy bietet unterschiedliche Suchmöglichkeiten an, darunter beispielsweise die Suche anhand von Autorennamen. Das bedeutet, dass man für einen spezifischen Autor als Ergebnis der Anfrage alle Publikationen dieses Autors erhält, die im Datensatz enthalten sind. Die Diversifizierung einer derartigen Menge kann auf zweierlei Weise erfolgen: Entweder man reichert diese Menge mit Publikationen aus einer anderen Quelle an oder aber man bildet eine diverse Submenge. In der vorliegenden Arbeit wird der zweitgenannte Ansatz realisiert.

Die Definition einer Diversitätsmetrik auf strukturierten Daten ist in mancher Hinsicht komplexer als auf nicht-strukturierten Daten, wie sie im Falle von Web Suchmaschinen vorliegen. Einer der Gründe hierfür ist der, dass man die unterschiedlichen Informationen (Metainformationen), die auch unabhängig voneinander sein können, in der Metrik vereint betrachten muss. Darüber hinaus steht man vor der Frage, welche der Metainformationen für die Diversifizierung geeignet sind und wie man diese kombiniert. *Vee* und Kollegen [VSAYAY09] kommen beispielsweise zu dem Schluss, dass bestimmte Attribute für die Diversifizierung von Items (oder allgemein Ressourcen) in einem Web Shop relevanter sind als andere. Als Konsequenz dieser Beobachtung definieren sie eine Ordnung der Attribute (oder der Metainformationen). Demnach werden die Items zunächst anhand des relevantesten Attributes diversifiziert, anschließend auf dem nächst-relevanten und so weiter. Dieser Weg wird in der vorliegenden Diplomarbeit nicht beschritten, stattdessen wird ein konkretes Anwendungsszenario samt der für das Szenario notwendigen Metainformationen definiert.

4.2 Anwendungsszenario für Diversität auf strukturierten Daten

Aufgrund der Vielfalt der Metainformationen, die Bibsonomy zur Verfügung stellt, sind zahlreiche unterschiedliche Szenarien denkbar. Auch denkbar ist die kombinierte Verwendung von Metainformationen aus Bibsonomy mit Metainformationen anderer Plattformen, wie beispielsweise *DBLP* [2]. Bei DBLP handelt es sich um ein Portal, das ganz analog zu Bibsonomy Informationen über Publikationen anbietet. Zwischen beiden Portalen gibt es im Wesentlichen zwei Unterschiede: Zum einen ist der DBLP Datensatz umfangreicher und zum anderen sind Attribute enthalten, die Bibsonomy nicht aufweist.

Da in der vorliegenden Arbeit das Thema Diversity viel mehr aus einem forschungstheoretischen Gesichtspunkt betrachtet wird, ist es nicht das Ziel, ein Anwendungsszenario zu definieren, das in einer realen Anwendung realisiert werden soll. Aus diesem Grund genügt ein einfaches, jedoch plausibles Szenario aus, um die Leistungsfähigkeit sowohl des Diversitätsmaßes als auch des Evaluationsverfahrens zu untersuchen. Ein für diesen Zweck geeignetes Szenario umfasst die drei *Attribute Tags*, *Autoren* und *Jahr* (Erscheinungsjahr). Tags sind, wie bereits erläutert, Schlagwörter, mit denen die Publikationen durch die Nutzer annotiert werden. Bei dem Attribut *Autoren* handelt es sich zumindest um einen Autor, jedoch in der Regel um mehrere Autoren, und das Erscheinungsjahr der jeweiligen Publikation wird durch das Attribut *Jahr* bestimmt.

Eine Diversifizierung der Publikationen anhand der Tags dürfte thematisch (zumindest in einigen Aspekten) diverse Publikationen liefern, während die Diversifizierung der Autoren u.U. unterschiedliche Blickwinkel auf ähnliche Themen liefert. Die Diversifizierung anhand der Erscheinungsjahre der Publikationen verhindert, dass ausschließlich aktuelle oder ausschließlich veraltete Publikationen in die diverse Ergebnisliste aufgenommen werden.

2 http://www.informatik.uni-Trier.de/ ley/db/

4.3 Entropie als Diversitätsmaß

Die Entropie kommt in zahlreichen Diversity-Ansätzen zum Einsatz und hat sich in diesen durchaus bewährt. Die wesentlichen Gründe hierfür, werden im Folgenden erläutert. Nach *Shannon* [Sha01] gibt die Entropie für eine Wahrscheinlichkeitsverteilung an, wie ähnlich diese hinsichtlich einer Gleichverteilung ist. Nimmt die der Wahrscheinlichkeitsverteilung zugrunde liegende Zufallsvariable jeden Wert mit gleicher Wahrscheinlichkeit an, dann ist die Entropie für die betrachtete Wahrscheinlichkeitsverteilung maximal. Dieser Sachverhalt lässt sich auch aus dem Blickwinkel der Diversität betrachten. Es gilt die begründete Annahme (vgl. [Del]), dass Dokumente beispielsweise durch Tags auf eindeutige Weise beschrieben werden. Oder anders ausgedrückt: Ein Tag kann den *latenten Topic*, also wesentlichen den Inhalt, eines Dokumentes widerspiegeln. Das lässt aber den Schluss zu, dass gemäß dem Fall, dass mehrere Dokumente mit gleichen Tags annotiert sind, die Dokumente einen ähnlichen Inhalt haben müssen – was wiederum bedeutet, dass diese Dokumente nicht divers sind. Eine Menge von Dokumenten, die mit Tags annotiert sind, lässt sich auch als eine Wahrscheinlichkeitsverteilung auffassen. Das kann man sich wie folgt vorstellen: Ist beispielsweise eine Menge von fünf Dokumenten gegeben, die alle mit einem unterschiedlichen Tag annotiert sind, dann hat die zugrunde liegende Wahrscheinlichkeitsverteilung die Gestalt:

$$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}) \tag{4.1}$$

Durch Anwendung der Definition für die einfache Entropie aus Kapitel 3 wird für diese Verteilung ein Entropiewert von 2.321 bestimmt, der zugleich die maximale Entropie für diese (Gleich-)Verteilung darstellt. Nimmt man weiterhin an, dass es eine zweite Menge von fünf Dokumenten gibt, von denen jedoch drei Dokumente identisch und die beiden restlichen untereinander und bezüglich der drei vorangegangenen unterschiedlich getaggt sind, ergibt sich die folgende Verteilung:

$$(\frac{3}{5}, \frac{1}{5}, \frac{1}{5}) \tag{4.2}$$

Für diese Verteilung wird lediglich ein Entropie-Wert von 1.370 bestimmt. Im direkten Vergleich ist die erstgenannte Menge von Dokumenten divers, während die zweite Menge drei gleiche oder ähnliche Dokumente enthält. Die beiden resultierenden Entropiewerte spiegeln diesen Sachverhalt entsprechend wider.

Das Beispiel, das so eben lediglich für Tags gezeigt wurde, lässt sich prinzipielle auf die unterschiedlichsten Attribute von Ressourcen übertragen.

4.4 Bedingte Entropie als Diversitätsmaß

Im vorangegangenen Unterkapitel wird an einem einfachen Beispiel gezeigt, dass die Entropie, angewendet auf einer Verteilung der Dokumente, prinzipiell eine Aussage hinsichtlich der Diversität der Dokumente ermöglicht. Dabei wird die Verteilung basierend auf einem den Dokumenten zugrunde liegenden Attribut erzeugt. Diese Beobachtung stellt die Grundlage für das Diversitätsmaß dar, das im nächsten Unterkapitel definiert wird. Doch bevor die eigentliche Definition erfolgen kann, muss diskutiert werden, wie die Diversifizierung von Publikationen gleichzeitig über mehrere Attribute erfolgen kann.

Prinzipiell lassen sich mit der im vorangegangenen Unterkapitel gezeigten Vorgehensweise auch mehrere Attribute für die Diversifizierung betrachten. Man könnte beispielsweise für jedes Attribut der betrachteten Menge von Publikationen die Entropie einzeln bestimmen und die Werte aggregieren. Der Nachteil dieses Ansatzes besteht jedoch darin, dass ein hoher Wert für ein Attribut sehr niedrige Werte der restlichen Attribute ausgleichen kann. Das bedeutet, dass zwei unterschiedlich diverse Mengen von Items unter Umständen denselben aggregierten Entropie-Wert erhalten. Um diesem Problem zu begegnen, ist es sinnvoll, alle Attribute des Anwendungsszenarios gleichzeitig oder besser gesagt bedingt

zu betrachten. Einen klaren und natürlichen Ansatz für die gemeinsame Betrachtung aller Attribute bietet die bedingte Entropie.

Die Definition der bedingten Entropie ist in Kapitel 3 zu finden und soll daher an dieser Stelle nicht nochmals aufgeführt werden. Wichtig zu verstehen ist, dass sich die Entropie mehrerer diskreter Zufallsvariablen, als Summe von bedingten Entropien bestimmen lässt. Die gemeinsame Entropie dreier Zufallsvariablen (X, Y, Z) lässt sich wie folgt bestimmen:

$$H(XYZ) = H(X) + H(Y|X) + H(Z|XY) \tag{4.3}$$

Hierbei ist die Entropie $H(Y|X)$ zu interpretieren als die Unsicherheit die über Y bleibt, wenn X bereits bekannt ist (analog für $H(Z|XY)$). Das bedeutet, die totale Unsicherheit, die über die Werte von X, Y und Z besteht, ist gleich der Summe der Unsicherheiten bezüglich der Werte von X, der durchschnittlichen Unsicherheit Werte von Y sobald X bekannt ist und der durchschnittlichen Unsicherheit der Werte von Z sobald X und Y bekannt sind.

Analog zu dem gezeigten Beispiel mit drei Zufallsvariablen, lassen sich unter Anwendung der Kettenregel für die Entropie (bzw. für die gemeinsame Entropie) beliebig viele Zufallsvariablen betrachten. Anzumerken bleibt, dass die Reihenfolge, in der man die Zufallsvariablen betrachtet, irrelevant ist für die Bestimmung der Unsicherheit, die bezüglich X, Y und Z besteht. Das heißt, dass unterschiedliche Abspaltungen im Ergebnis identisch sind:

$$\begin{aligned} H(XYZ) &= H(X) + H(Y|X) + H(Z|XY) \\ H(XYZ) &= H(X) + H(Z|X) + H(Y|XZ) \\ &\vdots \\ H(XYZ) &= H(Z) + H(Y|Z) + H(X|YZ) \end{aligned} \tag{4.4}$$

Die Reihenfolge der Abspaltungen ist irrelevant, da sie die gesamte Unsicherheit der drei Zufallsvariablen, stets durch die Summe der drei einzelnen Unsicherheiten (Restunsicherheiten) ergibt.

Zusammenfassend lässt sich feststellen, dass mit Hilfe der bedingten Entropie, die gemeinsame Entropie prinzipiell beliebig vieler Zufallsvariablen bestimmt werden kann. Da die Entropie von Grund auf ein Maß für Diversität darstellt, bietet die Verwendung der bedingten Entropie ein natürliches Mittel zur Bestimmung der Diversität einer Menge von Ressourcen, anhand der gleichzeitigen Betrachtung mehrerer Attribute der Ressourcen.

4.5 Definition des Diversitätsmaßes

Die Bestimmung einer diversen Submenge aus einer großen Menge von Items wird im Kontext dieser Arbeit als ein Optimierungsproblem behandelt. Das bedeutet, dass für die Bestimmung der diversesten Submenge, oder auch Kombination, alle möglichen Kombinationen sukzessive betrachtet werden müssen. Für jede Kombination wird die bedingte Entropie anhand der Attribute, die durch das Anwendungsszenario vorgegeben werden, bestimmt und über alle Kombinationen hinweg maximiert. Die Kombination mit dem höchsten Entropie-Wert gilt als die diverseste unter allen Kombinationen. Mit dieser Anmerkung sowie den Erläuterungen der beiden vorangegangenen Unterkapiteln kann im Folgenden die Definition des Diversitätsmaßes erfolgen. Tabelle 4.1 zeigt die wesentlichen Symbole und deren Interpretation, die für die Definition des Maßes zum Einsatz kommen.

Sei RES die Menge der Tupel, die man für eine Anfrage Q an eine Relation R erhält. Das Schema von R sei definiert durch die Attribute $A = \{A_1, ..., A_n\}$ und $Z = \{Z_1, ..., Z_m\}$ sei die Menge der Attribute, die durch das Anwendungsszenario vorgegeben werden, wobei $Z \subseteq A$ gilt. Ist M die Menge aller möglicher Kombinationen $\{M_1, ..., M_l\}$ von Tupeln $t \in RES$, dann gilt für eine Kombination $S \in M$:

Symbol	Interpretation
Q	Eine Query
R	Eine Relation mit dem vollständigen Bibsonomy Schema
RES	Eine Menge von Tupeln, die man für eine Anfrage Q an die Relation R erhält
$A = \{A_1, ..., A_n\}$	Menge der Schema-Attribute von R
$Z = \{Z_1, ..., Z_m\}$	Menge von Attributen, die durch das Anwendungsszenario definiert wird
M	Eine Menge von Kombinationen von Tupeln $t \in RES$
$H(Z)$	Die Entropie für eine Kombination von Tupeln über die Attribute aus Z

Tabelle 4.1: Die für die Definition des Diversitätsmaßes verwendeten Symbole sowie deren Interpretation.

$$S \text{ ist divers} \Leftrightarrow \forall L \in M : |L| = |S| \Rightarrow H_S(Z) \geq H_L(Z) \tag{4.5}$$

Dabei ist $H_S(Z)$ die bedingte Entropie für die Tupel-Kombination S und $H_L(Z)$ analog die Entropie aller restlichen Kombinationen aus M.

4.6 Motivationsbeispiel für das Diversitätsmaß

Im Folgenden wird auf einem kleinen Beispieldatensatz, der dem Anwendungsszenario genügt, die Funktionsweise des Diversitätsmaßes präsentiert. Das bedeutet konkret, dass für eine fiktive Anfrage fünf Tupel vom System zurückgegeben werden und das Schema der Tupel entspricht dem Schema, das durch das Anwendungsszenario definiert wird. Das fiktive Ergebnis oder besser gesagt die Ergebnisliste der Anfrage ist Tabelle 4.2 zu entnehmen.

Tupel	Autoren	Tags	Jahr
A	Staab, Sizov	semantic, ontology	2004
B	Staab, Sizov, Schulz	web, language	2004
C	Staab	ontology	2004
D	Staab, Saathoff	p2p	2010
E	Staab	ontology	2000

Tabelle 4.2: Stark vereinfachte Ergebnisliste für eine virtuelle Anfrage bzw. Suche nach den Publikationen von *Steffen Staab*. Die Tabelle zeigt die Attributwerte entsprechend dem Schema, das durch das Anwenundungsszenario definiert wird. Von links aus gesehen, sieht man zuerst den Tupelbezeichner, anschließend die Autoren sowie die Tags, mit denen die Publikation assoziiert ist. In der letzten Spalte steht jeweils das Erscheinungsjahr der Publikation.

Um die Übersichtlichkeit dieses einfachen Beispiels zu gewährleisten, wird bewusst diese kleine Ergebnisliste verwendet. Das Ziel besteht darin, aus den fünf Tupeln die diverseste Kombination von drei Tupeln zu bestimmen. Augenscheinlich bedeutet das, dass man alle möglichen Kombinationen der Größe 3 aus den fünf Tupeln $A, ..., E$ bilden muss, dem ist jedoch nicht so. Zum einen darf ein Tupel nicht häufiger als einmal in einer Kombination vertreten sein, da eine Wiederholung von Grund auf die Forderung nach Diversität konterkarieren würde und zum anderen ist die Reihenfolge der Tupel in der Kombination irrelevant. Mit diesen beiden Forderungen kann die Anzahl der zu betrachtenden Kombinationen wie folgt bestimmt werden:

$$\binom{n}{k} = \frac{n!}{(n-k)! * k!} \tag{4.6}$$

Hierbei bezeichnet n die Anzahl der vorhandenen und k die Anzahl der ausgewählten (d.h. die Größe einer Kombination von Elementen) Elemente. Für das aktuelle Beispiel mit $n = 5$ und $k = 3$ bedeutet das, dass 10 Kombinationen zu betrachten sind. Gesucht wird die Kombination, für die die Diversitätsmetrik im Vergleich zu allen anderen Kombinationen maximal ist. Eine derartige diverse Kombination dürfte genau dann vorliegen, wenn sich die Tupel der Kombination möglichst stark voneinander unterscheiden. Man könnte

auch sagen: Jedes Tupel sollte möglichst viele neue Informationen und möglichst wenige bereits vorhandene Informationen in die Kombination einbringen. Betrachtet man die fünf Tupel unter diesem Gesichtspunkt, dann ist offensichtlich, dass die Tupel C und E nicht in der diversesten Kombination gemeinsam vertreten sein dürfen, da sie sich lediglich im Erscheinungsjahr unterscheiden. Darüber hinaus sollten die Tupel A und B in der diversesten Kombination gemeinsam vertreten sein, da sie abgesehen von Tag *p2p* alle Tags und abgesehen vom Autor *Saathoff* alle Autoren abdecken. Das Paar (A, B) kann im Grunde nur noch durch das Einfügen von Tupel D mit neuem Wissen angereichert und damit diverser gemacht werden. Zusammenfassend bedeutet das, dass die diversesten Kombinationen die Tupel A und B enthalten sollten und die am wenigsten diversen die Tupel C und E.

Tabelle 4.3 zeigt die Ergebnisse der Anwendung des Diversitätsmaßes auf den fünf Tupeln. Links in der Tabelle sieht man die jeweilige Kombination und rechts steht der Wert, den man durch Anwendung von Definition 4.5 für die jeweilige Kombination erhält.

Der höchste Wert liegt bei 9.088 und wird durch die Kombination der Tupel A, B und D erreicht. Der niedrigste Wert liegt lediglich bei 2.584 und wird für die Kombination (CDE) ermittelt. Mit einer einzigen Ausnahme bestätigen die Ergebnisse die anfangs getätigte Annahme, dass Kombinationen mit den Tupeln A und B eine hohe Diversität und Kombinationen mit C und E eine geringe Diversität aufweisen. Die besagte Ausnahme bildet die Kombination (BCE), die mit einem Wert von 5.673 im Mittelfeld liegt, jedoch immer noch weit von der maximal diversen Kombination (ABD) entfernt ist. Die Ursache hierfür liegt sehr wahrscheinlich in der Anwesenheit von Tupel B, welches mit drei Autoren und zwei Tags eine dominierende Wirkung hat und damit die Diversität der Kombination als Ganzes deutlich steigert. Tupel D hat hingegen nicht dieselbe dominierende Wirkung wie B, was in dem sehr niedrigen Wert für (CDE) resultiert.

Kombination	Diversitätswert
ABC	7.010
ABD	9.088
ABE	8.088
ACD	4.754
ACE	3.754
ADE	5.584
BCD	6.673
BCE	5.673
BDE	6.754
CDE	2.584

Tabelle 4.3: Liste aller relevanten Kombinationen sowie des für jede Kombination ermittelten Diversitätswertes nach Definition 4.5. Je größer der Wert, umso diverser sind die Tupel der Kombination.

4.7 Implementierung

Im Folgenden wird die Implementierung des Frameworks vorgestellt, welches das Diversitätsmaß nutzt (siehe Definition 4.5) um für eine Anfrage und die resultierende Ergebnisliste eine diverse Submenge vorgegebener Größe zu bestimmen. Implementiert ist das Framework in der Programmiersprache JAVA und als Datenbank kommt Oracle 11g zum Einsatz. Beim Testdatensatz handelt es sich um die zum Zeitpunkt der Entstehung dieser Arbeit aktuellste Version des frei zugänglichen Bibsonomy Datensatzes [3] , der im gepackten Zustand eine Größe von 188MB hat und dabei ca. 650000 Publikationen (vgl. [BHJ+10]) samt zugehöriger Metainformationen umfasst. Alle Berechnungen werden auf einem SUSE 11.3 Linux Server mit 16 Kernen und 70GB Arbeitsspeicher durchgeführt.

In der aktuellsten Version ermöglicht das Framework die Suche nach Publikationen anhand von Autoren. Es handelt sich hierbei um eine Funktionalität, die Bibsonomy von Haus aus zur Verfügung stellt, der Unterschied ist jedoch der, dass das Framework die Ergebnisse nicht einfach nach Aktualität sortiert, sondern eine diverse Kombination von Ergebnissen zurück gibt. Die Größe dieser Kombination kann durch die Nutzer vorgegeben werden. Die einzige Beschränkung für die Größe der Kombination ist im Prinzip die Größe der Ergebnisliste, denn wie weiter vorne bereits erläutert wird, macht es im Hinblick auf Diversität wenig Sinn, Kombinationen zu betrachten, die genauso groß sind wie die Ergebnisliste.

Nachdem ein Nutzer eine Suchanfrage eingegeben hat, wird die Anfrage via SQL-Query an die Datenbank geschickt. Da das Framework in der aktuellen Version lediglich als ein Testwerkzeug fungiert, werden nur die Attribute erfragt, die durch das weiter vorne definierte Anwendungsszenario vorgegeben werden. Diese Einschränkung reduziert zwar die Funktionalität des Frameworks erheblich, jedoch ist es nicht das Ziel dieser Arbeit eine voll funktionierende Anwendung zu entwickeln, sondern zum einen ein Diversitätsmaß und zum anderen ein Evaluationsmaß für Diversitätsmaße (oder Diversitätsverfahren) zu definieren und deren Leistungsfähigkeit auf realen Daten zu untersuchen.

Sobald die Ergebnisliste für die Anfrage vorliegt, werden die Daten aufbereitet. Das ist

3 http://www.kde.cs.uni-kassel.de/bibsonomy/dumps

notwendig, da für die Berechnung der Diversitätswerte der einzelnen Kombinationen von Tupeln die Attributwerte der Tupel miteinander vergleichbar sein müssen. In der Regel tragen aber unterschiedliche Nutzer die Daten auf unterschiedliche Weise ein, sodass ein direkter Vergleich zunächst nicht möglich ist. Für das Anwendungsszenario bedeutet das beispielsweise, dass die einzelnen Autoren jeder Publikation bestimmt werden müssen. In der Regel hat eine Publikation mehrere Autoren, die als ein langer String vorliegen und die einzelnen Namen im String sind durch Kommata oder Bindewörter separiert. Ein direkter Vergleich derartiger Strings ist nicht sinnvoll, da die Namen in unterschiedlicher Reihenfolge vorliegen können oder aber die Vornamen entweder voll ausgeschrieben oder in abgekürzter Form vorliegen. Um derartige Einträge doch vergleichbar zu machen, werden die einzelnen Nachnamen extrahiert und in einer speziellen Datenstruktur abgelegt.

Nach der Aufbereitung der Daten beginnt die Berechnung der Kombinationen von Publikationen und der den Kombinationen zugrunde liegenden Diversitätswerte. Wie bereits im vorangegangenen Kapitel erläutert, kann die Anzahl der zu berechnenden Kombinationen sehr groß sein. Die Zahl der Kombinationen hängt direkt von der Größe der Kombinationen sowie der Größe der Ergebnisliste ab. Liegt beispielsweise eine Ergebnisliste von 50 Publikationen vor und soll die diverseste Kombination der Größe 5 bestimmt werden, sind bereits über 2 Millionen Kombinationen zu berechnen. Bei 10 aus 50 fallen mehr als 10 Milliarden Kombinationen an, aus denen die Kombination mit dem maximalen Entropiewert bestimmt werden muss. Eine sequentielle Berechnung aller Kombinationen und der zugehörigen Diversitätswerte würde bei dieser Anzahl an Kombinationen enorm viel Zeit in Anspruch nehmen und ist folglich nicht hinnehmbar. *Vee* und Kollegen [VSAYAY09] stellen ein Verfahren vor, das in Echtzeit die diverseste Menge von Items bestimmen kann. Für eine reale Anwendung des in dieser Arbeit vorgestellten Diversitätsmaßes, beispielsweise als Erweiterung der Funktionalität des Bibsonomy Portals, wäre der Ansatz von Vee durchaus überlegenswert, für den eher theoretischen Charakter dieser Arbeit führt er jedoch zu weit. Aus diesem Grund kommt in dieser Arbeit eine Kombination aus Parallelisierung und *Brute-Force* zum Einsatz. Mit *Brute-Force* ist die stupide Berechnung aller relevanter Kombinationen gemeint, wobei dieser Ansatz in der Regel leistungsstarke Hardware voraussetzt, die jedoch durch den anfangs beschriebenen Server gegeben ist. Parallelisierung wird realisiert durch die Multithreading Funktionalität, die JAVA von Haus aus zur Verfügung stellt.

Abbildung 4.1 zeigt den konzeptuellen Aufbau des Frameworks. Die Klasse *Main* enthält die Steuerungslogik und ist im Wesentlichen für das Starten der Threads und die Ausgabe des Ergebnisses zuständig. Bevor die Threads gestartet werden, wird jedoch eine Instanz der Klasse *DBInterface* aufgerufen, die via *JDBC* Schnittstelle eine Verbindung zur lokalen Oracle Datenbank herstellt und eine Anfrage an die Datenbank schickt. Jede Publikation wird als ein Objekt vom Typ *Entry* in einer Collection abgespeichert und anschließend zurückgegeben an die aufrufende Methode in *Main*. Nachdem die Ergebnisliste in Form einer Menge von *Entry*-Objekten erstellt wurde, werden in *Main* eine festgelegte Anzahl von Threads gestartet. Jeder Thread bekommt bei seinem Start eine Referenz auf ein Objekt vom Typ *Runnable*, wobei es sich hierbei um eine Java-Schnittstelle handelt, die im Framework von der Klasse *myRunnable* implementiert wird.

Die Threads rufen eine Instanz der Klasse *CombinationFinder* auf, um eine valide und noch nicht bearbeitete Kombination zu bestimmen. Passend zur Kombination holt sich jeder Thread die Daten, d.h. die Metainformationen der zur Kombination gehörenden Publikationen, aus der Collection und formatiert diese durch Aufruf einer Instanz der Klasse *DataFormater*. Für die Berechnung der bedingten Entropie einer Kombination sind zwei Klassen zuständig: Das ist zum einen die Klasse *DistributionsCalculater* und zum anderen *EntropyCalculater*. Die erstgenannte der beiden Klassen ist für die Generierung der bedingten Verteilungen der Zufallsvariablen zuständig. Bedingt durch das Anwendungsszenario werden die Verteilungen P_X, $P_{Y|Z}$, sowie $P_{X|YZ}$ generiert, wobei X dem Attribut *Jahre*, Y dem Attribut *Tags* und Z dem Attribut *Autoren* entsprechen. Basierend auf diesen Verteilungen, die für jede Kombination generiert werden, berechnet *EntropyCalculater* die bedingte Entropie der zugrunde liegenden Kombination.

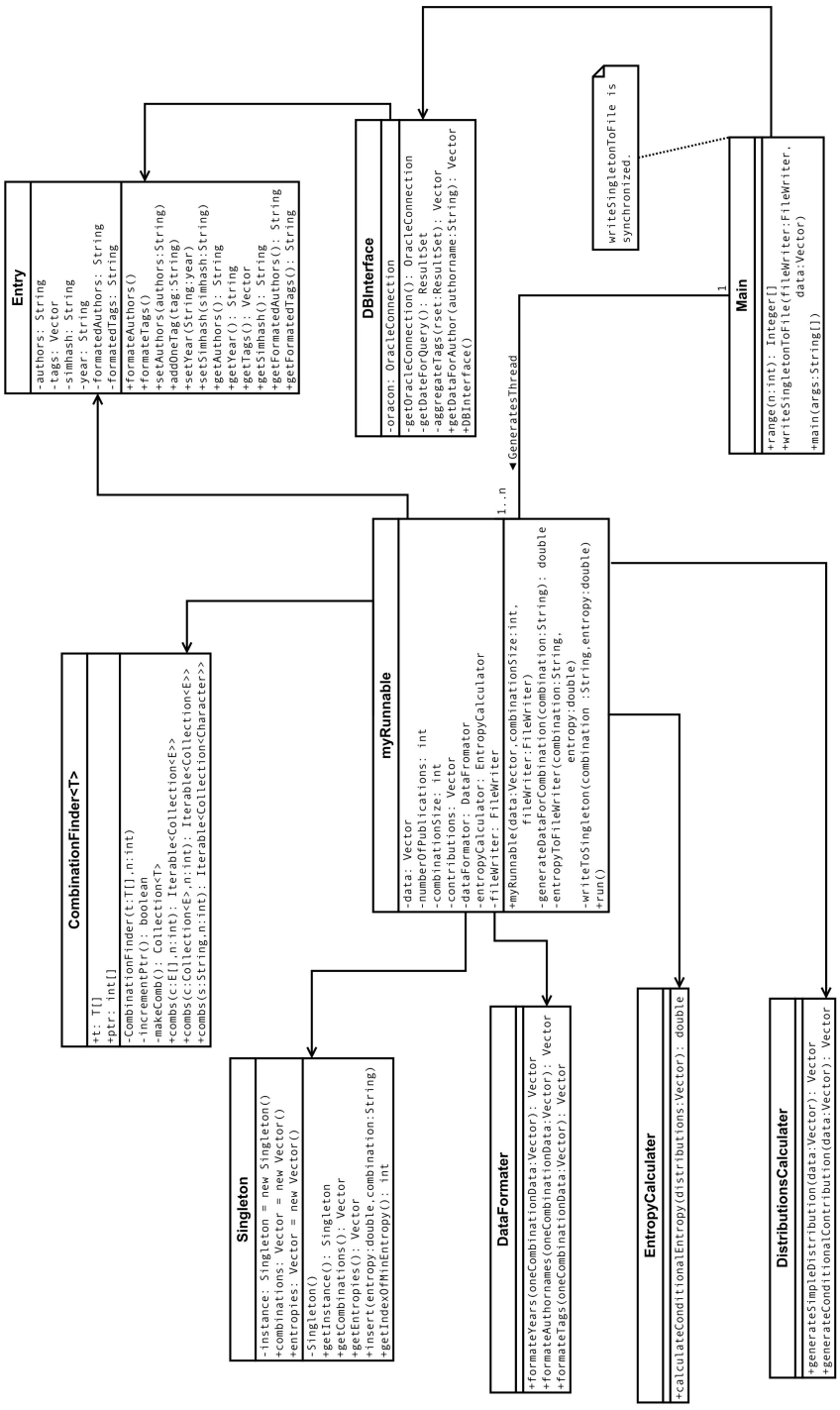

Abbildung 4.1: UML-Klassendiagramm. Konzeptuelle Beschreibung des Aufbaus des Frameworks. Zu sehen sind die wesentlichen Komponenten des Frameworks sowie deren Beziehung untereinander.

5 Evaluationsverfahren

Im Folgenden wird das zweite und wesentliche Kernthema der vorliegenden Arbeit behandelt: Die Definition eines neuartigen Verfahrens für die Evaluierung von Diversitätsmetriken. Zum Zeitpunkt der Entstehung dieser Diplomarbeit existieren zwar einige Ansätze für die Evaluierung von Diversitätsmetriken (im Bereich des Webs), jedoch quantifiziert keines davon die Diversität an sich. In der Regel wird lediglich versucht, insbesondere durch die Optimierung von *Precision*-Werten, die Nutzerzufriedenheit zu maximieren. Das bedeutet, dass Diversität gleichgesetzt wird mit Nutzerzufriedenheit, was aber nicht unbedingt korrekt ist. Diverse Suchergebnisse beeinflussen sicherlich die Nutzerzufriedenheit, schließlich möchten die meisten Nutzer keine fast-identischen Suchergebnisse erhalten. In die umgekehrte Richtung lässt sich jedoch nicht, einzig aufgrund der Zufriedenheit der Nutzer, eine quantitative Aussage hinsichtlich der Diversität der Ergebnisse herleiten.

Das Evaluationsverfahren kann entweder vorhandene Tag-Zuweisungen oder in Nutzerexperimenten durchgeführte Tag-Zuweisungen zu Ressourcen nutzen, um die Diversität von Mengen von Ressourcen zu messen. Auf diese Weise werden die Ergebnisse unterschiedlicher Diversitätsmetriken/Diversitätsverfahren auf direkte Weise vergleichbar gemacht. Doch bevor Tag-Zuweisungen hierfür überhaupt verwendet werden können, muss gezeigt werden, dass Tags die wesentlichen Inhalte von Ressourcen widerspiegeln und sich für die thematische Gruppierung der Ressourcen eignen. Das geschieht im nachfolgenden Unterkapitel. Im Anschluss daran wird eine Metrik definiert, die anhand der Tag-Zuweisungen zu einer Menge von Ressourcen einen Wert generieren kann, der ein Maß für die Diversität der Menge darstellt. Die Metrik ist eine Adaption des C4,5 Verfahrens, dessen Grundlage der Information Gain Ratio darstellt, der bereits in Kapitel 3 definiert wurde. Nach der Definition der Metrik wird die Funktionsweise des Evaluationsverfahrens anhand eines Beispiels auf einfachen Daten demonstriert.

5.1 Grundlagen des Evaluationsverfahrens

Die wesentliche Aufgabe dieser Arbeit liegt in der Definition eines Verfahrens, mit dem unterschiedliche Diversitätsverfahren bzw. Diversitätsmaße evaluiert werden können. Wie bereits mehrfach angedeutet wurde, werden in dem Verfahren Tag-Zuweisungen zu einer Menge von Ressourcen genutzt, um die Diversität dieser Menge zu messen.

Die theoretischen Grundlagen für das Verfahren liefert eine noch nicht veröffentlichte Arbeit von *Klaas Dellschaft* [Del] - die Arbeit wird im Folgenden vorgestellt. Das in dem Verfahren zum Einsatz kommende Maß wird im nachfolgenden Unterkapitel definiert.

5.1.1 Verwandte Arbeit

In seiner Arbeit untersucht *Klaas Dellschaft* [Del] den Einfluss unterschiedlicher Arten von Tagvorschlägen auf die Eigenschaft von *Tagging-Systemen*, Ressourcen zu organisieren und zu indexieren. Diese Eigenschaft definiert er als Maß der Korrelation zwischen der von den Nutzern der Tagging-Systeme wahrgenommenen Ähnlichkeit der Ressourcen mit der Ähnlichkeit der Tag-Zuweisungen der Nutzer zu diesen Ressourcen.

Um den Einfluss unterschiedlicher Arten von Tagvorschlägen auf dieses Maß zu untersuchen, führt er ein 2-stufiges web-basiertes Nutzerexperiment mit über 800 Testpersonen durch. In der ersten Stufe des Experiments wurden drei Gruppen von Testpersonen 10 Screenshots von Webseiten präsentiert. Die erste Gruppe bekam keinerlei Tagvorschläge, der zweiten Gruppe wurden populäre Tags vorgeschlagen, während der dritten Gruppe nur die Tags angezeigt wurden, die die jeweilige Testperson im Laufe des Experiments bereits einmal verwendet hatte. In der zweiten Stufe des Experiments sollten die Testpersonen aller drei Gruppen die ihnen präsentierten Screenshots der 10 Webseiten anhand der empfundenen Ähnlichkeit der Webseiten klassifizieren.

In dem Experiment wurden folglich zwei unterschiedliche Arten von Informationen generiert: Auf der einen Seiten stehen die Tag-Zuweisungen zu den Ressourcen in Form von Tagvektoren. Für jede Ressource enthält ein Tagvektor die Anzahl der Tags, die der

Ressource durch die Testpersonen zugeordnet wurden. Auf der anderen Seite liefert das Experiment eine Verteilung (Clustering) der Ressourcen, basierend auf deren Ähnlichkeit (jede der drei Gruppen erzeugt ein derartiges Clustering).

Wie anfangs bereits gesagt wird, will *Dellschaft* die Korrelation zwischen der von den Nutzern wahrgenommenen Ähnlichkeit von Ressourcen mit der Ähnlichkeit der Tag-Zuweisungen zu diesen Ressourcen messen. Dabei gilt, dass je größer die Korrelation ist, desto besser können Ressourcen in einem Tagging-System organisiert und indexiert werden. Gemessen wird die Korrelation mit dem *Silhouette Coefficient* (siehe Definition 5.1). Hierbei handelt es sich um eine Funktion, die bei der Evaluierung von Clustering-Algorithmen zum Einsatz kommt.

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{5.1}$$

Der Silhouette Coefficient s_i wird definiert für ein Objekt i aus dem gegebenen Clustering. a_i ist die durchschnittliche Distanz von Objekt i zu allen anderen Objekten des Clusters von i. Für die Bestimmung von b_i werden alle Cluster betrachtet, die das Objekt i nicht enthalten. Für jedes dieser Cluster wird die durchschnittliche Distanz von i zu den Objekten des jeweiligen Clusters ermittelt. Die minimale dieser Distanzen ist gerade b_i.

Im klassischen Sinne wird der Silhouette Coefficient genutzt, um zu bestimmen, wie gut ein Clustering die zugrunde liegenden Daten widerspiegelt. Für ein gegebenes Clustering von Objekten sowie eine beliebige Distanzfunktion, mit der der Abstand zwischen je zwei Objekten gemessen werden kann, kann mit dem Silhouette Coefficient die Dichte der Objekte in den Clustern des Clusterings sowie die Eindeutigkeit der Cluster bestimmt werden.

Dellschaft modifiziert die klassische Anwendung in die Richtung, dass er nicht zwei unterschiedliche Clusterings vergleicht, sondern zwei unterschiedliche Mengen von Tagvektoren (jede Menge von Tagvektoren entspricht den Tag-Zuweisungen einer der drei Gruppen), während das Clustering als fix betrachtet wird. Für die Berechnung der Distanz

zwischen je zwei Tagvektoren verwendet er das im Bereich des Information Retrieval weit verbreitete *Kosinusmaß*:

$$cosim(v_i, v_j) = cos\theta = \frac{v_i \cdot v_j}{\parallel v_i \parallel \cdot \parallel v_j \parallel} \tag{5.2}$$

Hierbei sind v_i und v_j zwei Tagvektoren und θ der Winkel, den die beiden Vektoren einschließen. Ist der Kosinus von θ gleich 0, dann sind die beiden Vektoren identisch. Bei einem Kosinus von 90 Grad sind die Vektoren orthogonal zu einander.

Für je zwei Mengen von Tagvektoren gilt, dass die Menge mit dem größeren s_i-Wert konsistenter bezüglich des Clusterings ist (d.h. konsistenter hinsichtlich der Gruppierung der 10 Webseiten).

Als die wichtigste Vorbedingung für die Anwendung des Silhouette Coefficient nennt *Dellschaft* die durchschnittliche Ähnlichkeit der Clusterings. Das bedeutet, dass er explizit die Annahme trifft, dass die Nutzer der drei Gruppen die Gruppierung der Webseiten anhand ihrer Ähnlichkeit durchschnittlich identisch durchgeführt haben. Um diese Bedingung zu beweisen, musste *Dellschaft* zeigen, dass die binären Entscheidungen der drei Gruppen hinsichtlich der inhaltlichen Gruppierung der 10 Webseiten im Durchschnitt identisch sind. Das Ergebnis des für diesen Zweck verwendeten χ^2-Tests bestätigte diese Annahme.

5.1.2 Interpretation im Diversitätskontext

Im Wesentlichen lassen sich aus den Ergebnissen der Arbeit von *Dellschaft* zwei Erkenntnisse gewinnen: Zum einen hat sich gezeigt, dass sich populäre Tagvorschläge gegenüber Tagvorschlägen bereits verwendeter Tags im Hinblick auf die Organisations- und Indexierungseigenschaft von Tagging-Systemen eher hinderlich auswirken. Zum anderen offenbart das web-basierte Nutzerexperiment, dass Nutzer voneinander unabhängig eine

im Durchschnitt identische Gruppierung von Ressourcen anhand ihrer Ähnlichkeit vornehmen und darüber hinaus die Ähnlichkeit von Ressourcen durch die Vergabe gleicher oder ähnlicher Tags ausdrücken.

Während die erste Erkenntnis für die vorliegende Arbeit von geringer Bedeutung ist, bildet die zweite die eigentliche Grundlage für das verwendete Evaluationsverfahren. Um Tags für die Messung der Diversität einer Menge von Ressourcen nutzen zu können, müssen Ressourcen anhand von Tags thematisch gruppierbar sein. Genau genommen müssen Tags zwei Bedingungen erfüllen, um für den beschriebenen Zweck verwendet werden zu können:

1. Tag-Zuweisungen müssen konsistent sein, d.h. unterschiedliche Nutzer müssen gleiche oder ähnliche Ressourcen auf gleiche oder ähnliche Weise taggen.
2. Anhand der Tags muss eine thematische Gruppierung von Ressourcen möglich sein.

Dass Tag-Zuweisungen die beiden Bedingungen erfüllen, wird durch das Nutzerexperiment von *Dellschaft* hinreichend demonstriert. Offen bleibt jedoch die Frage, wie man die Diversität einer Menge von Ressourcen anhand von Tag-Zuweisungen messen kann. In gewisser Weise gibt *Dellschaft* auch hierfür einen Hinweis: Wenn mehrere Testpersonen eine Anzahl von Ressourcen mit Tags annotieren sollen, dann ist das Ergebnis ein Tagvektor für jede Ressource. Die Ähnlichkeit zweier Tagvektoren lässt sich, wie *Dellschaft* es beispielsweise vorschlägt, durch Anwendung des Kosinusmaßes bestimmen. Die Ähnlichkeit kann man direkt zum Messen der Diversität heranziehen: Je unähnlicher zwei Tagvektoren sind, desto diverser sind die mit den Tagvektoren assoziierten Ressourcen. Die Diversität einer Menge könnte man folglich messen, in dem man für alle Paare von Tagvektoren das Kosinusmaß bestimmt und die Einzelergebnisse aggregiert.

Alternativ zum Kosinusmaß wird in dieser Arbeit ein entropie-basierter Ansatz verwendet, der im Wesentlichen auf dem C4,5 Algorithmus basiert. Im folgenden Unterkapitel wird das verwendete Maß definiert und anschließend das Verfahren anhand eines einfachen Beispiels demonstriert.

5.2 Definition des Evaluationsverfahrens

In Kapitel 3 wird der Information Gain Ratio (*gainratio*) als Kern des C4,5 Algorithmus präsentiert. Im Kontext des Evaluationsverfahrens wird *gainratio* zum Messen der Diversität einer Menge von Ressourcen verwendet. Hierfür wird die Gruppierungseigenschaft von Tags ausgenutzt. Das Messen der Diversität kann folglich auf eine einzige Fragestellung reduziert werden: *Wie eindeutig partitionieren (oder clustern) die Tag-Zuweisungen die zugrunde liegenden Publikationen?* Sind die Tag-Zuweisungen eindeutig, d.h. dass bestimmte Tags nur an bestimmte Publikationen und an keine anderen vergeben werden, dann kann unter Zuhilfenahme der Erkenntnisse von *Dellschaft* argumentiert werden, dass die betrachteten Publikationen divers sind. Genau für diesen Zweck eignet sich *gainratio*, denn es bietet im Grunde eine Aussage darüber, wie eindeutig ein bestimmtes Attribut den betrachteten Datensatz partitioniert. Je höher der Informationsgewinn, also der Wert, den *gainratio* annimmt, umso eindeutiger sind die Tag-Zuweisungen und damit die Partitionierung (oder Klassifizierung) der Publikationen durch die Tag-Zuweisungen.

Wie bereits erläutert, ist der Grundgedanke bei dem Evaluationsverfahren der, dass einer Menge von Testpersonen U zum einen eine Menge von Publikationen D_{prs} und zum anderen eine Menge von Tags T_{prs} präsentiert wird. Die Aufgabe der Testpersonen $u \in U$ besteht darin, den Publikationen $d \in D_{prs}$ Tags $t \in T_{prs}$ zuzuweisen. Die Definition des Information Gain Ratio Kriteriums (siehe Definition 3.11) lässt sich jedoch nicht ohne Weiteres auf das beschriebene Szenario der Evaluation übertragen. Beispielsweise fehlt ein Trainingsdatensatz, mit dessen Hilfe im klassischen Fall ein Entscheidungsbaum generiert wird, der anschließend für die Klassifizierung neuer Daten genutzt wird. Im Kontext der Evaluation fällt das Fehlen eines Trainingsdatensatzes jedoch nicht weiter ins Gewicht, da es nicht das Ziel ist, einen Entscheidungsbaum zu generieren und neue Daten zu klassifizieren. D.h. das Vorhandensein des Evaluationsdatensatzes, bestehend aus den Attributen *Tags*, *Publikationen* und *Testpersonen* sowie den *Einträgen*, die Zuweisungen von Tags an Publikationen durch die Testpersonen darstellen, reicht für die Evaluation vollkommen aus. Gravierender ist jedoch, dass keines der Attribute im Datensatz ein für die Definition des Information Gain Ratio Kriteriums notwendiges kategorisches Attribut darstellt, mit dem die Daten klassifiziert werden können.

Um dem letztgenannten Problem zu begegnen, werden alle Tags $t \in T_{prs}$ für sich allei-

ne betrachtet. Das bedeutet, dass *gainratio* für jeden Tag definiert wird, was aber nichts anderes heißt, als dass sich zwei Partitionen, oder besser gesagt Klassen, ergeben: Zum einen die Klasse C_1, die aus den Publikationen besteht, denen der Tag t zugewiesen wurde und zum anderen steht eine Klasse C_2 mit den Publikationen, denen t nicht zugewiesen wurde. Diese Adaption entspricht offensichtlich der Klassifizierung der Daten anhand eines kategorischen Attributes, so wie es in der Definition des *gainratio* gefordert wird.

Zum besseren Verständnis der folgenden Definitionen soll an dieser Stelle festgehalten werden, dass die Definitionen in der Regel einen Bezug zu einem konkreten Tag $t \in T_{prs}$ oder einer konkreten Testperson $u \in U_{prs}$ haben. Dieser Tatbestand äußert sich darin, dass stets tag-spezifische bzw. testperson-spezifische Indizes oder Parameter angegeben werden. Darüber hinaus soll auf Tabelle 5.1 hingewiesen werden. Diese enthält eine Auflistung der wesentlichen verwendeten Symbole samt ihrer Interpretation im Kontext der Evaluation und soll der Übersichtlichkeit dienlich sein.

Der Informationsgewinn *gain* ist definiert als die Differenz zwischen $info$ und $infoX$. Wie bereits erwähnt, wird *gain* auf Basis eines Tags t sowie im Kontext der Tag-Zuweisungen durch eine Testperson u bestimmt. In diesem Zusammenhang ist *gain* ein Maß dafür, wie eindeutig eine Testperson $u \in U$ den Publikationen $d \in D_{prs}$ den Tag $t_k \in T_{prs}$ (mit $k \in [1, ..., |T_{prs}|]$) zugewiesen hat. Im nachfolgenden Unterkapitel wird dieser Sachverhalt anhand eines plastischen Beispiels näher beleuchtet. Bevor *gain* im Sinne der Evaluation definiert wird, sollen einige Grundlagen formalisiert werden, die für alle nachfolgenden Definitionen relevant sind.

Symbol	Interpretation
U	Eine Menge von Testpersonen u
D_{all}	Menge der Publikationen d, die man für eine Query Q erhält
D_{prs}	Menge der Publikationen d, die den Testpersonen präsentiert werden
B	Ein Bag von Tags, mit denen die Publikationen $d \in D_{all}$ annotiert sind
T_{prs}	Menge der Tags t, die den Testpersonen präsentiert werden
T_{zug}	Menge der Tags $t \in T_{prs}$, die alle Testpersonen $u \in U$ den Publikationen $d \in D_{prs}$ zugewiesen haben.
T_{u_j}	Menge der Tags $t \in T_{prs}$, die eine Testperson $u_j \in U$ den Publikationen $d \in D_{prs}$ zugewiesen hat
P_{t_i}	Wahrscheinlichkeitsverteilung für eine Partition der Daten auf Basis von Tag t_i
V	Partition, basierend auf einem nicht-kategorischen Attribut
g_{t_k}	Gewichtungsfaktor für $gainratio(t_k)$

Tabelle 5.1: Die für die Definition des Evaluationsmaßes verwendeten Symbole sowie deren Interpretation.

Gegeben sei eine Menge U von Testpersonen mit $|U| = m$. Weiterhin sei für eine Testperson u_j mit $j \in [1, ..., m]$ die Menge T_{u_j} mit $|T_{u_j}| = n$ die Menge der Tags, die u_j den Publikationen $d \in D_{prs}$ zugewiesen hat. Mit $i \in [1, ..., n]$ ist für einen Tag t_i, im Kontext der Tag-Zuweisungen von Testperson u_j, der Informationsgewinn $gain_{u_j}(t_i)$ wie folgt definiert:

$$gain_{u_j}(t_i) = info_{u_j}(P_{t_i}) - infoX_{u_j}(t_i) \tag{5.3}$$

Festgehalten werden soll, dass P_{t_i} die Wahrscheinlichkeitsverteilung einer Partition ist, die durch den Tag t_i definiert wird. In anderen Worten lässt sich dieser Sachverhalt auch beschreiben als die Aufteilung der Einträge des Datensatzes in zwei Klassen C_1 und C_2, wobei C_1 die Einträge enthält, die einer Zuweisung von t_i entsprechen und C_2 alle restlichen Zuweisungen. Die Wahrscheinlichkeitsverteilung P_{t_i} hat die folgende Gestalt:

$$P_{t_i} = (\frac{|C_1|}{|C_1 \cup C_2|}, \frac{|C_2|}{|C_1 \cup C_2|}) \tag{5.4}$$

Analog zur Definition von $info$ in Kapitel 3 handelt es sich bei $info_{u_j}(P_{t_i})$ um die Information, die die Verteilung P_{t_i} und damit die Partitionierung durch t_i,bietet. Im Sinne der Evaluation ist $info_{u_j}(P_{t_i})$ wie folgt definiert:

$$\begin{aligned} info_{u_j}(P_{t_i}) = &-\frac{|C_1|}{|C_1 \cup C_2|} \cdot log_2 \frac{|C_1|}{|C_1 \cup C_2|} - \\ &- \frac{|C_2|}{|C_1 \cup C_2|} \cdot log_2 \frac{|C_2|}{|C_1 \cup C_2|} \end{aligned} \tag{5.5}$$

Für die Definition von $infoX_{u_j}(t_i)$ müssen die Daten zunächst auf Basis eines nichtkategorischen Attributes X (siehe Erläuterungen zu Definition 3.8) partitioniert werden. In

der Evaluation erfolgt die Partitionierung durch die Publikationen $d \in D_{prs}$. Das bedeutet, dass die Tag-Zuweisungen von Testperson u_j für jede Publikation $d \in D_{prs}$ betrachtet werden. Formell bedeutet das: Ist $|D_{prs}| = h$, dann existieren h Partitionen $V_1, ..., V_h$, wobei die Partition V_1 die Tag-Zuweisungen enthält, die Publikation d_1 durch Testperson u_j zugewiesen wurden. Analog enthält V_2 die Tag-Zuweisungen, die d_2 zugewiesen wurden usw. In diesem Zusammenhang sei V_{all} eine Partition, die alle Tag-Zuweisungen aus $V_1, ..., V_h$ enthält. $infoX_{u_j}(t_i)$ kann damit wie folgt definiert werden:

$$infoX_{u_j}(t_i) = \sum_{k=1}^{h} \frac{|V_k|}{|V_{all}|} \cdot info_{u_j}(P_{t_i}) \tag{5.6}$$

Anzumerken bleibt, dass die Wahrscheinlichkeitsverteilung P_{t_i} für jede Partition V_k definiert werden muss. Da jede Partition V_k die Tag-Zuweisungen für eine Publikation d_k darstellt und ein Tag t_i für eine Publikation durch eine Testperson lediglich ein einziges Mal vergeben werden kann, kann die Wahrscheinlichkeitsverteilung in diesem Kontext lediglich zwei unterschiedliche Formen haben:

1. $P_{t_i} = (\frac{1}{|V_k|}, \frac{|V_k|-1}{|V_k|})$

2. $P_{t_i} = (0, 1)$

Das bedeutet, dass der Tag t_i der Publikation entweder zugewiesen oder nicht zugewiesen wird.

In den Erläuterungen von Definition 3.10 wird u.A. erläutert, dass *gain* solche Attribute favorisiert, die viele Ausprägungen haben können. Die Konsequenz dieser negativen Eigenschaft wäre im Kontext der Evaluation die, dass Publikationen mit zahlreichen Tag-Zuweisungen einen dominierenden Einfluss auf das Ergebnis von *gain* hätten. Diese Problematik ist durchaus bekannt und die Fachliteratur weist auf einen Lösungsansatz, der eine Normalisierung von *gain* vorsieht. In diesem Zusammenhang spricht man von *gainratio*.

Die Normalisierung von *gain* erfolgt durch die Information, die eine Partition V_k enthält. Diese Information wird bezeichnet als $splitinfo$ und ist definiert über das nichtkategorische Attribut X:

$$splitinfo_{u_j}(X) = -\sum_{k=1}^{h} \frac{|V_k|}{|V_{all}|} \cdot log_2 \frac{|V_k|}{|V_{all}|} \tag{5.7}$$

Setzt man $gain_{u_j}(t_k)$ ins Verhältnis zu $splitinfo_{u_j}(X)$, erhält man die normalisierte Form des $gain_{u_j}(t_k)$:

$$gainratio_{u_j}(t_k) = \frac{gain_{u_j}(t_k)}{splitinfo_{u_j}(X)} \tag{5.8}$$

In dieser Form ist *gainratio* lediglich definiert über den Tag-Zuweisungen einer Testperson u_j. Da *gainratio* jedoch über den gesamten Datensatz, d.h. die Tag-Zuweisungen aller Testpersonen, definiert werden soll, muss $gainratio_{u_j}(t_k)$ über alle $j \in [1, ..., m]$ aggregiert werden:

$$gainratio(t_k) = \sum_{j=1}^{m} gainratio_{u_j}(t_k) \tag{5.9}$$

Wie bereits erwähnt, ist die Normalisierung von *gain* ein Hilfsmittel, mit dem die Favorisierung von Attributen mit zahlreichen Ausprägungen beschränkt werden soll. Daneben kann jedoch ein weiteres Problem auftreten, falls die Testpersonen seltene Tags überdurchschnittlich häufig und insbesondere auf chaotische Weise vergeben. In diesem Fall kann man davon ausgehen, dass derartige Tags die betroffenen Publikationen nicht adäquat beschreiben und sich damit nicht für eine Klassifizierung der Publikationen eignen. Falls ausschließlich schwache Tag-Zuweisungen vorliegen, diese jedoch einheitlich sind, kann durchaus eine adäquate Aussage hinsichtlich Diversität erfolgen. Sind jedoch schwache und zudem chaotische mit starken und korrekten Zuweisungen gepaart, dominieren die

schwachen Tag-Zuweisungen und verfälschen damit den Diversitätswert.

Eine Möglichkeit zur Beschränkung des Einflusses schwacher Tag-Zuweisungen auf die resultierenden Diversitätswerte besteht darin, die Tags zu gewichten. Im einfachsten Fall kann man die relative Frequenz der Tags im Korpus als Gewichtungsfaktor verwenden, was im Grunde nichts anderes als das weit verbreitete *TF-IDF* Maß darstellt. Aus Gründen der Konsistenz wird in dieser Arbeit ein etwas anderer Weg beschritten: Es wird ein Gewichtungsfaktor g_{t_k} für $gainratio(t_k)$ definiert, der für jeden Tag t_k den Informationsgehalt angibt, den dieser bezüglich des Korpus enthält. Damit wird sich der Mittel bedient, die der C4,5 Algorithmus von Haus aus mitbringt. Der Gewichtungsfaktor wird im Folgenden definiert.

Sei D_{all} eine Menge von Publikationen, die man für eine Anfrage Q erhält. D_{prs} ist eine Untermenge von D_{all}, für die das Diversitätsmaß, das in dieser Arbeit vorgestellt wird, maximal ist. Sei weiterhin B ein Bag von Tags t, mit denen die Publikationen $d \in D_{all}$ annotiert sind. Der Gewichtungsfaktor g_{t_k} wird definiert für $gainratio(t_k)$ auf Basis einer Partitionierung von B durch den Tag t_k. Eine derartige Partitionierung liefert zwei Partitionen: Zum einen eine Partition W_1, die alle Tags umfasst, die gleich dem Tag t_k sind, und zum anderen eine Partition W_2, die alle restlichen Tags umfasst. Damit kann der Gewichtungsfaktor g_{t_k} wie folgt definiert werden:

$$g_{t_k} = -\frac{|W_1|}{|B|} \cdot log_2 \frac{|W_1|}{|B|} - \frac{|W_2|}{|B|} \cdot log_2 \frac{|W_2|}{|B|} \tag{5.10}$$

Wie man an Definition 5.9 sehen kann, wird *gainratio* über jeden zugewiesenen Tag definiert. Da der Gewichtungsfaktor ebenfalls tag-spezifisch ist und lediglich Werte zwischen 0 und 1 annehmen kann, kann *gainratio* für einen Tag t_k definiert werden als das Produkt des Gewichtungsfaktors g_{t_k} für den Tag t_k mit der Summe der nutzer-spezifischen *gainratio*-Werte für t_k:

$$gainratio(t_k) = g_{t_k} \cdot \sum_{j=1}^{m} gainratio_{u_j}(t_k) \tag{5.11}$$

Im nachfolgenden Unterkapitel wird ein Beispiel betrachtet, in dem die Anwendung von Definition 5.11 auf einer einfachen (virtuellen) Ergebnisliste demonstriert wird.

5.3 Motivationsbeispiel für das Evaluationsverfahren

In dem folgenden Beispiel wird ein Szenario durchgespielt, in dem ein System für eine virtuelle Anfrage eine Anzahl von Publikationen liefert. Einer Kombination von drei Publikationen werden, durch drei virtuelle Testpersonen, eine Reihe von fünf möglichen Tags zugewiesen. Die Tag-Zuweisungen sind zwar nicht eineindeutig, aber eine Tendenz ist erkennbar, sodass eine subjektive Einschätzung der Diversität der drei Publikationen gemacht werden kann.

Das Ziel des Beispiels ist es, durch Anwendung des Evaluationsverfahrens, so wie es im vorangegangenen Unterkapitel definiert ist, die subjektive Einschätzung hinsichtlich der Diversität der drei Publikationen zu bestätigen und damit die theoretische Funktionsfähigkeit des Verfahrens zu demonstrieren. Ein derartiges Vorgehen ist natürlich kein Ersatz für eine großangelegte Evaluation mit zahlreichen Experimenten. Aufgrund der zeitlichen Beschränkung dieser Arbeit muss ein derartiges Unterfangen als eine zukünftige Aufgabe angesehen werden. Für eine Vergleichbarkeit der Ergebnisse werden jedoch zwei weitere Szenarien durchgespielt, in denen jeweils eine extreme Form der Tag-Zuweisungen untersucht wird. Zum einen werden eineindeutige und zum anderen chaotische Tag-Zuweisungen untersucht. Aus Gründen der Übersichtlichkeit sind beide Szenarien im Anhang dieser Arbeit zu finden und nicht in diesem Unterkapitel.

Im Folgenden soll die Annahme gelten, dass für eine Anfrage an das System eine Menge von 30 Publikationen mit insgesamt 41 Tags zurückgegeben wird. Insgesamt enthält die Menge fünf unterschiedliche Tags $t_1, t_2, ..., t_5$. Tabelle 5.2 sind die Häufigkeiten der fünf Tags zu entnehmen.

Tag	Frequenz
t_1	16
t_2	12
t_3	8
t_4	4
t_5	1

Tabelle 5.2: Frequenz der Tags $t_1, ..., t_5$ im Bag B von Tags. Der Bag enthält 41 Tags, mit denen die 30 Publikationen annotiert sind.

Tag	Gewicht
t_1	0.964
t_2	0.870
t_3	0.712
t_4	0.460
t_5	0.164

Tabelle 5.3: Tag-spezifische Gewichte für die Tags $t_1, ..., t_5$, bestimmt auf Basis der Taghäufigkeiten aus Tabelle 5.2 unter Verwendung von Definition 5.10. Ein größerer Wert impliziert ein höheres Gewicht des Tags und der mögliche Wertebereich liegt zwischen $[0...1]$.

Auf Basis der Taghäufigkeiten lassen sich mit Anwendung von Definition 5.10 die tag-spezifischen Gewichte bestimmen, die für die anschließende Bestimmung der *gainratio*-Werte notwendig sind. Die einzelnen Gewichte sind Tabelle 5.3 zu entnehmen. Mit einer Frequenz von 16 in einem Korpus von insgesamt 41 Tags ist Tag t_1 der am häufigsten vorkommende Tag und erreicht ein Gewicht von 0.964. Einen entgegengesetzten Fall stellt Tag t_5 dar, der mit einem einzigen Vorkommen im Korpus, lediglich auf ein Gewicht von 0.164 kommt. Das bedeutet, dass der für Tag t_5 ermittelte *gainratio*-Wert lediglich zu ca. 16% in die Aggregation der *gainratio*-Werte einfließen wird.

Wie erwähnt, wird eine Kombination von drei Publikationen $d_1, .., d_3$ von drei Testpersonen

	u_1	u_2	u_3
d_1	t_1, t_2, t_3	t_1	t_1, t_2
d_2	t_4	t_3, t_4	t_3, t_4
d_3	t_4, t_5	t_5	t_5

Tabelle 5.4: Zu sehen sind die Tag-Zuweisungen der drei Testpersonen $u_1, ..., u_3$ zu den Publikationen $d_1, ..., d_3$. Jede Testperson kann je Publikation beliebig viele Tags zuweisen, wobei ein Tag von einer Testperson nicht doppelt zugewiesen werden kann. Damit die Vergleichbarkeit der Zuweisungen gewährleistet wird, ist die Anzahl der Tag-Zuweisungen je Dokument über alle Testpersonen hinweg identisch.

$u_1, .., u_3$ mit Tags versehen. Die Tag-Zuweisungen der Testpersonen zu den Publikationen sind Tabelle 5.4 zu entnehmen. Wie man sehen kann, sind die Zuweisungen weder eineindeutig noch chaotisch. Eine Tendenz ist jedoch offensichtlich, sodass die folgenden Beobachtungen festgehalten werden können:

1. Publikation d_1 wird mehrheitlich mit den Tags t_1 und t_2 versehen.
2. Publikation d_2 wird in erster Linie mit t_4 assoziiert, wobei auch t_3 stark vertreten ist.
3. Publikation d_3 wird auf eindeutige Weise mit Tag t_5 assoziiert.
4. Es gibt nur wenige Überschneidungen der Tags. Konkret heißt das, dass t_3 sowohl mit d_1 als auch d_2 assoziiert wird und t_4 mit d_2 und d_3. In beiden Fällen könnte die Überschneidung jedoch zufälliger Natur sein.

Neben den wenigen Überschneidungen deuten die Tag-Zuweisungen auf eine relativ hohe Diversität der drei Publikationen hin. Aufgrund der Arbeit von *Dellschaft* liegt die Vermutung nahe, dass das hier beschriebene Szenario in einer realen Evaluation durchaus vorkommen könnte, vorausgesetzt das zugrunde liegende Diversitätsverfahren liefert entsprechend diverse Kombinationen von Ressourcen.

Im nächsten Schritt wird geprüft, in welcher Form sich die soeben gemachten Beobachtungen in der Metrik aus Definition 5.11 widerspiegeln. Hierfür muss jedoch zunächst Definition 5.8 auf den testperson-spezifischen Tag-Zuweisungen angewendet werden. Die Ergebnisse sind Tabelle 5.5 zu entnehmen. Wie man sehen kann, enthält die Tabelle die *gainratio*-Werte, die im Kontext der Tag-Zuweisungen jeder Testperson für jeden Tag

bestimmt werden. Damit ist die erste Zeile beispielsweise so zu deuten, dass Tag t_1 im Kontext der Tag-Zuweisungen von Testperson u_1 einen *gainratio*-Wert von 0.130 generiert.

Eine Betrachtung der Werte aus Tabelle 5.5 offenbart direkt, dass die anfangs gemachten Beobachtungen mit den Werten korrelieren. Am offensichtlichsten trifft dies für die Zuweisung von Tag t_5 zu, das durch die zweite und dritte Testperson maximal eindeutig der Publikation d_3 zugewiesen wird. Eine aussagekräftige Deutung der Ergebnisse kann jedoch lediglich durch Betrachtung der Einzelwerte nicht erfolgen. Aus diesem Grund werden die Werte gemäß der Definition 5.11 aggregiert und gewichtet. Das Ergebnis dieses Vorgangs ist Tabelle 5.6 zu entnehmen.

Mit den Ergebnissen kann untersucht werden, inwieweit die am Anfang dieses Unterkapitels gemachten Beobachtungen durch das Evaluationsverfahren untermauert werden.

Unter Punkt 1 wird gesagt, dass die Tags t_1 und t_2 die Publikation d_1 eindeutig beschreiben. Für t_1 wird ein sehr hoher Wert von 0.849 generiert, während t_2 jedoch einen verhältnismäßig niedrigen Wert von 0.296 erhält. Der Wert für t_1 scheint durchaus begründet, da alle Testpersonen diesen Tag der Publikation d_1 und keiner anderen Publikation zuweisen. t_2 wird zwar ebenfalls ausschließlich mit d_1 assoziiert, jedoch nicht von allen Testpersonen. Zudem wird t_2 einmal in Kombination mit zwei weiteren Tags verwendet und damit die Eindeutigkeit dessen Zuweisung deutlich reduziert.

Im zweiten Punkt wird die Zuweisung von t_3 und t_4 angesprochen, die in einer gewissen Wechselwirkung die Publikation d_2 charakterisieren. Eingeschränkt wird die Eindeutigkeit der Zuweisungen dieser beiden Tags dadurch, dass sie an zwei Stellen als Ausreißer interpretiert werden können. Trotz dieser Eigenschaft sowie der relativ niedrigen Gewichtung der beiden Tags kommen sie auf noch beachtlich gute Werte, was mit der Eindeutigkeit in deren Zuweisung zur Publikation d_2 begründet werden kann.

Unter Punkt 3 wird die Eindeutigkeit der Zuweisung von Tag t_5 erwähnt. Auf den ersten Blick zeigt sich jedoch, dass das Verfahren für t_5 den niedrigsten Wert generiert. Diese Beobachtung trügt jedoch. Zum einen zeigen die einzelnen Werte aus Tabelle 5.5, dass t_5

Tag	Testperson	*gainratio*
t_1	u_1	0.130
t_2	u_1	0.130
t_3	u_1	0.130
t_4	u_1	0.400
t_5	u_1	0.217
t_1	u_2	0.540
t_2	u_2	0.000
t_3	u_2	0.207
t_4	u_2	0.207
t_5	u_2	0.540
t_1	u_3	0.211
t_2	u_3	0.211
t_3	u_3	0.211
t_4	u_3	0.211
t_5	u_3	0.474

Tabelle 5.5: Zu sehen sind die ermittelten Tag-spezifischen *gainratio*-Werte basierend auf den Tag-Zuweisungen jeder Testperson. Die ersten fünf Einträge beziehen sich auf die erste Testperson, die nächsten fünf auf die zweite Testperson usw.

Tag	*gainratio*
t_1	0.849
t_2	0.296
t_3	0.390
t_4	0.376
t_5	0.201

Tabelle 5.6: Aggregierte und gewichtete *gainratio*-Werte. Bestimmt nach Definition 5.11 unter Verwendung der Werte aus Tabelle 5.5 sowie der Tag-spezifischen Gewichte aus Tabelle 5.3.

sehr eindeutig d_5 charakterisiert. Zum anderen muss das extrem niedrige Gewicht von t_5 beachtet werden, das dazu führt, dass lediglich ca. 16% des aggregierten *gainratio*-Wertes in das Ergebnis einfließen. An dieser Stelle könnte man argumentieren, dass die niedrige Gewichtung eines Tags eine sehr eindeutige Zuweisung zu stark degradiert. Dem mag so sein, jedoch ist bei eindeutiger Zuweisung eines schwachen Tags, der Einfluss auf das Ergebnis dennoch relativ groß, während eine chaotische Vergabe eines schwachen Tags keinen nennenswerten Einfluss hat.

Damit lässt sich zusammenfassend sagen, dass die anfangs gemachten Beobachtungen durch das Evaluationsverfahren durchaus bestätigt werden. Wie bereits erwähnt, werden im Anhang zwei weitere Beispiele auf ähnliche Weise diskutiert. Auch hierbei werden die Annahmen hinsichtlich der Diversität der betrachteten Publikationen, die sich aus den Tag-Zuweisungen eindeutig ergeben, durch das Evaluationsverfahren bestätigt.

6 Experiment auf realen Daten

Im Folgenden wird in einem ersten Schritt anhand eines ausgewählten Beispiels demonstriert, dass das in der Arbeit definierte Diversitätsmaß unter realen Bedingungen in der Lage ist, aus einer Menge von Ressourcen eine diverse Kombination von Ressourcen auszuwählen. Zu Vergleichszwecken werden zwei Kombinationen bestimmt, von denen die erste maximal divers, die zweite hingegen minimal divers ist. Auf beiden Kombinationen wird anschließend das Evaluationsverfahren angewendet, mit dem Ziel, den Unterschied hinsichtlich der Diversität zwischen beiden Kombinationen zu bestätigen oder zu widerlegen.

Das Ziel dieses Vorgehens ist es zum einen, die Ergebnisse zweier unterschiedlicher Diversitätsverfahren unter Zuhilfenahme des Evaluationsverfahrens miteinander zu vergleichen. Auf diese Weise wird beispielhaft geprüft, ob das entropie-basierte Diversitätsverfahren ein besseres, d.h. diverseres, Ergebnis generiert als ein einfaches, auf dem Zufallsprinzip basierendes Verfahren. Zum anderen kann durch den Vergleich der drei Kombinationen sowie die bewusst einfach gehaltene Ergebnisliste die Konsistenz der Diversitätsmessung durch das Evaluationsverfahren beispielhaft geprüft werden.

6.1 Erprobung des Diversitätsmaßes auf realen Daten und Messung der Diversität der Ergebnisse

In diesem Unterkapitel wird zunächst gezeigt, dass das Diversitätsverfahren, das in dieser Arbeit definiert wird, auf realen Bibsonomy-Daten und damit unter realistischen Bedingungen, aus einer Menge von Publikationen diejenige Kombination von Publikationen bestimmen kann, die ein hohes Maß an Diversität aufweist. Zu diesem Zweck werden zwei Kombinationen bestimmt: Bei der ersten Kombination handelt es sich um die Kombination, für die das Verfahren den maximalen Werte ermittelt hat. Die zweite Kombination bildet den krassen Gegensatz zur ersten, d.h. sie weist den minimalen Diversitätswert auf.

Da die Ergebnisliste überschaubar groß ist, werden lediglich Kombinationen mit drei Publikationen bestimmt. Der Vorteil dieser Begrenzung liegt darin, dass sie eine einfache Einschätzung der Diversität der jeweiligen Kombination ermöglicht: Für 3er Kombination lässt sich der Grad der Diversität leichter erkennen, als das beispielsweise für 5er oder 10er Kombinationen der Fall ist.

Anschließend wird für beide Kombinationen die Diversität, unter Verwendung des Evaluationsverfahrens, ermittelt und die Ergebnisse diskutiert.

Zur Gewinnung einer Ergebnisliste von Publikationen wird an das System eine Anfrage gerichtet, die 10 Publikationen liefern soll, in denen der Autor *Andreas Hotho* mitgewirkt hat. Angefragt werden die Attribute, die durch das Anwendungsszenario aus Kapitel 4 definiert werden. Es handelt sich hierbei um die Attribute *Autoren*, *Tags* und *Jahr* (Erscheinungsjahr). Da eine Einschätzung der Diversität der Kombinationen auch aufgrund des subjektiven Empfindens erfolgen soll, werden zusätzlich die *Titel* der Publikationen erfragt. Die vollständige und nummerierte Ergebnisliste ist Tabelle 6.1 zu entnehmen. Es bleibt noch anzumerken, dass der Tabelle auch die Taghäufigkeiten entnommen werden können - diese befinden sich in den Klammern hinter den Tags.

An dieser Stelle soll festgestellt werden, dass Tags im Bibsonomy-Datensatz in der Regel nicht in der Klarheit vorliegen, wie dies in Tabelle 6.1 der Fall ist. Um die Lesbarkeit und

Nr.	Titel	Autoren	Tags	Erscheinungsjahr
1	KAON - Towards a large scale Semantic Web	Andreas Hotho, Alexander Maedche, Christoph Schmitz, Steffen Staab, Rudi Studer, Gerd Stumme, York Sure	kaon (3x), web (3x), semantic (4x), watchdog (1x)	2002
2	Ontology-based Text Clustering	Andreas Hotho, Steffen Staab, Alexander Maedche	clustering (3x), text (2x), ontology (3x)	2001
3	Personalized Information Access in a Bibliographic Peer-to-Peer System	Peter Haase, Marc Ehrig, Andreas Hotho, Björn Schnizler	bibliographic (1x), p2p (3x), personalization (1x), recommender (2x)	2006
4	Text Clustering Based on Good Aggregations	Andreas Hotho, Alexander Maedche, Steffen Staab	clustering (4x), kmeans (1x), text (3x), ontology (1x)	2001
5	Text Classification by Boosting Weak Learners based on Terms and Concepts	Stephan Bloehdorn, Andreas Hotho	boosting (1x), classification (3x), ontology (1x), learning (3x), text (2x)	2004
6	Tag Recommendations in Folksonomies	Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt, Gerd Stumme	bibsonomy (1x), folksonomy (5x), recommender (3x), tag(3x), social (1x), web (1x)	2007
7	AEON - An approach to the automatic evaluation of ontologies	Johanna Völker, York Sure, Andreas Hotho	aeon (2x), automatic (2x), ontology (4x), web (3x), evaluation (1x)	2008
8	Semantic Network Analysis of Ontologies	Bettina Hoser, Andreas Hotho, Robert Jäschke, Christoph Schmitz, Gerd Stumme	personalization (1x), semantic (4x), mining (1x), ontology (3x)	2006
9	Boosting for Text Classification with Semantic Features	Stephan Bloehdorn, Andreas Hotho	boosting (1x), text (4x), semantic (2x), classification (3x)	2004
10	Conceptual User Tracking	Daniel Oberle, Bettina Berendt, Andreas Hotho, Jorge Gonzalez	tracking (4x), humanities (1x), conceptual (2x), mining (2x), user (1x)	2003

Tabelle 6.1: Ergebnisliste von Publikationen für den Autor *Andreas Hotho* entsprechend dem Schema, das in dem Anwendungsszenario definiert wird.

die Klarheit der Ergebnisse zu verbessern, werden beispielsweise spam-ähnliche Tags entfernt und die Tags bezüglich der Rechtschreibung homogenisiert. Darüber hinaus hat diese Bereinigung den Zweck, dass die Evaluierung direkt auf den vorhandenen Tags durchgeführt werden kann und man auf diese Weise nicht auf Testpersonen angewiesen ist. Aus Sicht des Evaluationsverfahrens und des zugrunde liegenden Maßes ist der Verzicht auf Testpersonen irrelevant. Viel wichtiger als das Vorhandensein von Testpersonen ist die Eindeutigkeit der Tags, da diese automatisch auf Gleichheit überprüft werden – dieser Forderung wird jedoch, wie gesagt, durch die Homogenisierung der Tags entsprochen.

Bei den Publikationen aus der Ergebnisliste sind einige Tendenzen und Wiederholungen zu beobachten. Beispielsweise sind die Autoren *Steffen Staab*, *Alexander Maedche*, *Gerd Stumme* und *Robert Jäschke* überdurchschnittlich häufig vertreten. Bei den Tags sind es *ontology*, *text*, *web* und *semantic*, die die stärkste thematische Ausrichtung der Publikationen darstellen. Was die Erscheinungsjahre betrifft, ist die Diversität bereits relativ hoch – kein Jahr kommt häufiger als zweimal und insgesamt kommen lediglich drei Jahre doppelt vor. Da jedoch eine diverse Kombination von nur drei Publikationen gesucht ist, kann bereits ein doppeltes Vorkommen großen Einfluss auf die Diversität der Kombination als Ganzes haben.

Im Folgenden werden einige, aus Sicht der Diversität, wesentliche Beobachtungen festgehalten:

1. Die Publikationen 2 und 4 ähneln sich stark. Der latente Topic ist offensichtlich das *Text Clustering* im Zusammenhang mit *Ontologien*. Die Autoren sowie das Erscheinungsjahr sind identisch. Die Tag-Zuweisungen überschneiden sich deutlich und unterstützen die Annahme bezüglich des latenten Topics der beiden Publikationen. Keine diverse, und erst recht nicht die maximal diverse, Kombination dürfte somit die Publikationen 2 und 4 gemeinsam enthalten. Man kann sogar davon ausgehen, dass die minimal diverse Kombinationen diese beiden Publikationen enthalten wird.

2. Die Publikation mit der Nummer 1 weist zum einen, eine hohe Anzahl von Autoren auf und zum anderen sind die mit der Publikation assoziierten Tags relativ selten vertreten. Unter *seltenen Tags* sind derartige Tags zu verstehen, die in erster Linie mit

einer einzigen Publikation assoziiert sind und keiner anderen aus der Ergebnisliste.

3. Seltene Tags sind ebenfalls bei den Publikationen 6 und 10 zu beobachten.
4. Die thematische Ausrichtung der Publikationen 1, 3, 6 und 10 scheint sich von den restlichen Publikationen abzuheben. Während die vorherrschenden Themen das *Text Clustering* bzw. *Text Classification* im Zusammenhang mit *Ontologien* zu sein scheint, werden in Publikation 1 das *Semantic Web*, in Publikation 3 P*eer-to-Peer Systeme*, in Publikation 6 *Recommender Systeme* und in Publikation 10 das *User Tracking* behandelt.

Anhand der so eben aufgestellten Beobachtungen, kann eine, wenn auch subjektiv bewertete, Tendenz hinsichtlich der zu erwartenden maximal bzw. minimal diversen Kombinationen abgegeben werden:

1. Aufgrund der sehr großen Ähnlichkeit der Publikationen 2 und 4, ist zu erwarten, dass die minimal diverse Kombination beide Publikationen enthalten wird.

2. Die große Anzahl unterschiedlicher Autoren der Publikation 1 sowie die relativ seltenen Tags, die mit der Publikation assoziiert sind, legt die Vermutung nahe, dass diese Publikation bei den meisten Kombinationen mit einem hohem Diversitätswert vertreten sein wird.

3. Generell ist davon auszugehen, dass die maximal diverse Kombination aus einer Kombination der Publikationen 1, 3, 6 und 10 bestehen wird.

Im nächsten Schritt wird geprüft, inwieweit diese Prognose durch das Diversitätsverfahren untermauert wird. Zu diesem Zweck wird das Diversitätsverfahren auf den 10 Publikationen der Ergebnisliste angewendet. Um eine Vergleichbarkeit der Ergebnisse zu ermöglichen, werden zum einen die 10 diversesten (*Top 10*) und zum anderen die 10 am wenigsten diversen (*Worst 10*) Kombinationen durch das Verfahren ermittelt. Die Kombination, mit dem höchsten Wert für die Diversität, ist Tabelle 6.2 und die mit dem niedrigsten Wert Tabelle 6.3 zu entnehmen. Die restlichen Kombinationen werden im nächsten Unterkapitel, im Kontext der Diskussion der Ergebnisse, diskutiert.

Nr.	Titel	Autoren	Tags	Erscheinungsjahr
1	KAON - Towards a large scale Semantic Web	Andreas Hotho, Alexander Maedche, Christoph Schmitz, Steffen Staab, Rudi Studer, Gerd Stumme, York Sure	kaon (3x), web (3x), semantic (4x), watchdog (1x)	2002
6	Tag Recommendations in Folksonomies	Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt, Gerd Stumme	bibsonomy (1x), folksonomy (5x), recommender (3x), tag(3x), social (1x), web (1x)	2007
10	Conceptual User Tracking	Daniel Oberle, Bettina Berendt, Andreas Hotho, Jorge Gonzalez	tracking (4x), humanities (1x), conceptual (2x), mining (2x), user (1x)	2003

Tabelle 6.2: Maximal diverse Kombination von Publikationen, bestimmt durch das Diversitätsverfahren, das auf den 10 Einträgen in der Ergebnisliste angewendet wird. Der Kombination wird durch das Diversitätsverfahren ein Wert von 43.652 zugewiesen.

Bevor die Ergebnisse genauer analysiert werden, kann bereits festgestellt werden, dass die Prognose und damit die subjektive Erwartung bezüglich der maximal bzw. minimal diversen Kombinationen, durch das Diversitätsverfahren bestätigt wird. Wie man an den Ergebnissen sehen kann, besteht die maximal diverse Kombination aus den Publikationen 1, 6 und 10. Die minimal diverse Kombination besteht aus den Publikationen 2, 4 und 9. Damit wird zum einen die Erwartung vollständig bestätigt, dass die maximal diverse Kombination aus den Publikationen 1, 3, 6 oder 10 bestehen sollte. Zum anderen enthält die minimal diverse Kombination die beiden fast-identischen Publikationen 2 und 4. Der erstgenannten Kombination weißt das Diversitätsverfahren einen Wert von 43.652 zu, während die minimal diverse Kombination lediglich auf 15.208 kommt. Der durchschnittliche Wert, der als Mittelwert aller Kombinationen bestimmt wird, liegt bei 29.182.

Nr.	Titel	Autoren	Tags	Erscheinungsjahr
2	Ontology-based Text Clustering	Andreas Hotho, Steffen Staab, Alexander Maedche	clustering (3x), text (2x), ontology (3x)	2001
4	Text Clustering Based on Good Aggregations	Andreas Hotho, Alexander Maedche, Steffen Staab	clustering (4x), kmeans (1x), text (3x), ontology (1x)	2001
9	Boosting for Text Classification with Semantic Features	Stephan Bloehdorn, Andreas Hotho	boosting (1x), text (4x), semantic (2x), classification (3x)	2004

Tabelle 6.3: Minimal diverse Kombination von Publikationen, bestimmt durch das Diversitätsverfahren, das auf den 10 Einträgen in der Ergebnisliste angewendet wird. Der Kombination wird durch das Diversitätsverfahren ein Wert von 15.208 zugewiesen.

Neben der diversen thematischen Ausrichtung, die bereits weiter vorne erläutert wird, ist an der maximal diversen Kombination zu beobachten, dass zum einen, sehr viele unterschiedliche Autoren abgedeckt werden. Selbst der, in der Ergebnisliste, stark vertrete Autor *Steffen Staab*, kommt lediglich ein einziges mal vor. Gleiches gilt für *Alexander Maedche* und *Robert Jaeschke*. Lediglich der ebenfalls stark präsente *Gerd Stumme* ist häufiger als einmal vertreten.

Neben der Heterogenität der Autoren, fallen die diversen Erscheinungsjahre auf - kein Jahr ist doppelt vertreten.

Zusammenfassend lässt sich damit für die maximal diverse Kombination sagen, dass es durchaus plausibel scheint, dass diese Kombination den höchsten Diversitätswert aufweist. Sowohl die thematische Ausrichtung, die zum einen aus den Titeln und zum anderen aus den Tags der Publikationen hergeleitet werden kann, als auch die Diversität der Autoren und der Erscheinungsjahre sprechen für eine hohe Diversität der drei Publikationen.

Mit den beiden fast-identischen Kombinationen bleibt im Prinzip nur die Frage offen, welche dritte Publikation die größte Ähnlichkeit zu den beiden erstgenannten Publikationen aufweist - das scheint die Publikation mit der Nummer 9 zu sein.

Die thematische Ausrichtung scheint zumindest ähnlich zu sein: Der in der Kombination starke, das heißt mit einer relativ hohen Frequenz vorhandene, Tag *text*, scheint eine dominante Wirkung zu haben.

Neben der thematischen Ähnlichkeit fällt auf, dass die Publikation 9 lediglich zwei Autoren aufweist, wobei nur der Autor *Stephan Bloehdorn* zur Diversifizierung der Autoren beiträgt. Generell ist die Anzahl der unterschiedlichen Autoren als gering zu bezeichnen, insbesondere im direkten Vergleich mit der maximal diversen Kombination.

Bei den Erscheinungsjahren passiert wenig überraschendes. Da in der Ergebnisliste kein Erscheinungsjahr mehr als zweimal vorkommt und Publikationen 2 und 4 bereits identische Erscheinungsjahre aufweisen, kann keine dritte Publikation gewählt werden, die ein identisches Jahr und damit eine geringere Diversität der Kombination bewirken könnte.

Mit der Bestätigung der subjektiven Einschätzung durch das Diversitätsverfahren bezüglich einer diversen und einer nicht-diversen Kombination von Publikationen, besteht der nächste Schritt darin, auf den gemachten Ergebnissen das Evaluationsverfahren anzuwenden. Das bedeutet konkret, dass die vergleichbaren Tag-Zuweisungen, mit denen die Publikationen bereits versehen sind, dazu genutzt werden, um eine Aussage hinsichtlich der Diversität der beiden Kombinationen herzuleiten. Bestätigt das Evaluationsverfahren die subjektive Einschätzung, die bereits durch das Diversitätsverfahren untermauert wird, so kann die erste Stufe des Experiment als erfolgreich abgeschlossen gelten.

Für das Evaluationsverfahren sind zunächst die tag-spezifischen Gewichte zu bestimmen. Der Korpus enthält insgesamt 100 Tags, davon sind 27 Tags unterschiedlich. In Tabelle 6.4 werden diese aufgelistet. Zudem zeigt die Tabelle die Frequenz der Tags sowie das bestimmte Gewicht.

Tag	Frequenz	Gewicht
boosting	2	0.132
classification	6	0.326
ontology	12	0.529
learning	3	0.191
text	11	0.499
bibsonomy	1	0.080
folksonomy	5	0.286
recommender	5	0.286
tag	3	0.191
social	1	0.080
web	7	0.365
aeon	2	0.132
automatic	2	0.132
personalization	2	0.132
semantic	10	0.468
tracking	4	0.241
humanities	1	0.080
conceptual	2	0.132
mining	3	0.191
user	1	0.080
evaluation	1	0.080
kmeans	1	0.080
kaon	3	0.191
watchdog	1	0.080
p2p	3	0.191
clustering	7	0.365
bibliographic	1	0.080

Tabelle 6.4: Tag-spezifische Gewichte für die Menge der Tags aus der Ergebnisliste. Zu sehen ist von links nach rechts der Tagbezeichner, dessen Frequenz im Korpus sowie das, für den Tag, bestimmte Gewicht.

Tag	*gainratio*-Wert	gewichteter *gainratio*-Wert
bibsonomy	0.024	0.001
folksonomy	0.137	0.039
recommender	0.077	0.022
tag	0.077	0.014
social	0.024	0.001
web	0.062	0.022
kaon	0.099	0.018
semantic	0.137	0.064
watchdog	0.031	0.002
tracking	0.149	0.035
humanities	0.033	0.002
conceptual	0.069	0.009
mining	0.069	0.013
user	0.033	0.002

Tabelle 6.5: Aufgeführt sind die *gainratio*-Werte sowie die gewichteten *gainratio*-Werte der Tags, die den drei Publikationen, der maximal diversen Kombination, zugewiesen sind.

Zunächst wird für die maximal diverse Kombination, auf Basis der Tag-Zuweisungen und unter Verwendung der Gewichte aus Tabelle 6.4, die tag-spezifischen *gainratio*-Werte je zugewiesenen Tag bestimmt. Das Ergebnis ist Tabelle 6.5 zu entnehmen. Anschließend wird derselbe Vorgang wiederholt, jedoch auf den Tag-Zuweisungen der minimal diversen Kombination. Diese Ergebnisse sind Tabelle 6.6 zu entnehmen.

Aggregiert man die einzelnen Werte, so erhält man als *gainratio*-Wert, für die maximal diverse Kombination, einen Wert von 0.244. Für die minimal diverse Kombination wird der Wert 0.178 ermittelt. Auch wenn der Unterschied der beiden Werte nicht groß zu sein scheint, ist dennoch ein Unterschied vorhanden, der zudem die subjektive Einschätzung

Tag	*gainratio*-Wert	gewichteter *gainratio*-Wert
boosting	0.034	0.004
text	0.007	0.003
semantic	0.071	0.033
classification	0.112	0.036
clustering	0.134	0.048
ontology	0.097	0.051
kmeans	0.038	0.003

Tabelle 6.6: Aufgeführt sind die *gainratio*-Werte sowie die gewichteten *gainratio*-Werte der Tags, die den drei Publikationen, der minimal diversen Kombination, zugewiesen sind.

als auch die Ergebnisse des Diversitätsverfahrens untermauert.

6.2 Messung der Diversität einer randomisiert generierten Menge von Publikationen

In der ersten Stufe des Experiments wurden mit dem Evaluationsverfahren die Ergebnisse des Diversitätsverfahrens untersucht. Konkret bedeutet das, dass für zwei unterschiedlich diverse Kombinationen, die vom Diversitätsverfahren generiert wurden, die Diversität mit dem Evaluationsverfahren gemessen wurde. Die Messung hat sowohl den subjektiv empfundenen Unterschied in der Diversität der beiden Kombinationen als auch die Ergebnisse des Diversitätsverfahrens bestätigt.

In der zweiten Stufe des Experiments, der dieses Unterkapitel gewidmet ist, wird zunächst eine zufällige Auswahl von drei Publikationen getroffen. Die Auswahl der Publikationen

Nr.	Titel	Autoren	Tags	Erscheinungsjahr
2	Ontology-based Text Clustering	Andreas Hotho, Steffen Staab, Alexander Maedche	clustering (3x), text (2x), ontology (3x)	2001
5	Text Classification by Boosting Weak Learners based on Terms and Concepts	Stephan Bloehdorn, Andreas Hotho	boosting (1x), classification (3x), ontology (1x), learning (3x), text (2x)	2004
7	AEON - An approach to the automatic evaluation of ontologies	Johanna Völker, York Sure, Andreas Hotho	aeon (2x), automatic (2x), ontology (4x), web (3x), evaluation (1x)	2008

Tabelle 6.7: Eine Kombination von drei zufällig (randomisiert) ausgewählten Publikationen.

erfolgt mit einem Zufallsgenerator[4], der Zahlen zwischen 1 und 10 ausgibt. Der Zufallsgenerator lieferte die Zahlenkombination 2, 5 und 7 und die mit den Nummern assoziierten Publikationen sind Tabelle 6.7 zu entnehmen.

Kurz erwähnt werden soll, dass sich die zufällige Auswahl von Ressourcen durchaus für die Generierung diverser Kombinationen eignen kann. Nicht auszuschließen ist hierbei jedoch, dass identische oder fast-identische Ressourcen in einer Kombination vertreten sein können. Dennoch zeigt sich bei einer Betrachtung der Kombination in Tabelle 6.7, dass das Ergebnis, zumindest nach subjektivem Ermessen ein gewisses Maß an Diversität aufweist. Sowohl die Autoren als auch die Erscheinungsjahre kann man als divers interpretieren. Das vorherrschende Thema scheinen Ontologien zu sein, wie man sowohl an den Titeln (siehe Publikationen 2 und 7) als auch an der Zuweisung des Tags *ontology* zu allen drei Publikationen erkennen kann. Damit scheint die thematische Diversität der Publikationen beschränkt zu sein.

Nach subjektivem Ermessen kann man festhalten, dass die zufällig ausgewählte Kom-

4 $http://www.cognitive-tools.de/.../zufallszahlen\_erzeugen.html$

Tag	*gainratio*-Wert	gewichteter *gainratio*-Wert
clustering	0.136	0.050
text	0.069	0.034
ontology	0.037	0.019
boosting	0.034	0.004
classification	0.111	0.036
learning	0.111	0.021
aeon	0.059	0.007
automatic	0.059	0.007
web	0.092	0.033
evaluation	0.028	0.002

Tabelle 6.8: Aufgeführt sind die *gainratio*-Werte sowie die gewichteten *gainratio*-Werte der Tags, die den drei Publikationen der randomisiert generierten Kombination zugewiesen sind.

bination, im Hinblick auf die Diversität, der minimal diversen Kombination aus dem vorangegangenen Unterkapitel überlegen ist. Zugleich scheint jedoch die maximal diverse Kombination diverser zu sein. Zu prüfen bleibt, ob die Messung der Diversität der zufällig generierten Kombination durch das Evaluationsverfahren diese Annahme mit einem entsprechenden Wert untermauert. Die tag-spezifischen, gewichteten und ungewichteten *gainratio*-Werte für die zufällig generierte Kombination sind Tabelle 6.8 zu entnehmen.

Aggregiert man die einzelnen Werte aus Tabelle 6.8, erhält man einen *gainratio*-Wert von 0.213. Zur Erinnerung: Die maximal diverse Kombination erhielt einen Wert von 0.244 und die minimal diverse Kombination kam lediglich auf 0.178. Damit wird die Annahme bestätigt, dass das Diversitätsverfahren diverse Mengen von Publikationen bestimmen kann.

6.3 Diskussion der Ergebnisse

In den beiden letzten Unterkapiteln wird ein Experiment präsentiert, in dem die theoretische Funktionsfähigkeit des in dieser Arbeit definierten Evaluationsverfahrens gezeigt werden sollte. Zu diesem Zweck wurde das Evaluationsverfahren auf den Ergebnissen (Kombinationen von Publikationen) zweier Diversitätsmaße angewendet. Für beide Ergebnisse wurde auf diese Weise die Diversität gemessen und die Diversitätsmaße damit vergleichbar gemacht. Hierbei gab es zwei grundlegende Erkenntnisse:

1. Alle Messergebnisse haben die subjektive Wahrnehmung hinsichtlich der Diversität der betrachteten Kombinationen von Ressourcen bestätigt.
2. Das in dieser Arbeit definierte Diversitätsmaß ist prinzipiell dazu in der Lage, eine diverse Kombination von Publikationen zu bestimmen. Zumindest im Vergleich zu dem einfachen, auf dem Zufallsprinzip basierenden Diversitätsmaß gilt die begründete Annahme, dass das entropie-basierte Diversitätsverfahren in den meisten Fällen eine diversere Kombination bestimmen kann.

Ein einfaches Experiment mit lediglich drei Messungen kann die Frage, ob das Evaluationsverfahren die Diversität von Kombinationen adäquat messen kann, natürlich nicht beantworten. Daher besteht eine der wichtigsten zukünftigen Aufgaben darin, eine Evaluation mit möglichst vielen Testpersonen durchzuführen, um für die besagte Frage eine hinreichende und wissenschaftlich fundierte Antwort zu erhalten.

In dem Experiment wurde das entropie-basierte Diversitätsmaß dazu genutzt, um zwei unterschiedlich diverse Kombinationen von Ressourcen zu bestimmen. Die erste Kombination wies einen maximalen und die zweite einen minimalen Diversitätswert auf. Eine genaue Betrachtung beider Kombinationen hat gezeigt, dass die minimal diverse Kombination fast-identische Publikationen enthielt. Der beobachtete und begründete Unterschied der Diversität wurde durch die Messergebnisse des Evaluationsverfahrens eindeutig bestätigt.

Da das Evaluationsverfahren lediglich Tag-Zuweisungen betrachtet, lohnt sich ein Blick auf die Tag-Zuweisungen der drei Kombinationen. Den Tag-Zuweisungen der drei Kombinationen aus Tabelle 6.9 kann entnommen werden, dass die maximal diverse Kombination

Kombination	Tags Publikation 1	Tags Publikation 2	Tags Publikation 3
max	kaon(3x), web (3x), semantic (4x), watchdog (1x)	bibsonomy (1x), folksonomy (5x), recommender (3x), tag(3x), social (1x), web (1x)	tracking (4x), humanities (1x), conceptual (2x), mining (2x), user (1x)
min	clustering (3x), text (2x), ontology (3x)	clustering (4x), kmeans (1x), text (3x), ontology (1x)	boosting (1x), text (4x), semantic (2x), classification (3x)
random	clustering (3x), text (2x), ontology (3x)	boosting (1x), clasification (3X), ontology (1x), learning (3x), text (2x)	aeon (2x), automatic (2x), ontology (4x), web (3x), evaluation (1x)

Tabelle 6.9: Die Tabelle zeigt die Tag-Zuweisungen der maximal und der minimal diversen sowie der randomisiert generierten Kombination. Die Taghäufigkeiten sind den Klammern hinter den jeweiligen Tags zu entnehmen.

die höchste Anzahl von Zuweisungen aufweist: 35 Tag-Zuweisungen im Vergleich zu 26 für die minimal diverse und 30 für die randomisiert generierte Kombination. Darüber hinaus zeigt sich, dass die Tag-Zuweisungen der maximal diversen Kombination ein hohes Maß an Eindeutigkeit aufweisen. Lediglich der Tag *web* wird zwei von drei Publikationen zugewiesen, wobei Publikation 2 lediglich eine einzige Zuweisung des Tags aufweist. Betrachtet man im Vergleich dazu die Zuweisungen der randomisiert generierten Kombination, so fallen sogar zwei Überschneidungen auf: Zum einen wird *text* sowohl Publikation 1 als auch Publikation 2 zugewiesen. Zum anderen wird der Tag *ontology* allen Publikationen zugewiesen, wobei die Zuweisung zur Publikation 2 mit einer einzigen Zuweisung kaum ins Gewicht fällt. Besonders stark fallen aber die überschneidenden Tag-Zuweisungen von *clustering* und *text* in der minimal diversen Kombination auf. *text* ist in allen drei Publikationen stark vertreten und *clustering* in den ersten beiden Publikationen.

Zusammenfassend lässt sich sagen, dass die Tag-Zuweisungen der drei Kombinationen die Messergenbnisse des Evaluationsverfahrens bestätigen.

7 Zusammenfassung und Ausblick

Das Kernthema der vorliegenden Arbeit ist die Diversifizierung von Suchergebnissen im Web Kontext, wobei die Arbeit selbst zwei Schwerpunkte aufweist. Den ersten Schwerpunkt bildet die Definition eines auf der bedingten Entropie basierenden Diversitätsmaßes. Es wird ein Framework implementiert, das für ein konkretes Anwendungsszenario auf dem Bibsonomy Datensatz Nutzern die Suche nach Publikationen ermöglicht und aus einer Menge von Ergebnissen für eine Anfrage die Kombination von Publikationen bestimmt, für die das Maß und damit die Diversität maximal ist. Den zweiten und wesentlichen Schwerpunkt bildet die Entwicklung eines auf Tag-Zuweisungen basierenden Verfahrens für die Evaluierung von Diversitätsmaßen.

Anhand eines Experiments auf ausgewählten realen Bibsonomy Daten wird gezeigt, dass das in der Arbeit vorgestellte Diversitätsverfahren aus einer Ergebnismenge von Publikationen eine diverse Submenge bestimmen kann. In dem Experiment werden die Ergebnisse des Diversitätsverfahrens mit denen eines einfachen, auf dem Zufallsprinzip basierenden Verfahrens verglichen und mit dem in der Arbeit definierten Evaluationsverfahren evaluiert.

Auch wenn die Messergebnisse des Experiments argumentativ untermauert werden können, bleibt die grundlegende Frage offen, inwieweit sich das Evaluationsverfahren für die Evaluierung von Diversitätsmaßen eignet. Eine der offenen Aufgaben ist daher die Erprobung des Evaluationsverfahrens. Das kann beispielsweise anhand eines Datensatzes erfolgen, der bereits im Hinblick auf die Diversität der Ressourcen evaluiert wurde. Ein derartiger Datensatz ist der *TREC 2009 Web Track* Datensatz [CCS09], der seit vielen Jahren einen Standard in der Evaluierung von Retrieval Systemen darstellt.

Eine weitere offene Aufgabe besteht in der Implementierung eines Plug-ins für die Bibsonomy Plattform, das neben den unterschiedlichen Suchmöglichkeiten auch die Suche diverser Publikationen ermöglicht. In diesem Zusammenhang lassen sich weitere interessante Fragestellungen untersuchen. Eine mögliche Fragestellungen betrifft beispielsweise den Einfluss diverser Suchergebnisse auf die Nutzerzufriedenheit, d.h. in welchem Umfang sind die Nutzer von Web 2.0 Plattformen überhaupt an diversen Suchergebnissen interessiert und inwieweit decken diverse Suchergebnisse den Informationsbedarf der Nutzer.

A Anhang

Als Zusatz zu dem plastischen Beispiel aus Kapitel 5 werden in diesem Anhang zwei weitere Beispiele angeführt, die zwei extreme Szenarien darstellen. Im ersten Beispiel wird ein Szenario betrachtet, in dem alle Testpersonen exakt gleiche Tag-Zuweisungen tätigen. Im Gegensatz dazu sind die Tag-Zuweisungen im zweiten Szenario chaotisch und ermöglichen im Grunde keine Aussage hinsichtlich der Diversität der betrachteten Publikationen.

Um eine Vergleichbarkeit des Beispiels aus Kapitel 5 mit den beiden folgenden zu gewährleisten, werden die bereits für das erste Beispiel formulierten Tag-Gewichte auch für die kommenden beiden Beispiele herangezogen. Außerdem ist die Anzahl der Tag-Zuweisungen in allen drei Beispielen identisch.

Tabelle A.1 sind die Tag-Zuweisungen für das erstgenannte Szenario zu entnehmen. Wie man sehen kann, sind die Zuweisungen aller Testpersonen identisch und eindeutig. Das bedeutet, dass beispielsweise die Tags t_1 und t_2 eindeutig die Publikation d_1 klassifizieren usw.

	u_1	u_2	u_3
d_1	t_1, t_2	t_1, t_2	t_1, t_2
d_2	t_3, t_4	t_3, t_4	t_3, t_4
d_3	t_5	t_5	t_5

Tabelle A.1: Zu sehen sind die einheitlichen Tag-Zuweisungen der drei Testpersonen $u_1, ..., u_3$ zu den Publikationen $d_1, ..., d_3$.

Derartige homogene Tag-Zuweisungen sind in der Realität kaum zu erwarten, dennoch bietet die Betrachtung dieses Szenarios einige wichtige Erkenntnisse. So lässt sich beispielsweise auf eindrucksvolle Weise der Einfluss des Gewichtungsfaktors auf die *gainratio*-Werte beobachten.

Betrachtet man die Werte in Tabelle A.2, die durch Anwendung von Definition 5.8 zu Stande kommen, so fällt sofort auf, dass die Werte für die Tags t_1 bis t_4 identisch sind. Deren gewichteten und aggregierten Werte (siehe Tabelle A.3) unterscheiden sich aber deutlich voneinander. Im extremsten Fall, also bei den Tags t_1 und t_4, liegt der Unterschied immerhin bei 0.319.

Noch eindrucksvoller ist die gewicht-bedingte Abwertung von t_5: Hierbei handelt es sich um den einzigen Tag, der die Publikation d_3 klassifiziert, was sich in einem überdurchschnittlich hohem *gainratio*-Wert niederschlägt. Der resultierende gewichtete und aggregierte Wert hingegen ist der niedrigste von allen fünf betrachteten Werten. Man muss allerdings auch anmerken, dass in einer derart klaren Klassifizierung, wie sie im Falle von Tag t_5 und Publikation d_3 vorliegt, die starke Abwertung des *gainratio*-Wertes von t_5 nicht ganz gerechtfertigt ist - sie hat aber auch keine verfälschenden Auswirkungen. Im Falle zahlreicher Zuweisungen eines schwach-gewichtigen Tags jedoch würde dieser Tag den *gainratio*-Wert zu stark beeinflussen. Ein derartiger Fall wird im nächsten Beispiel erläutert.

Das zweite Szenario, das hier präsentiert werden soll, stellt eine Situation dar, in der die Testpersonen chaotische Tag-Zuweisungen vornehmen. Tabelle A.4 zeigt auf gewohnte Weise die Zuweisungen der drei virtuellen Testpersonen.

Wie man sehen kann, wird keine Publikation durch ein Tag oder eine Kombination von Tags eindeutig klassifiziert. Die Tag-Zuweisungen jeder Testperson deuten zwar zunächst darauf hin, dass die drei betrachteten Publikationen nicht divers sind - die Betrachtung der Tag-Zuweisungen über alle Testpersonen hingegen verzerrt diese Beobachtung und verhindert eine eindeutige Aussage bezüglich der Diversität der Publikationen.

Tag	Testperson	*gainratio*
t_1	u_1	0.211
t_2	u_1	0.211
t_3	u_1	0.211
t_4	u_1	0.211
t_5	u_1	0.474
t_1	u_2	0.211
t_2	u_2	0.211
t_3	u_2	0.211
t_4	u_2	0.211
t_5	u_2	0.474
t_1	u_3	0.211
t_2	u_3	0.211
t_3	u_3	0.211
t_4	u_3	0.211
t_5	u_3	0.474

Tabelle A.2: Zu sehen sind die ermittelten Tag-spezifischen *gainratio*-Werte basierend auf den Tag-Zuweisungen jeder Testperson. Die ersten fünf Einträge beziehen sich auf die erste Testperson, die nächsten fünf auf die zweite Testperson usw.

Tag	*gainratio*-Wert (gewichtet)
t_1	0.610
t_2	0.550
t_3	0.450
t_4	0.291
t_5	0.233

Tabelle A.3: Aggregierte und gewichtete *gainratio*-Werte je Tag. Bestimmt anhand der Werte aus Tabelle A.2 und unter Verwendung der tag-spezifischen Gewichte aus Kapitel 5.

	u_1	u_2	u_3
d_1	t_1, t_2	t_3, t_4	t_5
d_2	t_2	t_3, t_5	t_1, t_4
d_3	t_1, t_2	t_3	t_4, t_5

Tabelle A.4: Zu sehen sind die chaotischen Tag-Zuweisungen der drei Testpersonen $u_1, ..., u_3$ zu den Publikationen $d_1, ..., d_3$. Jede Testperson kann je Publikation beliebig viele Tags zuweisen, wobei ein Tag von einer Testperson für eine Publikation nicht doppelt zugewiesen werden kann.

Tabelle A.5 sind die einzelnen *gainratio*-Werte und Tabelle A.6 die resultierenden aggregierten und gewichteten Werte zu entnehmen.

Zusammenfassend lässt sich festhalten, dass die Beobachtungen aus Kapitel 5 durch die beiden so eben vorgestellten Szenarien weiter bekräftigt werden.

Tag	Testperson	*gainratio*
t_1	u_1	0.112
t_2	u_1	0.112
t_3	u_1	0.000
t_4	u_1	0.000
t_5	u_1	0.000
t_1	u_2	0.000
t_2	u_2	0.000
t_3	u_2	0.112
t_4	u_2	0.211
t_5	u_2	0.211
t_1	u_3	0.211
t_2	u_3	0.000
t_3	u_3	0.000
t_4	u_3	0.112
t_5	u_3	0.375

Tabelle A.5: Zu sehen sind die ermittelten Tag-spezifischen *gainratio*-Werte basierend auf den Tag-Zuweisungen jeder Testperson. Die ersten fünf Einträge beziehen sich auf die erste Testperson, die nächsten fünf auf die zweite Testperson usw.

Tag	*gainratio*-Wert (gewichtet)
t_1	0.610
t_2	0.550
t_3	0.450
t_4	0.291
t_5	0.233

Tabelle A.6: Aggregierte und gewichtete *gainratio*-Werte je Tag. Bestimmt anhand der Werte aus Tabelle A.5 und unter Verwendung der tag-spezifischen Gewichte aus Kapitel 5.

B Danksagung

An dieser Stelle möchte ich mich ausdrücklich bedanken bei Prof. Staab dafür, dass er mir ein interessantes und abwechslungsreiches Thema für meine Abschlussarbeit vorgeschlagen hat. Bei Sergej Sizov möchte ich mich für die langjährige Zusammenarbeit bedanken. Für das Korrekturlesen gebührt ein tiefer Dank Magdalena Rohrbeck.

In erster Linie möchte ich mich jedoch bei meinem Vater bedanken, dem ich so viel zu verdanken habe und leider nicht mehr die Gelegenheit dazu habe mich zu revangieren. *Daher widme ich ihm diese Arbeit.*

Literaturverzeichnis

[AGH+09] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, Samuel Ieong, and Samuel Ieong. Diversifying search results. In *WSDM*, 2009.

[BHJ+10] Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, Gerd Stumme, and Gerd Stumme. The social bookmark and publication management system bibsonomy - a platform for evaluating and demonstrating web 2.0 research. 2010.

[CCS09] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. Overview of the trec 2009 web track. 2009.

[CGG98] Jaime G. Carbonell, Jade Goldstein, and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[CKC+08] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, Ian MacKinnon, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, 2008.

[CMZ+09] Olivier Chapelle, Donald Metlzer, Ya Zhang, Pierre Grinspan, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, 2009.

[Del] Klaas Dellschaft. The influence of tag suggestions on users during tagging. Unpublished paper.

[DPP10] Marina Drosou, Evaggelia Pitoura, and Evaggelia Pitoura. Search result diversification. 2010.

[LSC09] Ziyang Liu, Peng Sun, and Yi Chen. Structured search result differentiation. 2009.

[Mor02] Tatsunori Mori. Information gain ratio as term weight: The case of summarization of ir results. In *COLING*, 2002.

[SCS10] Tetsuya Sakai, Nick Craswell, and Ruihua Song. Simple evaluation metrics for diversified search results. In *EVIA*, 2010.

[SDPP10] Kostas Stefanidis, Marina Drosou, Evaggelia Pitoura, and Evaggelia Pitoura. Perk: personalized keyword search in relational databases through preferences. In *EDBT*, 2010.

[Sha01] Claude E. Shannon. A mathematical theory of communication. 2001.

[SMM01] Barry Smyth, Paul McClave, and Paul McClave. Similarity vs. diversity. In *ICCBR*, 2001.

[VSAYAY09] Erik Vee, Jayavel Shanmugasundaram, Sihem Amer-Yahia, and Sihem Amer-Yahia. Efficient computation of diverse query results. 2009.