Brandenburgische Technische Universitat Cottbus
Institut für Informatik
Lehrstuhl Datenbank- und Informationssysteme

Bachelor-Arbeit

Analyse von Kalibrierungsansätzen für die CQQL-Auswertung

Robert Kuban
MatrikelNr.: 2911138
Informatik

07.06.2012
05.10.2012

Betreuer: Prof. Dr.-Ing. Ingo Schmitt

Inhaltsverzeichnis

1 Einleitung

1.1 Fakten- und Information-Retrieval

Datenbanken und Information-Retrieval entwickelten sich längere Zeit unabhängig voneinander. Während erste Datenbankanwendungen im Umfeld von Versicherungen und Banken entstanden, findet man die Ursprünge des Information-Retrieval im Bibliothekswesen [Hen08, S. 27]. Dementsprechend unterscheiden sich diese in einigen Eigenschaften.

Anfragen an eine Datenbank, auch Fakten-Retrieval genannt [Hen08, S. 33], ergeben für einen bestimmten Datenbankzustand eine eindeutig bestimmbare Lösung, welche aus den strukturiert in der Datenbank abgelegten Daten berechnet wird. Datenbankanfragen werden meist in einer künstlichen Sprache formuliert, welche an formelle Sprachen wie relationaler Algebra oder Tupel- beziehungsweise Bereichs-Kalkül angelehnt sind.

Beispiel für Fakten-Retrieval: In Tabelle 1.1 sieht man einen möglichen Zustand einer Datenbank für Nahrungsmittel. Die Tabelle enthält deren Namen, einen Preis pro Kilogramm und einen kurzen Beschreibungstext. Eine typische Fakten-Retrieval-Anfrage auf diese Tabelle in SQL ist „`SELECT Name, Preis FROM Nahrungsmittel WHERE Preis < 1.5`“ und liefert als Antwort Tabelle 1.2.

Im Information-Retrieval kann hingegen auf eine Anfrage oft keine eindeutige Lösung berechnet werden. Vielmehr liefert eine solche Anfrage eine Ordnung nach Relevanz bezüglich der Anfrage [Hen08, S. 33]. Diese Anfragen können dabei meist in natürlicher Sprache formuliert werden. Da die zu durchsuchenden Dokumente im Vergleich zu Datenbanken oft weniger stark strukturiert sind, werden im Information-Retrieval-System vereinfachte Dokument-Repräsentationen genutzt, welche die Semantik der Dokumente ausdrücken sollen. Die Nähe dieser Repräsentationen zu der Repräsentation einer Anfrage lässt sich dann berechnet und durch einen Score-Wert ausdrücken.

Beispiel für Information-Retrieval: Eine typische Information-Retrieval-Anfrage auf den in Tabelle 1.1 gespeicherten Beschreibungstexten wäre „`blaue schmackhafte Frucht`“. Tabelle 1.3 zeigt eine mögliche Repräsentation der Dokumente und der Anfrage und Tabelle 1.4 ein nach Scores absteigend sortiertes Ergebnis, wobei die Scores dem Nutzer üblicherweise nicht mitgeteilt werden.

Name	Preis	Beschreibung
Apfel	1.90	Ein schmackhaftes Tafelobst.
Birne	1.30	Schmackhaft und frisch aus der Gegend.
Pflaume	1.20	Eine blaue Frucht.

Tabelle 1.1: Möglicher Inhalt einer Datenbank

Name	Preis
Birne	1.30
Pflaume	1.20

Tabelle 1.2: Ergebnis der Fakten-Retrieval-Anfrage

Dokument	schmackhaft	Tafelobst	frisch	Gegend	blau	Frucht
Anfrage	1	0	0	0	1	1
Apfel	1	1	0	0	0	0
Birne	1	1	1	1	0	0
Pflaume	0	0	0	0	1	1

Tabelle 1.3: Repräsentation der Anfrage und der Dokumente des Beispiels

Neben dem klassischen Information-Retrieval, welches sich mit der Suche in Textdokumenten beschäftigt, gibt es auch das Multimedia-Retrieval, welches zur Erschließung von multimedialen Inhalten, also Bildern, Audio- und Videoinhalten, genutzt wird.

Zunehmend wird jedoch deutlich, dass es Probleme gibt, die weder mit reinem Fakten-Retrieval noch mit Information-Retrieval allein zu lösen sind [Hen08, S. 33]. Man kann sich zum Beispiel leicht vorstellen, dass es den Wunsch gibt, die Ähnlichkeit zwischen multimedialen Inhalten in eine Datenbankanfrage einzubeziehen. Damit ergibt sich die Frage nach Systemen, die Charakteristiken beider Retrieval-Arten aufweisen.

Beispiel für gemischtes Retrieval: Es kann ein Obst gesucht sein, welches einen Kilopreis von ungefähr 1,50 € hat und dessen Beschreibung „`blaue schmackhafte Frucht`“ ähnelt.

1.2 CQQL

CQQL steht für Commuting Quantum Query Language und ist ein Teil der in [Sch08] vorgestellten Anfragesprache QQL, welche die Kombination von scharfen, unscharfen und Retrievalanfragen erlaubt. Dabei bestehen die Anfragen aus booleschen Verknüpfungen von Prädikaten, welche sowohl scharfe als auch unscharfe Attributvergleiche erlauben.

Beispiel von Misch-Retrieval in CQQL: Die in Beispiel 1.1 beschriebene Anfrage kann in CQQL als (Preis ≈ 1.5) $\wedge$ (Beschreibung $\approx$ „blaue schmackhafte Frucht“) ausgedrückt werden.

Rang	Score	Name
1	0.816	Pfaume
2	0.408	Apfel
3	0.289	Birne

Tabelle 1.4: Ergebnis der Information-Retrieval-Anfrage des Beispiels

Dokument/Zeile	(Preis ≈ 1.5)	(Beschreibung ≈ „. . .“)	Auswertung
Apfel	0.912	0.408	0.372
Birne	0.992	0.289	0.287
Pflaume	0.983	0.816	0.802

Tabelle 1.5: Auflistung möglicher Ähnlichkeitswerte und das Ergebnis der CQQL-Auswertung

(Preis ≈ 1.5)	(Beschreibung ≈ „. . .“)	Gesamt-Score	intuitives Ergebnis
Birne	Pflaume	Pflaume	Pflaume
Pflaume	Apfel	Apfel	Birne
Apfel	Birne	Birne	Apfel

Tabelle 1.6: Ranking nach verschiedenen Merkmalen

Tabelle 1.5 zeigt mögliche Werte der in der Formel enthaltenen Prädikate und den berechneten Score.

1.3 Kalibrierung

Als Kalibrierung wird im Folgenden die Normalisierung von Ähnlichkeitswerten im Bezug auf die Auswertung in CQQL-Formeln bezeichnet.

Beispiel für ein Problem, welches Kalibrierung nötig macht : Wenn man die Rangfolge der Objekte bezüglich der einzelnen Werte der Prädikate und die Rangfolge des Endergebnisses in Tabelle 1.6 betrachtet, fällt auf, dass scheinbar nur die Rangfolge des Prädikats (Beschreibung ≈ „blaue schmackhafte Frucht“) Einfluss auf die Gesamtreihenfolge hat. Da aber beide Prädikate gleichberechtigt und ungewichtet in der Formel stehen, widerspricht dies der vermuteten Intention des Nutzers.

Das im Beispiel gezeigte Problem tritt in vielen Disziplinen auf, bei denen Scores aus verschiedenen Quellen zusammengeführt werden. Im Bereich des Information-Retrievals wurde das Problem insbesondere in Bezug auf Meta-Suchmaschinen untersucht, wobei bessere Ergebnisse mithilfe von Normalisierungsverfahren erhalten wurden. Normalisierungsverfahren bereiten Scores aus verschiedenen Quellen für eine zusammenfassende Auswertung vor. Sie werden beispielsweise auch im Bereich des Data-Mining [SSK06] oder in der Biometrie [JNR05] eingesetzt.

1.4 Gliederung der Arbeit

Im zweiten Kapitel der Arbeit wird CQQL kurz vorgestellt. Dabei wird auf den Aufbau und die Auswertung von CQQL-Formeln eingegangen und die Anforderungen an Scores behandelt, welche als Eingabe für die Auswertung genutzt werden können. Danach wird auf übliche Anforderungen eingegangen, mit denen im Information-Retrieval die Nutzbarkeit von Scores für die Anfragebearbeitung bewertet werden. Im Anschluß werden Normalisierungsverfahren aus der Literatur vorgestellt, welche die Qualität der Scores verbessern sollen.

Das dritte Kapitel geht darauf ein, ob und wie die im zweiten Kapitel vorgestellten Normalisierungsverfahren bei der CQQL-Auswertung genutzt werden können.

Im vierten Kapitel wird eine Definition des Kalibrierungsproblem erarbeitet und eine Methode entwickelt, um Kalibrierungsfehler bei der Auswertung einfacher CQQL-Formeln zu messen. Dieses Maß wird dann genutzt, um mögliche Ursachen für das Kalibrierungsproblem zu untersuchen.

Das fünften Kapitel liefert eine Zusammenfassung der in der Arbeit behandelten Aspekte und einen Ausblick.

2 Übersicht über das Themengebiet

2.1 Commuting Quantum Query Language

In [Sch08] wird CQQL als Teil der Anfragesprache QQL vorgestellt. Dieser Artikel ist die Grundlage der folgenden Übersicht. In diesem Teil der Arbeit wird auf den Aufbau von CQQL-Anfragen eingegangen und kurz die Auswertung der Anfragen beschrieben. Auf den theoretischen Hintergrund der Anfragesprache wird dabei nur so weit eingegangen, wie er die Möglichkeiten der Kalibrierung beschränkt.

2.1.1 Aufbau einer CQQL-Anfrage

Wie bereits in der Einleitung erwähnt, unterstützt CQQL sowohl Fakten- als auch Information-Retrieval-Anfragen. Diese beiden Anfragetypen werden in CQQL mithilfe zweier Attributtypen unterstützt: Datenbank-Attribute sind Attribute, bei denen bei der Auswertung einer Anfrage Abstände zwischen ihnen keine Rolle spielen. Diese Art von Attributen treten bei der Verarbeitung klassischer Datenbankanfragen auf. Ein einfaches Beispiel für ein Datenbank-Attribut ist ein boolescher Wahrheitswert. Im Unterschied dazu gibt es Retrieval-Attribute, deren Abstand bei der Auswertung der Anfrage bedeutsam ist. Dieser Attributtypbietet bietet die Möglichkeit, unscharfe Anfragen und Information-Retrieval-Anfragen in eine CQQL-Anfrage zu integrieren.

CQQL ist rekursiv definiert. Zuerst müssen Atome als kleinste mögliche CQQL-Anfragen definiert werden:

- Eine Selektionsbedingung $A_i = c$, wobei A_i ein Attribut ist und c eine Konstante.
- Eine Gleichheitsbedingung $A_{i_1} = \cdots = A_{i_n}$, wobei die Attribute $A_{i_1}, \ldots, A_{i_n}$ zum gleichen Attributtyp gehören.
- Eine Elementbedingung $A_i \in C$, wobei A_i ein Datenbank-Attribut ist und C eine konstante Menge.

Eine Menge von Atomen ist kommutierend, wenn kein Retrieval-Attribut in zwei verschiedenen Atomen in der Menge enthalten ist.

Über einer Menge A von kommutierenden Atomen kann nun eine CQQL-Anfrage definiert werden. Dabei gilt folgende Definition, welche aus [Sch08] entnommen ist:

1. Jedes Atom $a \in A$ ist eine CQQL-Anfrage.
2. Ist a eine CQQL-Anfrage, dann ist auch $(\neg a)$ eine CQQL-Anfrage.
3. Sind a und b CQQL-Anfragen, dann sind auch $(a \wedge b)$ und $(a \vee b)$ CQQL-Anfragen.

Die Forderung der Kommutativität der Atome ergibt sich dabei daraus, dass verschiedene Werte von Retrieval-Attributen auf nicht-orthogonale Vektoren abgebildet werden können. Die auf CQQL aufbauende Sprache QQL besitzt diese Einschränkung nicht mehr und erlaubt den Einsatz von Quantoren und Relationen. In dieser Arbeit wird darauf nicht näher eingegangen, [Sch08] behandelt dies umfassend.

2.1.2 Auswertung

Obwohl CQQL auf Quantenlogik basiert, können CQQL-Anfragen einfach arithmetisch ausgewertet werden. Eine CQQL-Anfragen lässt sich auf einen Tupel t auswerten, wenn sie bestimmte Voraussetzungen erfüllt: Die beiden Teilanfragen einer jeden Konjunktion dürfen keine gemeinsamen Retrieval-Attribute enthalten. Disjunktionen unterliegen auch dieser Beschränkung, wenn sie nicht exklusiv sind.

Beispiel für die Auswertbarkeit von CQQL-Anfragen: Seien a, b, c eine kommutierende Menge von Atomen, welche Retrieval-Attribute enthalten.

Die CQQL-Anfragen $(a \wedge a)$ und $(a \vee a)$ sind nicht auswertbar, da das Atom a und damit das in a enthaltene Retrieval-Attribut jeweils in beiden Teilanfragen vorkommt.

Im Unterschied dazu ist die Anfrage $(a \wedge b) \vee ((\neg a) \wedge c)$ auswertbar: Die Teilanfrage von $(a \wedge b)$ und $((\neg a) \wedge c)$ besitzen jeweils keine gemeinsamen Retrieval-Attribute. $(a \wedge b)$ und $((\neg a) \wedge c)$ sind durch eine Disjunktion verknüpft und schließen sich aus.

Wenn diese Voraussetzungen erfüllt sind, folgt die Auswertung von CQQL-Anfragen mithilfe der folgenden Vorschriften den Regeln einer booleschen Algebra [Sch08]:

- $\text{eval}^t(\neg a) = 1 - \text{eval}^t(a)$
- $\text{eval}^t(a \wedge b) = \text{eval}^t(a) \cdot \text{eval}^t(b)$
- $\text{eval}^t(a \vee b) = \text{eval}^t(a) + \text{eval}^t(b) - \text{eval}^t(a) \cdot \text{eval}^t(b)$

Es existiert ein Algorithmus, welcher alle CQQL-Anfragen in eine auswertbare Form überführen kann [Sch08]. Da dies im Kontext von CQQL auch als Normalisierung bezeichnet wird, dient der Begriff Kalibrierung auch der Abgrenzung zu diesem Vorgang.

Atome werden ausgewertet, indem die Attributwerte auf Vektoren abgebildet werden und der quadrierten Kosinus des eingeschlossenen Winkel errechnet wird. Verschiedene Datenbank-Attributen werden auf paarweise orthogonale Vektoren abgebildet, so das die Auswertung nur 0 oder 1 ergeben kann. Damit sind Atome, die Datenbankattribute enthalten, für die Betrachtung des Kalibrierungsproblem nicht relevant. Retrieval-Attribute, die in Atomen enthalten sind, werden mithilfe einer Funktion $f : \text{Dom}(A) \to \mathbb{R}^n$ auf einen Vektor abgebildet. Hierbei ist Dom der Wertebereich des Attributs.

Da die Auswertung dabei dem quadrierten Kosinus des zwischen den Vektoren eingeschlossenen Winkels entspricht, ergeben sich folgende Konsequenzen für die so errechneten Werte [Sch08]:

1. $\forall a, b \in \mathrm{Dom}(A) : \mathrm{eval}(a = b) \in [0, 1]$
2. $\forall a \in \mathrm{Dom}(A) : \mathrm{eval}(a = a) = 1$
3. $\forall a, b \in \mathrm{Dom}(A) : \mathrm{eval}(a = b) = s(b, a)$

Wie man sieht, ist die Auswertung des Atoms ein Ähnlichkeitsmaß auf den verglichenen Attributen.

In der praktischen Anwendung möchte man übliche Ähnlichkeitsmaße verwenden, ohne konkret eine Abbildung f anzugeben zu müssen. Trotzdem soll sichergestellt sein, das eine Abbildung f existiert. Eine Bedingung dafür ist die positive Semidefinitheit der Ähnlichkeitsmatrix, welche wie folgt definiert wird: Seien $a_1, \ldots, a_n \in \mathrm{Dom}(A)$ und s ein Ähnlichkeitsmaß auf $\mathrm{Dom}(A)$, dann ist M eine Ähnlichkeitsmatrix definiert als:

$$M = (s(a_i, a_j))_{i,j} = \begin{pmatrix} s(a_1, a_1) & \ldots & s(a_1, a_n) \\ \vdots & \ddots & \vdots \\ s(a_n, a_1) & \ldots & s(a_n, a_n) \end{pmatrix} \tag{2.1}$$

2.2 Ansätze für Normalisierungsverfahren

Die in späteren Teilen der Arbeit vorgestellten Normalisierungsverfahren stammen aus dem Bereich der Meta-Suche. Dort werden sie verwendet, um Scores aus verschiedenen Quellen auf eine anschließende Fusion vorzubereiten. Dabei gibt es die Möglichkeit, mithilfe von bewerteten Dokumentkollektionen, wie die TREC-Kollektionen, empirisch zu ermitteln, wie die Anwendung verschiedener Verfahren die Qualität von Suchergebnissen beeinflusst. Dieses Vorgehen liefert aber kaum Anhaltspunkte, warum sich einige Normalisierungsansätze besser für bestimmte Anwendungen eignen als andere.

2.2.1 Angleichung an die Wahrscheinlichkeit der Relevanz

Im Information-Retrieval werden Dokumente oft bezüglich einer Anfrage eines Nutzers in relevante und irrelevante Dokumente eingeteilt. Wu, Crestani und Bi bezeichnen Normalisierungsverfahren als ideal für bestimmte Fusionsmethoden, wenn der normalisierte Scores eines Dokuments, bis auf einen für alle Quellen gleichen Faktor, der Wahrscheinlichkeit entspricht, dass das Dokument relevant ist [WCB06]. In [WCB06] wird auf dieser Basis ein Test entwickelt, um Normalisierungsverfahren empirisch auf ihre Eignung zu prüfen. Diese Herangehensweise liefert allerdings kaum Anhaltspunkte, wie ein Normalisierungsverfahren aufgebaut sein sollte, um dieses Kriterium zu erfüllen, insbesondere wenn keine nach Relevanz bewerteten Trainingsdaten vorliegen.

2.2.2 Angleichung von Scoreverteilungen

Eine anderer Ansatz ist die Angleichung bestimmter Eigenschaften der Verteilungen normalisierter Scores. Viele Veröffentlichungen fordern die Abbildung aller Scores aus verschiedenen Quellen in einen gemeinsamen Bereich [SSK06], da Scores aus verschiedenen Quellen, welche in unterschiedliche Bereiche fallen, als schlecht vergleichbar gelten [ARK09, MA01, JNR05]. Die Anforderung, dass alle Scores in das Intervall $[0, 1]$ abgebildet werden, ergibt sich bei der Kalibrierung schon aus den Anforderungen der CQQL-Auswertung.

Diese Forderung nach einem gemeinsamen Wertebereich kann auch dazu ausgeweitet werden, dass normalisierte Scores aus verschiedenen Quellen gleiche Verteilungen aufweisen sollten, da auch unterschiedliche Score-Verteilungen als Ursache für Unvergleichbarkeit ausgemacht wurden [MA01, ARK09, OC03]. Fernández, Vallet und Castells scheinen diesen Ansatz in der mir bekannten Literatur am konsequentesten zu verfolgen, da sie eine Normalisierungsmethode vorstellen, welche Scores aus Quellen so abbildet, dass sie eine exakt gleiche Verteilung besitzen.

Im vierten Kapitel wird überprüft, ob die Angleichung von Score-Verteilungen eine Lösung für das Kalibrierungsproblem ist.

2.2.3 Entfernung von semantisch nicht bedeutsamen Einflüssen

Eine Ansatz es, nicht allen möglichen Komponenten eines Scores semantische Bedeutung zuzugestehen. Montague und Aslam fordern unter anderem, dass Verschiebungen und Skalierungen aller eingegebenen Werte aus einer Quelle wenig Einfluss auf die ausgegebenen normalisierten Werte haben sollten. Diese Eigenschaften bezeichnen sie als Verschiebungsinvarianz beziehungsweise Skalierungsinvarianz [MA01]. Ist $\mathrm{src}(a)$ eine Score-Quelle und $\mathrm{src}'(a) = \mathrm{src}(a) + c$ mit $c \in \mathbb{R}$ eine um c verschobene Score-Quelle, dann ist ein Normalisierungverfahren verschiebungsinvariant, wenn für alle a gilt: $\mathrm{norm}(\mathrm{src}(a)) = \mathrm{norm}(\mathrm{src}'(a))$. Analog dazu ist norm skalierungsinvariant, wenn für $\mathrm{src}(a)$ und $\mathrm{src}'(a) = c \cdot \mathrm{src}(a)$ mit $c \neq$ gilt, dass $\mathrm{norm}(\mathrm{src}(a)) = \mathrm{norm}(\mathrm{src}'(a))$ [MA01].

Avampatzis und Kamps betrachten einen Score-Wert als Summe aus einer semantisch bedeutsamen Signal- und einer unbedeutenden Noise-Komponente. Sie gehen davon aus, dass das das Verhältnis von Signal und Noise nur von der Größe des Gesamt-Scores abhängt [AK09]. Dabei ist das Ziel der Normalisierung das Verringern oder Entfernen des Einflusses einer solchen Noise-Komponente. Bezieht man diese Betrachtungsweise auf die von Montague und Aslam definierte Verschiebungs- und Skalierungsinvarianz, bemerkt man, dass diese einen linearen Zusammenhang von Gesamt-Score und Noise-Komponente implizieren. Diese Linearität wird von Avampatzis und Kamps nicht gefordert.

2.3 Robustheit von Normalisierungsverfahren

Eine übliche Anforderung an Normalisierungverfahren ist, dass diese nicht zu stark von einzelnen Werten eines einzelnen Scores abhängen sollten. Montague und Aslam fordern

in [MA01, S. 3], dass Normalisierungensverfahren unsensibel gegenüber Ausreißern sind. Ausreißer bezeichnen dabei einzelne, ungewöhnlich hohe oder niedrige Werten.

Dieses Kriterium ähnelt dabei stark einem Bewertungskriterium von Schätzern aus der Statistik. Schätzer werden zur Bestimmung von Lage- und Streuungsparametern von Verteilungen einer Grundgesamtheit aus einer Stichprobe genutzt. Eine Sensibilitäts-Kurve beschreibt dabei „den Einfluss einer einzelnen, wandernden Beobachtung auf eine gerade interessierende Schätzung“ [Ham80]. Wandernd bezieht sich dabei darauf, dass sich die Beobachtung von einem Wertebereich entfernt, der gut zu dem genutzten Modell, also einer vermuteten Verteilung der Scores, passt.

Neben einer flachen Sensibilitäts-Kurve gibt es zwei weitere Hauptziele für einen robusten Schätzer: Ein hoher Bruchpunkt und eine „hohe Effizienz unter dem parametrischem Modell für ‚gute‘ Daten“ [Ham80]. Der Bruchpunkt ist der Anteil „gerade noch tolerierter Ausreißer, bevor eine Schätzung ‚zusammenbricht‘ und total unzuverlässig wird“ [Ham80]. Die Effizienz ist ein Maß für die „die Nähe der Schätzung zu einer optimalen Schätzung, bei der die Verteilung der Daten bekannt ist“ [JNR05]. Sie ist auch außerhalb des Kontext der Robustheit ein gebräuchliches Kriterium zur Beurteilung der Qualität von Schätzern [wika].

Jain, Nandakumara und Ross fordern von einer guten Normalisierung, dass genutzte Parameter effizient und robust geschätzt werden [JNR05]. Das erscheint insbesondere sinnvoll, wenn man davon ausgeht, dass die im Retrieval-System vorhandenen Dokumente eine Stichprobe einer größeren Grundgesamtheit sind. Deshalb werden im Weiteren Angaben zur Robustheit und Effizienz von vorgestellten Verfahren gemacht. Diese sind aus [JNR05] entnommen.

Robustheit und Effizienz als alleiniges Bewertungskriterium für Normalisierungsverfahren zu benutzen, erscheint mir nicht sehr sinnvoll, da man dort verwendete Schätzer potenziell auch für andere Verfahren benutzen kann.

2.4 Normalisierungen

Dieser Abschnitt soll eine Übersicht über Normalisierungsverfahren aus der Literatur bieten.

Wie bereits bei der Betrachtung der Anforderungen an Normalisierungsverfahren bemerkt, ist die Normalisierung im Kontext des Information Retrieval insbesondere im Bereich der Meta-Suche untersucht worden. Dabei sind die Ausgaben von Suchmaschinen häufig Scores, so dass viele der folgenden Verfahren zur Normalisation von Scores dienen. Im dritten Kapitel wird dann auf die Anpassung zur Verwendung mit Ähnlichkeitswerten eingegangen.

2.4.1 Lineare Normalisierungen

Eine lineare Normalisierung transformiert Scores mithilfe einer linearen Funktion in der Form $\text{norm}(x) = a_1 \cdot x + a_0$. In [WCB06] wird eine andere Schreibweise für lineare Normalisierungen benutzt, in der sich sämtliche hier vorgestellten linearen Normalisierungen

Name	$y_{\min}$	$y_{\max}$	$x_{\min}$	$x_{\max}$	Anmerkung
Min-Max	0	1	$\min(S)$	$\max(S)$	
Fitting	a	b	$\min(S)$	$\max(S)$	$0 < a < b < 1$
Sum	0	1	$\min(S)$	Summe(S)	
ZMUV	0	1	μ_S	σ_S	
ZMUV2	2	3	μ_S	σ_S	
MAD	0	1	Median(S)	MAD(S)	
NORMEXP	0	c	$\min(S)$	$\sigma_{\text{exponential}}$	$c > 0$ konstant
Decimal Scaling	0	1	0	10^n	$n = \lceil \ln(\max(S)) \rceil$

Tabelle 2.1: Parameter der linearen Normalisierungen

darstellen lassen:

$$\text{norm}(s) = y_{\min} + \frac{s - x_{\min}}{x_{\max} - x_{\min}}(y_{\max} - y_{\min})$$

Diese Darstellung macht sich zu Nutze, dass zwei Punkte ausreichen, um eine lineare Funktion eindeutig zu bestimmen. Die Bezeichnungen $x_{\min}, x_{\max}, y_{\min}$ und $y_{\max}$ sind durch die Definition von Min-Max und Fitting motiviert und sind nur bei diesen beiden Verfahren auch das tatsächlichen Minimum beziehungsweise Maximum der Scores. Im Weiteren werden diese Werte als Normalisierungsparameter bezeichnet. Natürlich lassen sich beide Darstellungen auch leicht in einander umwandeln, wobei $a_0 = y_{\min} - \frac{x_{\min}}{x_{\max} - x_{\min}}$ und $a_1 = \frac{y_{\max} - y_{\min}}{x_{\max} - x_{\min}}$ ist.

Eine tabellarische Übersicht der Parameter der vorgestellten lineare Normalisierungsverfahren aus der Literatur findet sich in Tabelle 2.1. S bezeichnet dabei die aus einer Quelle erhaltenen Scores beziehungsweise die bei der CQQL-Auswertung eines Atoms auf einer Menge von Tupeln erhaltenen Ähnlichkeitswerte.

In Tabelle 2.2 sind einige Eigenschaften der aufgeführten linearen Normalisierungen zusammengefasst. Die Angaben zu Robustheit und Effizienz im statistischen Sinne sind aus [JNR05] entnommen. Die Angaben „unsensibel“ und „sensibel“ beschreiben die Sensibilität gegenüber Ausreißern nach [MA01]. Die so gekennzeichneten Verfahren sind nicht robust.

Min-Max

Dieses Verfahren wird in verschiedenen Veröffentlichungen als Min-Max [JNR05], Zero-One [WCB06] oder auch als Standard-Normierung [MA01] bezeichnet. Es bildet das Maximum aller Scores einer Anfrage auf 1 und das Minimum auf 0 ab. Damit bildet es alle Scores in das Intervall $[0, 1]$ ab [JNR05].

Da dieses Verfahren stark vom Minimum und Maximum der Scores einer Anfrage abhängt, ist es anfällig bezüglich Ausreißern [MA01, S. 3], also nicht robust [JNR05]. Damit ist es am Besten für Fällen geeignet, in denen obere und untere Schranken von Scores bekannt sind [JNR05], da diese dann nicht aus den Scores ermittelt werden müssen.

Name	Scalierungs- und Verschiebungsinvarianz	Robustheit	Effizienz
Min-Max	✓	Sensibel	/
Fitting	✓	Sensibel	/
Sum	✓	Unsensibel	/
ZMUV	✓	Unsensibel	Hoch
ZMUV2	✓	Unsensibel	Hoch
MAD	✓	Robust	Moderat
NORMEXP	✓	Unsensibel	/
Decimal Scaling	-	Sensibel	/

Tabelle 2.2: Eigenschaften der linearen Normalisierungsverfahren

Die beiden anderen Kriterien nach Montague und Aslam, die Skalierungs- und die Verschiebungsinvarianz, werden durch Min-Max erfüllt [MA01].

Fitting

Dokumente mit maximalem Score sind nicht immer relevant und Dokumente mit minimalem Score nicht immer irrelevant. Diese Überlegung führt in [WCB06] zur Kritik an der Min-Max-Normalisierung und der Betrachtung des Fittings als alternatives Verfahren. Dabei werden die Scores nicht wie bei Min-Max in das Intervall $[0, 1]$ abgebildet, sondern in ein kleineres Intervall $[a, b]$ mit $0 < a < b < 1$. Dabei wurden mit $a = 0.06$ und $b = 0.6$ gute Ergebnisse in einer experimentellen Überprüfung erzielt, wobei nicht ausgeschlossen wird, dass mit anderen Werten eventuell bessere Ergebnisse erreicht werden können [WCB06].

Fittings wird gelegentlich auch als Min-Max bezeichnet [SSK06], was sicher an der starken Ähnlickeit zu diesem Verfahren liegt. Aufgrund dieser Ähnlichkeit zu Min-Max ist es auch leicht zu sehen, dass Fitting skalierungs- und verschiebungsinvariant ist, aber sensibel gegenüber Ausreißern ist.

Sum

Sum wurde entwickelt, um für die Meta-Suche neben der Skalierungs- und Verschiebungsinvarianz auch Unsensibilität gegenüber Ausreißern zu erhalten [MA01]. Namensgebend ist dabei die Tatsache, dass die Summe aller Scores, die bei einer Anfrage erhalten werden, auf 1 abgebildet wird. Allerdings wird auch die Abbildung des arithmetischen Mittels auf 1 als ähnlich wirkungsvoll beschrieben [MA01]. Da die Summe sich aus allen Scores zusammensetzt, wird sie als unsensibel gegenüber Ausreißern angesehen [MA01]. Das Minimum der Scores wird als unsensibel angesehen, da bei der vorgesehenen Verwendung in einer Meta-Suchmaschine ein Schwellwert bei der Quell-Suchmaschine vermutet wird [MA01]. Bei der Auswertung von Prädikaten auf Datenbankinhalten kann dies nicht vorausgesetzt werden. Analog dazu mag die Summe der Scores in dem Anwendungsfall unsensibel gegenüber Ausreißern sein, allerdings wird der ähnlich zu ermittelnde arithmetische Mittel nicht als robust im statistischen Sinne angesehen [JNR05].

Dokument	Wert	ZMUV	ZMUV2
1	0.1	-1.286	0.714
2	0.5	-0.076	1.924
3	0.6	0.227	2.227
4	0.9	1.135	3.135
Erwartungswert	0.525	0	2
Varianz	0.109	1	1
Standardabweichung	0.330	1	1

Tabelle 2.3: Beispiel für ZMUV und ZMUV2

ZMUV

ZMUV steht für „Zero-Mean, Unit-Variance“ [MA01], was zwei Eigenschaften einer Standardnormalverteilung beschreibt: Einen Erwartungswert μ von 0 und eine Varianz σ^2 von 1. Sind Erwartungswert und Varianz bekannt, entspricht dieses Verfahren der in der Statistik verwendeten Standardisierung oder auch z-Transformation, welche normalverteilte Werte auf standard-normal-verteilte Werte abbildet. Es ist deshalb auch unter der englischen Bezeichnung „z-Score“ [JNR05] anzutreffen.

$$Z = \frac{X - \mu}{\sigma}$$

In [JNR05] wird ZMUV insbesondere bei bekannten Parametern empfohlen, da arithmetisches Mittel und die Standartabweichung nicht robust im statistischen Sinne [JNR05] sind. Montague und Aslam bezeichnen die Parameter aber als unsensibel gegenüber Ausreißern [MA01].

Wie alle von Montague und Aslam in [MA01, S.3] vorgestellten Verfahren ist ZMUV skalierungs- und verschiebungsinvariant.

Man beachte auch, dass ZMUV die Scores nicht in ein $[0, 1]$-Intervall abbildet, wie in Tabelle 2.3 beispielhaft gezeigt wird. Insbesondere werden bei diesem Verfahren ein großer Teil aller Scores negativ, weshalb in [WCB06, S. 2] ZMUV2 eingeführt wird, welches zum mit ZMUV normalisierten Score noch 2 hinzu addiert. Dies führt dazu, das die Werte überwiegend im $[0, 3]$-Intervall liegen, schließt aber normalisierte Werte außerhalb dieses Intervalls nicht aus.

Median-MAD-Normalisierung

Das Problem der fehlenden Robustheit von ZMUV kann behoben werden, indem robustere Parameter zur Beschreibungen der Lage und Streuung gewählt werden. Bei der Median-MAD-Normalisierung wird diese durch den Median und den Median der absoluten Abweichungen, kurz MAD, ersetzt.

$$\mathrm{MAD}(X) = \mathrm{Median}(|X - Median(X)|)$$

Die Median-MAD-Normalisierung ist deutlich robuster als ZMUV, die verwendeten Schätzer sind jedoch nicht so effizient [JNR05]. Wie ZMUV erfüllt die Median-MAD-Normalisierung alle Anforderungen nach Montague und Aslam.

NORMEXP

Viele genutzte Normalisierungsverfahren sind eher heuristisch motiviert [MS02]. Im Unterschied dazu basiert das von Manmatha und Sever [MS02] vorgestellte NORMEXP auf dem Normal-Exponential-Modell.

Das Normal-Exponential-Modell beschreibt die Score-Verteilung für einzelne Anfragen im Information-Retrieval. Das Modell geht von der im Information-Retrieval verbreiteten Annahme aus, dass Dokumente sich bezüglich einer Anfrage entweder als relevant oder als irrelevant einordnen lassen. Das Normal-Exponatial-Modell besagt, dass die Scores der relevanten Dokumente normalverteilt sind, während die Score der irrelevanten Dokumente eine Exponentialverteilung haben [ARK09, AK09]. Für kurze Text-Anfragen trifft die Normalverteilung für relevante Scores in der Praxis nicht immer zu [ARK09], wobei die Verteilung schnell mit steigender Anfragelänge gegen die Normalverteilung konvergiert [AK09]. Die Wahl der Exponentialverteilung für Scores nicht relevanter Dokumente ist bisher nur empirisch gerechtfertig [ARK09].

Da irrelevanten Dokumente nicht in Zusammenhang zu einer Anfrage an eine Suchmaschine stehen, können diese Informationen über die Verteilung von Scores von zufällig ausgewählten Dokumenten liefern. Anscheinend gehen Manmatha und Sever dabei wie Montague und Aslam davon aus, das die betrachteten Scores bereits nur von Dokumenten stammen, dessen Scores einen bestimmten Grenzwert überschreiten [MA01], wobei der Anteil der relevanten Dokumente größer als in der gesamten Dokumentkollektion angenommen werden kann.

Zusätzlich wird von der Annahme ausgegangen, dass eine gute Normalisierung eine zufällige Auswahl an Dokumenten für unterschiedliche Quellen auf die gleiche Weise abbildet. Daraus wird geschlussfolgert, dass die Angleichung der Verteilung von zufälligen Dokumenten ein gutes Mittel zur Normalisierung von Scores ist [MS02, S. 3].

Um dieses Ziel zu erreichen, wird das Minimum der Scores einer Anfrage auf 0 und der Erwartungswert auf einen festen Wert c abgebildet. Bei vorliegenden Relevanz-Informationen für die Dokumente bezüglich einer Anfrage wird dieses Verfahren als EXPML bezeichnet [MS02].

Da in der Praxis keine Einteilung der Dokumente nach Relevanz vorliegt, schlagen Manmatha und Sever 3 Methoden zur Schätzung des Erwartungswerts vor [MS02].

Abschätzung durch Mixture Model Fit Mixture Model Fit ist eine auf Expectation Maximation basierende iterative Methode, die verwendet werden kann, um die Parameter der Normalverteilung und der Exponentialverteilung aus der Gesamtverteilung zu ermitteln [MRF01]. Der so ermittelte Erwartungswert der Exponentialverteilung kann dann zur Normalisierung genutzt werden, dieses Verfahren heißt EXPEM [MS02].

s	$\text{norm}_{\text{Decimal-Scaling}}(s)$	$s+5$	$\text{norm}_{\text{Decimal-Scaling}}(s+5)$	$3 \cdot s$	$\text{norm}_{\text{Decimal-Scaling}}(3 \cdot s)$
1	0.1	6	0.06	3	0.03
5	0.5	10	0.10	15	0.15
6	0.6	11	0.11	18	0.18
9	0.9	14	0.14	27	0.27

Tabelle 2.4: Ein Beispiel für fehlende Verschiebungs- und Skalierungsinvarianz bei Verwendung von Decimal-Scaling.

Abschätzung durch die Gesamtverteilung Die Verteilung der Scores der nicht-relevanten Dokumente geht stärker in die Gesamtverteilung der Scores ein, da relevante Dokumente oftmals auch im Ergebnis einer Anfrage an eine Suchmaschine in viel geringerer Anzahl auftreten [MS02]. Daher kann man den Erwartungswert der Scores der nicht-relevanten Dokumente durch das arithmetische Mittel der Scores aller Dokumente abschätzen [MS02]. Bei Nutzung dieser Abschätzungen und einer Wahl von $c = 1$ entspricht NORMEXP dem Normalisierungsverfahren Sum bei Verwendung des arithmetischen Mittels [MS02].

Durchschnitt der vorherigen Abschätzung Manmatha und Sever stellten fest, dass der Fehler der oben genannten Abschätzungen oft in unterschiedliche Richtungen auftritt [MS02]. Deshalb wurde die Benutzung des Durchschnitts beider Abschätzungen vorgeschlagen. Mithilfe dieses, EXPARV genannten, Verfahrens wurden ähnlich gute Ergebnisse wie bei der Verwendung von EXPML erzielt [MS02].

Decimal-Scaling

Decimal-Scaling geht davon aus, dass sich die Scores aus verschiedenen Quellen nur um Zehnerpotenzen als Faktor unterscheiden. Bei diesem Verfahren wird 0 auf sich selbst und die erste Zehnerpotenz, welche größer als der größte Score ist, auf 1 abgebildet [SSK06]. Wie auch Min-Max ist dieses Verfahren sensibel gegenüber Ausreißern, bietet aber keine Skalierungs- und Verschiebungsinvarianz. Dies ist beispielhaft in Tabelle 2.4 deutlich gemacht.

2.4.2 Nichtlineare Normalisierungen

Hier wird zuerst auf zwei Normalisierungsverfahren eingegangen, die prinzipiell den linearen Normalisierungen ähnlich sind, aber durch die Anwendung einer Sigmoid-Funktion nicht linear sind. Eine Sigmoid-Funktionen ist „eine beschränkte und differenzierbare reelle Funktion mit einer durchweg positiven oder durchweg negativen ersten Ableitung und genau einem Wendepunkt“ [wikc]. Charakteristisch für eine Sigmoidfunktion ist eine „S-Form“, wie sie in Abbildung 2.1 zu sehen ist.

Es folgen drei weitere Verfahren, deren Konzept sich stärker von denen der linearen Normalisierungen unterscheidet.

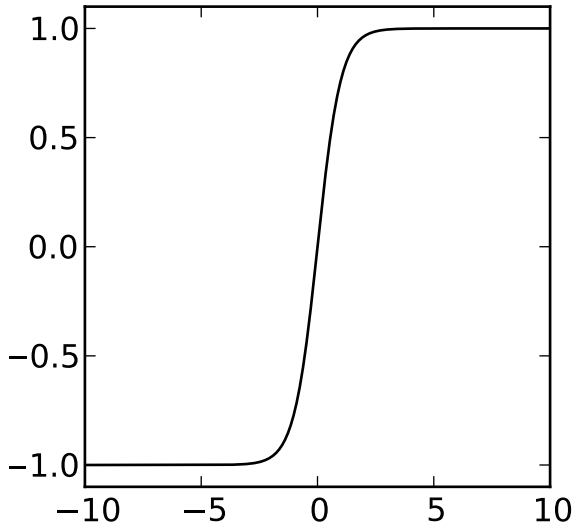

Abbildung 2.1: Die typische „S“-Form einer Sigmoid-Funktion am Beispiel von tanh

tanh-Estimator

Der tanh-Estimator basiert auf einer Normierung der Standard-Abweichung und des Erwartungswertes mit einer anschließenden Anwendung des Tangens Hyperbolicus. Es besteht damit eine starke Ähnlichkeit zu ZMUV. Durch die Verwendung des Hampel-Estimators zum Bestimmen von Erwartungswert und Standardabweichung ist dieses Verfahren robust und effizient [JNR05]. Es spricht jedoch nichts dagegen, diesen Schätzer auch in beliebigen anderen Normalisierungsverfahren zur Schätzung der Parameter einzusetzen.

$$s' = \frac{1}{2}\left(\tanh\left(0.01\left(\frac{s-\mu}{\sigma}\right)\right)+1\right) \tag{2.2}$$

Double-Sigmoid Normalisierung

Dieses in [JNR05] beschrieben Verfahren zielt auf eine klare Trennung von Scores von relevanten und irrelevanten Dokumenten. Dafür wird eine Sigmoid-Funktion verwendet, die aus zwei Teilen logistischer Funktionen zusammengesetzt ist.

$$\text{norm} = \begin{cases} \frac{1}{1+e^{-2*(s-t)/r_1}} & \text{wenn } s < t \\ \frac{1}{1+e^{-2*(s-t)/r_2}} & \text{sonst} \end{cases} \tag{2.3}$$

t wird als Referenzpunkt bezeichnet. Wie unschwer zu erkennen ist, wird ein Score $s = t$

immer auf $\frac{1}{2}$ abgebildet. r_1 und r_2 bestimmen mit $(t - r_1, t - r_2)$ das Intervall der Scores, in denen die Sigmoid-Funktion annähernd linear verläuft [JNR05].

Um t, r_1, r_2 sinnvoll auswählen zu können, muss die Verteilung der normalisierenden Scores bekannt sein. t wird gewöhnlich in der Überlappung von relevanten und nicht relevanten Scores gewählt und r_1, r_2 jeweils an den Kanten der Überlappung [JNR05]. Dieses Vorgehen scheint allerdings nur möglich zu sein, wenn die Verteilungen der Scores relevanter und irrelevanter Dokumente relativ gut zu trennen sind.

Aggregate Historical CDF

Fernándes, Vallet und Castells schlagen in [FVC06] eine Form der Normalisierung vor, die Scores aus verschiedenen Quellen so abbildet, dass sie eine bestimmten „ideale“ Verteilung besitzen.

Der Name der „Aggregate Historical CDF“ stammt aus [AK09] und weist auf den Umstand hin, dass die Verteilungen der Scores S_{HIS} aus den einzelnen Quellen über verschieden „historische“ Anfragen ermittelt werden.

Diese ideale Verteilung S_{ideal} ist die Verteilung einer Scoring-Funktion, welche „den Rang eines Ergebnisses auf die Relevanz abbildet“ [FVC06] und wird durch die durchschnittliche Verteilung mehrerer guter Scoring-Systemen geschätzt. Dazu werden die Ausgaben der Systeme für verschiedene Anfragen mit Min-Max normalisiert und dann gesammelt. Die Verteilung all dieser Scores wird als Schätzung der idealen Verteilung betrachtet.

Die Normalisierung erfolgt in der Form $\text{norm}_{\text{HIS}} = F^{-1}(D(s))$, wobei $D(s) = P(S_{\text{HIS}} \leq s)$ die kumulative Verteilungsfunktion der Scores aus dieser Quelle ist und $D(s) = P(S_{\text{ideal}} < s)$ die kumulative Verteilungsfunktion der idealen Verteilung ist.

Da F und D kumulative Verteilungsfunktionen sind, ist dieses Normalisierungsverfahren monoton steigend. Da die Eingabescores für die Ermittelung von F Min-Max-normalisiert sind, liegen auch die normalisierten Scores im Intervall $[0, 1]$ [FVC06].

Arampatzis und Kamps stellen sich mit „Aggregate Historical CDF Simplified“ eine einfachere Variante vor, die auf die Anwendung von F^{-1} verzichtet, stellen aber heraus, dass dies auf bei nicht Rang-basierenden Fusionsmethoden durchaus negativen Einfluss auf das Gesamtergebnis haben könnte [AK09]. Zudem äußern sie Kritik an der unklaren Definition der idealen Scoreverteilung.

Signal-to-Noise

Avampatzis und Kamps stellen in [AK09] drei verschieden Normalisierungsverfahren vor, die alle auf der Annahme basieren, dass sich ein Score-Wert eine Summe aus einer Signal- und einer Noise-Komponente zusammensetzen, wobei das Verhältnis der Summanden nur von dem Gesamt-Score abhängt. Die Signal-Komponente ist dabei semantisch bedeutsam, während die Noise-Komponente keine semantische Bedeutung hat.

In [AK09] werden drei verschieden Methoden vorgestellt, die auf dem Signal-to-Noise-Ansatz basieren: norm_{SN1} bildet Scores nur auf das vermutete Verhältnis von Signal und Noise ab, berücksichtigt dabei aber nicht die Höhe des ursprünglichen Scores. Bei

norm_{SN2} fließt die Höhe des Scores über eine kumulative Verteilungsfunktion von historischen Anfragen ein. norm_{SN3} unterscheidet sich durch die Verwendung der Verteilung der Signal-Komponente anstelle der historischen Scoreverteilungen von norm_{SN2}.

$$\begin{aligned}
\text{norm}_{\text{SN1}} &= \frac{p_{\text{signal}}(s)}{p_{\text{signal}}(s) + p_{\text{noise}}(s)} \\
\text{norm}_{\text{SN2}} &= \frac{p_{\text{signal}}(s)}{p_{\text{signal}}(s) + p_{\text{noise}}(s)} \cdot P(S_{\text{HIS}} \geq s) \\
\text{norm}_{\text{SN2}} &= \frac{p_{\text{signal}}(s)}{p_{\text{signal}}(s) + p_{\text{noise}}(s)} \cdot P(S_{\text{Signal}} \geq s)
\end{aligned} \tag{2.4}$$

p_{signal} und p_{noise} sind dabei Dichtefunktionen der Verteilungen von der Signal- und der Noise-Komponente. Es stellt sich natürlich die Frage, wie p_{signal} und p_{noise} bestimmt werden können. Avampatzis und Kamps verwenden dafür eine Menge künstlicher Anfragen an dass zu untersuchende Retrieval-System und erhalten dabei Scoreverteilungen für sinnvolle Anfragen, die vor allem die Signal-Komponente enthalten und zufällige Anfragen, von denen angenommen wird, das sie die Noise-Komponente repräsentieren.

Um p_{noise} zu approximieren, stellen Avampatzis und Kamps zufällige Anfragen an das Retrieval-System. Diese erhalten sie, indem sie zufällig gleichverteilt Wörter aus der Anfrage-Sprache wählen, wobei ein zusätzliches Wort die Anfrage beendet. p_{signal} approximieren Avampatzis und Kamps, in dem sie die Verteilung von Anfragen ermitteln, wie sie Menschen formulieren könnten. Dazu verwenden sie Textfragment aus der Dokumentkollektion, wobei die Verteilung der Länge dieser Fragmente durch eine Zipf'schen Verteilung approximiert wird.

3 Einbindung der Kalibrierung in die Abfragebearbeitung

Die vier geforderten Eigenschaften für ein Ähnlichkeitsmaß s, dass für die CQQL-Auswertung genutzt werden kann, sind:

1. ein Wertebereich, der im Intervall $[0, 1]$ liegt: $\forall a, b \in \mathrm{Dom}(A) : s(a = b) \in [0, 1]$
2. die Feststellung der Identität: $\forall a \in \mathrm{Dom}(A) : s(a = a) = 1$
3. Symmetrie: $\forall a, b \in \mathrm{Dom}(A) : s(a = b) = s(b, a)$
4. Die Ähnlichkeitsmatrizen von s sind positiv semidefinit.

Obwohl nicht klar ist, ob praktisch für die CQQL-Auswertung genutzten Ähnlichkeitsmaße die Eigenschaften 2. bis 4. besitzen, sollte versucht werden, diese bei der Kalibrierung zu erhalten.

3.1 Wertebereich im Intervall $[0, 1]$

Die im zweiten Kapitel vorgestellten Verfahren sind nicht direkt für die Verwendung mit Ähnlichkeitswerten entwickelt worden und können dazu führen, dass Ähnlichkeitswerte auf Werte außerhalb des Intervalls $[0, 1]$ abgebildet werden. Dies trifft nicht auf die vorgestellten nicht-linearen Normalisierungen zu, da diese nach ihrer jeweiligen Definition auf einen Wertebereich abbilden.

Auch Min-Max und Fitting bilden nach ihrer Definition immer in das Intervall $[0, 1]$ ab. Dies ist allerdings nicht mehr sichergestellt, falls Minimum und Maximum nur durch beispielhafte Anfragen ermittelt werden. Ausreißer, die außerhalb der so ermittelten Grenzen liegen können, werden dann insbesondere von Min-Max auf Werte außerhalb des Intervalls abgebildet.

Sum, ZMUV, ZMUV2, MAD, NORMEXP bilden ihrer Definition nach nicht in das gewünschte Intervall ab. Durch geschickte abweichende Wahl von $y_{\min}$ und $y_{\max}$ lässt sich erreichen, dass nur eine kleine Anzahl von Ausreißer auf Werte außerhalb von $[0, 1]$ abgebildet wird.

Am Beispiel lässt sich dies leicht an ZMUV zeigen. Dafür setzt man $y_{\min} = 0.5$ und $y_{\max} = 0.5 + \frac{0.5}{t}$ und $t > 1$. Die normalisierten Ähnlichkeitswerte besitzen dann einen Erwartungswert μ von 0.5 und eine Standardabweichung σ von $\frac{0.5}{t}$. Die Tschebyscheff-Ungleichung [wikd] kann genutzt werden, um eine oberere Schranke für die Wahrscheinlichkeit zu finden, dass ein Wert nach außerhalb des Intervalls $[0, 1]$ abgebildet wird:

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \tag{3.1}$$

Setzt man die Parameter ein und $k = 0.5$ als Abstand der Intervallgrenze zu μ, dann erhält man:

$$\begin{aligned} P(|X - 0.5| \geq 0.5) &\leq \frac{\left(\frac{0.5}{t}\right)^2}{0.5^2} \\ &= \frac{1}{t^2} \end{aligned} \tag{3.2}$$

Setzt man also $t = 5$ ein, erhält man $y_{\min} = 0.5$ und $y_{\max} = 0.1$ und eine maximale Wahrscheinlichkeit von $\frac{1}{25}$, dass ein Wert außerhalb des $[0, 1]$-Intervalls liegt. Dabei sei angemerkt, dass die Tschebyscheff-Ungleichung meist eine schwache Äbschätzung ist [wikd], daher sollten solche Ausreißer bedeutend seltener auftreten.

Werte außerhalb des $[0, 1]$-Intervalls sind nicht nur durch formell korrekte CQQL-Auswertung unmöglich zu erzielen, sondern können auch nicht sinnvoll durch die Auswertungen der CQQL-Operatoren verknüpft werden. Als Beispiel sei die Disjunktion genannt: $\text{eval}(2 \vee 2) = 2 + 2 - 2 \cdot 2 = -2$.

Um solche Ausreißer zu vermeiden, kann man Werte außerhalb von $[0, 1]$ auf die jeweilige Grenze abbilden:

$$\text{clip}(s) = \begin{cases} 0 & \text{wenn } s < 0 \\ 1 & \text{wenn } s > 1 \\ s & \text{sonst} \end{cases} \tag{3.3}$$

Es ist offensichtlich, dass dieses Vorgehen zu einem Informationsverlust bei Scores führt, die außerhalb von $[0, 1]$ liegen.

Eine sanftere Lösung für das Problem ist der Einsatz einer Sigmoid-Funktion, wie sie beim Double-Sigmoid-Verfahren und und dem tanh-Estimator benutzt wird.

Die für diese zwei Lösungen genutzten Funktionen sind nicht linear, so dass Kalibrierungen, die sie verwenden, auch nicht linear sind. Allerdings betrifft dies im Fall der clip-Funktion nur Ausreißer, die sonst die gesamte Auswertung kontaminieren würden. Es ist auch möglich, die Sigmoid-Funktionen durch ihre Parameter so einzustellen, dass sich ihre nicht-Linearität nur auf Ausreißer stark auswirkt.

3.2 Feststellung der Identität

Auch dass Erhalten dieser Eigenschaft ist bei den im zweiten Kapitel vorgestellten Verfahren nicht unbedingt gegeben. Für die meisten Verfahren ist es kaum möglich, allgemein eine Aussage zum Erhalt dieser Eigenschaft zu treffen. Aggregate Historical CDF und Min-Max erhalten diese Eigenschaft, wenn in den zum Bestimmen der Parameter verwendeten Ähnlichkeitswerte zumindest ein einziger Wert von 1 enthalten ist. Im Unterschied dazu bilden das Double-Sigmoid-Verfahren und und der tanh-Estimator niemals einen Score auf 1 ab, so dass sie diese Eigenschaft auf jeden Fall zerstören.

3.3 Symmetrie

Viele der linearen Normalisierungsverfahren, die in der Literatur verwendet werden, ermitteln die Normalisierungsparameter für jede Anfrage neu werden.

Das Ermitteln der Parameter pro Anfrage kann dabei dazu führen, dass die zweite Eigenschaft zerstört wird. Dazu kann beispielsweise Min-Max betrachtet werden. Man stelle sich eine Menge von drei Attributwerten a_1, a_2, a_3 vor, wobei sich a_1 und a_2 sehr ähnlich sind, a_3 aber beiden anderen Werten unähnlich ist:

- $s(a_1, a_2) = s(a_2, a_1) = 0.9$
- $s(a_1, a_3) = s(a_3, a_1) = 0.2$
- $s(a_2, a_3) = s(a_3, a_2) = 0.1$

Für eine Anfrage alle drei Werte mit der Anfrage verglichen, dabei wird Min-Max zur Normalisierung eingesetzt. Betrachten wir eine Anfrage nach $A = a_1$. Die nicht normalisierten Ergebnisse sind: $a_1 \to 1$, $a_2 \to 0.9$, $a_3 \to 0.1$. Normalisiert man dieses Ergebnis mit Min-Max, erhält man: $a_1 \to 1$, $a_2 \to 0.875$, $a_3 \to 0$ Nun führt man eine ähnliche Anfrage mit $A = a_3$ durch. Die nicht normalisierten Ergebnisse sind: $a_1 \to 0.2$, $a_2 \to 0.1$, $a_3 \to 1$. Normalisiert man dieses Ergebnis mit Min-Max, erhält man: $a_1 \to 0.112$, $a_2 \to 0$, $a_3 \to 1$. Damit würde gelten $s_{\text{norm}}(a_1, a_3) = 0 \neq 0.112 = s_{\text{norm}}(a_3, a_1)$ und die zweite Eigenschaft ist durch die Normalisierung zerstört. Man kann leicht erkennen, dass dieses Problem bei fester Wahl der Kalibrierungsparametern für alle Anfragen nicht auftreten kann.

4 Kalibrierungsproblem

In diesem Kapitel soll das Kalibrierungsproblem bei der CQQL-Auswertung genauer betrachtet werden. Zuerst wird kurz auf die bei der Verwendung von Fuzzy-Operatoren auftretende Dominanz als Problem des fehlendes Einflusses eines Änlichkeitswertes betrachtet. Dieses Problem ist bereits hinreichend bekannt. Dann soll Einfluss von unterschiedlichen Score-Verteilungen auf die Ergebnisse der Auswertung von CQQL-Operatoren gemessen werden, um das Kalibrierungsproblem zu beschreiben. Obwohl das Kalibrierungsproblem durchaus störenden Einfluss auf die Auswertung komplexerer Formeln haben kann, werden hier nur einfache Formeln betrachtet, die entweder aus einer Konjunktion ($\wedge$) oder Disjunktion ($\vee$) zweier Ähnlichkeitswerte bestehen. Im Nachfolgenden wird als Operation $\circ$ notiert, wenn wahlweise eine Konjunktion oder Disjunktion eingesetzt werden kann.

4.1 Dominanz bei Fuzzy-Operatoren

Die Auswertung der Fuzzy-Operatoren erfolgt durch Berechnung des Minimums für die Konjunktion und Berechnung des Maximums für die Disjunktion. Wie in Tabelle 4.1 zu sehen, werden boolesche Werte damit semantisch korrekt ausgewertet.

Werden Fuzzy-Operatoren jedoch zur Verknüpfung von Ähnlichkeitswerten genutzt, tritt häufig der Fall auf, dass einer von den zwei Werten nicht maßgeblich in das Auswertungsergebnis eingeht. Diesen Effekt bezeichnet man als Dominanz.

Beispiel für Dominanz von Fuzzy-Operatoren: Seien a und b Ähnlichkeitswerte und $\max(a, b)$ die Auswertung einer Disjunktion.

Bei Betrachtung von Tabelle 4.2 ist zu bemerken, dass das Ergebnis der Disjunktion in der ersten und der zweiten Zeile gleich 0.9 ist, obwohl sich die Werte für a stark unterscheiden. Im Vergleich dazu unterscheidet sich das Ergebnis der Auswertung in der ersten zu der dritten Zeile, obwohl sich der Unterschied in b nur gering ist.

Wie man leicht aus der Definition des Maximums ableiten kann, dominiert bei der Anwendung der Disjunktion jeder größere Wert einen kleineren. Analog dominiert bei Betrachtung der Konjunktion jeder kleiner Wert einen größeren.

a	b	$a \wedge b$	$\min(a, b)$	$a \vee b$	$\max(a, b)$
0	0	0	0	0	0
0	1	0	0	1	1
1	0	0	0	1	1
1	1	1	1	1	1

Tabelle 4.1: Auswertung auf boolschen Werten mit Fuzzy-Operatoren

a	b	max(a,b)
0.8	0.9	0.9
0.2	0.9	0.9
0.8	0.8	0.8

Tabelle 4.2: Auswertung der Fuzzy-Disjunktion auf Ähnlichkeitswerten

a	b	$a + b - a \cdot b$
0.8	0.9	0.98
0.2	0.9	0.92
0.8	0.8	0.96

Tabelle 4.3: Die CQQL-Auswertung der Disjunktion zeigt keine Dominanz.

Es gilt also:

$$\begin{aligned} a \geq b &\Rightarrow \text{eval}_{\text{fuzzy}}(a \vee b) = \max(a, b) = a \\ a \leq b &\Rightarrow \text{eval}_{\text{fuzzy}}(a \wedge b) = \min(a, b) = b \end{aligned} \tag{4.1}$$

Bei der Verwendung der CQQL-Operatoren tritt das Problem der Dominanz nicht auf [Sch08], wie man beispielhaft in Tabelle 4.3 sehen kann.

Tatsächlich sind einzigen Fälle, in denen nur ein einzelner Operand das Ergebnis einer der Auswertung der CQQL-Operatoren festlegt, bereits durch die boolesche Semantik festgelegt.

$$\begin{aligned} a = 0 &\Rightarrow a \wedge b = a \cdot b = 0 \cdot b = 0 \\ a = 1 &\Rightarrow a \vee b = a + b - ab = 1 + b + 1 \cdot b = 1 \end{aligned} \tag{4.2}$$

Abgesehen von dieser Ausnahme gilt für alle $a, b, c \in [0, 1]$ mit $b \neq c$, dass $\text{eval}(a \circ b) \neq \text{eval}_{\circ}(a \circ c)$.

4.2 Erarbeitung des Kalibrierungsproblems

Obwohl die Dominanz bei Verwendung der CQQL-Operatoren nicht auftritt, kann bei der Betrachtung einer größeren Anzahl von Ähnlichkeitswerten ein Effekt beobachtet werden, der ähnliche Auswirkungen hat: Die Auswertungen einiger Atome haben größeren Einfluss auf die Auswertung der gesamten Anfrage als andere. Dieser Effekt lässt sich gut am Beispiel deutlich machen.

Beispiel für das Kalibrierungsproblem : In Tabelle 4.4 sind drei verschiedene Rangfolgen einer durchnummerierten Menge von Tupeln (a, b) aufgelistet, die durch Auswertungen von Atomen auf Tupeln erhalten werden könnten. Einmal sind die Tupel absteigend nach a sortiert, einmal nach b und in der letzten Spalte nach der Auswertung der Anfrage $(a \vee b)$. Der Wert,

Rang	Tupel	a	Tupel	b	Tupel	eval($a \vee b$)
1	8	0.467	6	0.568	8	0.741
2	3	0.466	8	0.514	9	0.731
3	4	0.463	9	0.505	6	0.704
4	9	0.456	5	0.461	5	0.693
5	2	0.455	7	0.408	10	0.667
6	10	0.446	10	0.400	3	0.643
7	5	0.431	1	0.344	2	0.616
8	1	0.388	3	0.330	1	0.598
9	7	0.314	2	0.297	7	0.594
10	6	0.313	4	0.237	4	0.591

Tabelle 4.4: Eine nach verschiedenen Kriterien sortierte Menge von Tupeln

nach dem sortiert wurde, ist in der jeweiligen Spalte angegeben. Die doppelte horizontale Linie trennt die obere Hälfte der Ränge von der unteren.

Betrachtet man die fünf nach a in die oberen Ränge sortierten Tupel, so fällt auf, dass nur zwei davon bei der Sortierung nach eval($a \vee b$) zu den oberen Rängen gehören. Im Vergleich dazu sind vier der fünf Tupel aus den oberen Rängen bei der Sortierung nach b auch bei einer Sortierung nach eval($a \vee b$) in den oberen Rängen vertreten. Analoge Beobachtungen ergeben sich bei der Betrachtung der jeweils unteren Ränge. Man kann also davon ausgehen, dass der Rang des Tupels auf einer nach eval($a \vee b$) sortierten Liste stärker von der Wahl von b abhängt als von der Wahl von a. Da a und b aber ohne jede Gewichtung in die Formel eingehen, ist es aus Nutzersicht zu erwarten, dass die Wahl jeder Variablen gleich starken Einfluss hat.

Um das Kalibrierungsproblem besser definieren zu können, soll der Begriff „Wahl der Ähnlichkeitswerte“ genauer beleuchtet werden. Dabei ist entscheidend, dass die betrachteten Ähnlichkeitswerte die Auswertung von Atomen sind. Im Normalfall geht man davon aus, dass die Auswertungen unterschiedlicher Atome voneinander unabhängig sind, wodurch man die untersuchten Tupel als Stichprobe einer größeren Grundgesamtheit werten kann. Die im Tupel enthaltenen Ähnlichkeitswerte können als voneinander unabhängige Zufallsvariablen mit einer jeweils für ein bestimmtes Atom typischen Verteilung aufgefasst werden. Das Stärke des Einflusses einer dieser Variablen lässt sich dann als Korrelation zur Auswertung der Formel ausdrücken.

Damit kann Kalibrierungsproblem genauer beschrieben werden: Seien A und B zwei voneinander abhängige Zufallsvariablen, dann ist A überbewertet gegenüber B, wenn eval($A \circ B$) bedeutend stärker mit eval($A \circ B$) korreliert als B. Ein Kalibrierungsfehler tritt auf, wenn eine Variable eines binären Operators überbewertet ist. Zwei Verteilungen von Ähnlichkeitswerten heißen kalibriert, wenn kein Kalibrierungsfehler auftritt. Das Kalibrierungsproblem entspricht dem Umstand, dass Kalibrierungsfehler bei der Benutzung von Operatoren auftreten.

An dieser Stelle sei angemerkt, dass Dominanz bei Operatoren immer auch zu Kalibrierungsfehlern bei bestimmten Kombinationen von Verteilungen führt. Aus zwei nichtleeren Mengen M_1, M_2 von Ähnlichkeitswerten mit $\forall m_1 \in M_1, m_2 \in M_2 : m_1$ dominiert$_\circ$ m_2

erhält man leicht durch gleichverteilte zufällige Wahl von A aus M_1 und B aus M_2 eine Überbewertung von A gegenüber B.

Beispiel für das Kalibrierungsproblem bei der Fuzzy-Konjunktion: Wählt man A gleichverteilt im Interval $[0, 0.5]$ und B gleichverteilt im Intervall $(0.5, 1]$, dann korreliert A nicht mit mit $\text{eval}_{\text{fuzzy}}(A \wedge B)$ Damit ist A gegenüber B überbewertet.

4.3 Auswahl eines Korrelationsmaßes

Um das Kalibrierungsproblem bei CQQL-Operatoren genauer untersuchen zu können, soll nun ein Maß für die Korrelation ausgewählt werden. Das Vorgehen in einführenden Beispiel ist zwar einfach nachvollziehbar, erscheint aber zu ungenau für ausführlichere Untersuchungen.

Da bei der Definition der Übersensibilität bereits Zufallsvariablen verwendet werden, liegt es nahe, ein Korrelationsmaß aus der Statistik zu verwenden. Ausgehend vom einführenden Beispiel bietet sich Spearmans Rangkorrelationskoeffizient an.

$$\rho_{X,Y} = \frac{\text{Cov}\left(\text{Rang}(X) \cdot \text{Rang}(X)\right)}{\sigma_{\text{Rang}(X)} \cdot \sigma_{\text{Rang}(Y)}} \tag{4.3}$$

$$\text{Cov}(X, Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)] \tag{4.4}$$

Der Spearman-Rangkorrelationskoeffizient zum Gesamtergebnis liegt im Beispiel bei 0.200 für a und bei 0.806 für b. Dieses Maß scheint also geeignet zu sein, um das Kalibrationsproblem zu beschreiben.

Die Nutzung von Rängen zur Betrachtung des Kalibrierungsproblems hat den Vorteil, dass dafür die gleichen Informationen genutzt werden, die ein Nutzer eines Retrieval-Systems durch die Ausgabe einer sortierten Liste erhalten würde. Auch aus Anschaulichkeitsgründen würde in dieser Arbeit bei der Vorstellung des Kalibrierungsproblem bisher immer mit Rängen gearbeitet.

Die Nutzung von Rängen hat aber den Nachteils, dass Informationen über die Abstände der Ähnlichkeitswerte verloren gehen. Das diese eine semantische Bedeutung besitzen, wurde auch im Bereich der Meta-Suche festgestellt [WCB06]. Insbesondere mit Blick auf die CQQL-Formeln mit mehr als zwei Prädikaten ist es sinnvoll, dass diese Informationen auch bei der Bewertung der Übersensibilität genutzt werden. Bei der rekursiven Auswertung einer solchen Formel werden die durch Auswertungen der Teilformel errechneten Ähnlichkeitswerte genutzt. Für eine Betrachtung des Kalibrierungsproblems sollte deshalb auch diese Information genutzt werden. Deshalb schlage ich die Nutzung der Pearson-Korrelation vor, welche die lineare Abhängigkeit zwischen zwei Zufallsvariablen misst:

$$\varrho_{X,Y} = \frac{\text{Cov}\left(X \cdot Y\right)}{\sigma_X \cdot \sigma_Y} \tag{4.5}$$

Der Pearson-Korrelationskoeffizient ermittelt Korrelationswerte im Intervall $[-1, 1]$, wobei -1 eine negative linearen Abhängigkeit impliziert und 1 eine positive lineare Abhängigkeit. Ein Wert von 0 weist auf darauf hin, dass keine lineare Abhängigkeit besteht

[wikb]. Der Pearson-Korrelationskoeffizient zum Auswertungsergebnis im Beispiel liegt bei 0.188 für a und bei 0.841 für b.

Die Semantik der Konjunktion und Disjunktion implizieren dabei eine positive Abhängigkeit.

$$\begin{aligned} \text{eval}(A \vee B) &= A + B - A \cdot B \\ \text{eval}(A \wedge B) &= A \cdot B \end{aligned} \tag{4.6}$$

Allerdings kann man sich auch Operatoren vorstellen, bei denen eine negative Abhängigkeit von einem Operanten gewünscht ist. Ein Beispiel dafür ist ein „but“-Operator. Dieser hat dieselbe Semantik wie $A \wedge (\neg B)$ und wird im Information-Retrieval genutzt, um zu verhindern, das Nutzer Anfragen stellen können, die eine zu große Ergebnismenge erzeugen würden.

4.4 Graphische Darstellung von Kalibrierungen

Es wird immer die Kalibrierung von zwei Verteilungen bezüglich eines Operators betrachtet. Die zwei durch Korrelation der Verteilungen ermittelten Werte lassen sich als Punkt in $[-1, 1]^n$ auffassen. Dabei wird die Korrelation der ersten Verteilung zu Gesamtergebnis auf die horizontale Achse abgetragen und die der zweiten auf der vertikalen Achse.

Die Verteilungen sind kalibriert, wenn sie gleich stark mit der Auswertung korrelieren. Die Gerade durch den Koordinatenursprung und die für den Operator gewünschte stärksten Korrelation der Variablen kommt damit besondere Bedeutung zu. Sie ist im Diagramm als eine gestrichelte Linie gekennzeichnet und wird im Weiteren als Kalibrierungslinie bezeichnet. Bei allen Punkten unterhalb dieser Linie besitzt die Wahl der ersten Variable mehr Einfluss, bei allen oberhalb die Wahl der anderen Variablen.

4.5 Definition eines Kalibrierungsfehlers

Die graphische Darstellung der Kalibrierung zweier Verteilungen eignet sich gut dazu, um die Kalibrierung einzelner Paare von Verteilungen sichtbar zu machen. Auch das Vergleichen der Kalibrierung einer Referenzverteilung zu mehreren anderen ist damit einfach möglich. Sollen viele verschiedene Paare von Verteilungen untersucht und deren Kalibrierung verglichen werden, kann diese Darstellung jedoch schnell unübersichtlich werden.

Es ist klar, dass das Ergebnis der Auswertung nicht vollständig von beiden Variablen linear abhängen kann. Wenn die Korrelation einer der beiden unabhängigen Variablen zunimmt, nimmt die Korrelation der anderen Variable ab. In Abbildung 4.2 ist zu beobachten, dass die Punkte in der Darstellung ungefähr den gleichen Abstand vom Koordinatenursprung haben. Das führt zu der Annahme, dass der Abstand zum Koordinatenursprung wenig Informationen über die Stärke des Kalibrierungsfehlers enthält.

Ich schlage vor, dass ein vorzeichenbehaftere Winkel zwischen der Kalibrierungslinie und dem Ortsvektor einer Kalibrierung zur Bestimmung des Kalibrierungsfehler genutzt

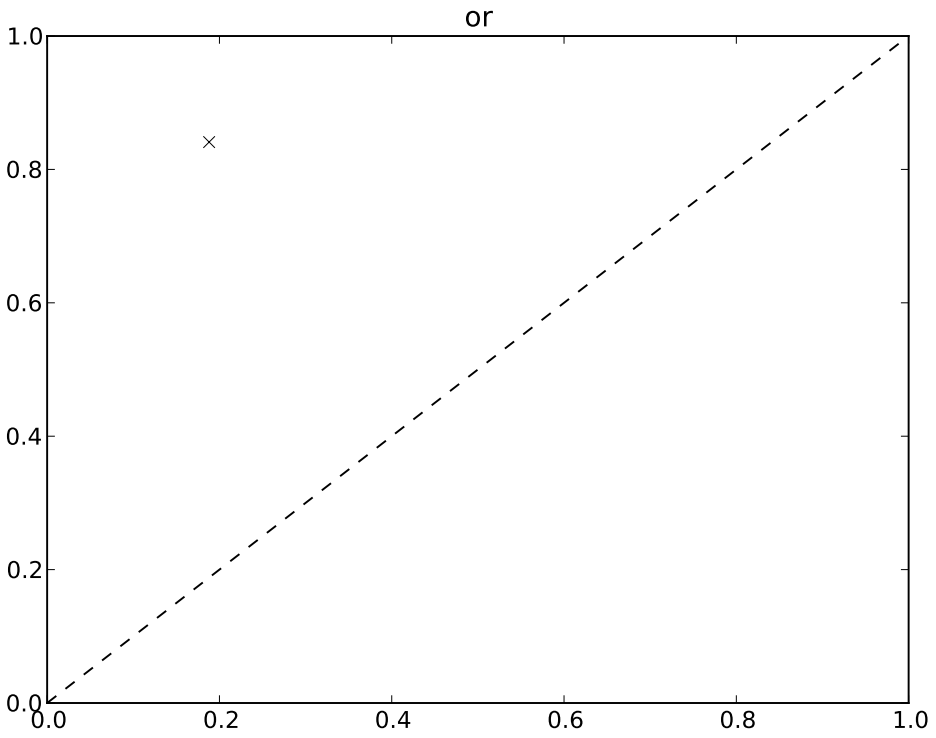

Abbildung 4.1: Graphische Darstellung der Kalibrierung der Verteilung des Beispiels mit a auf der horizontalen Achse und b auf der vertikalen Achse.

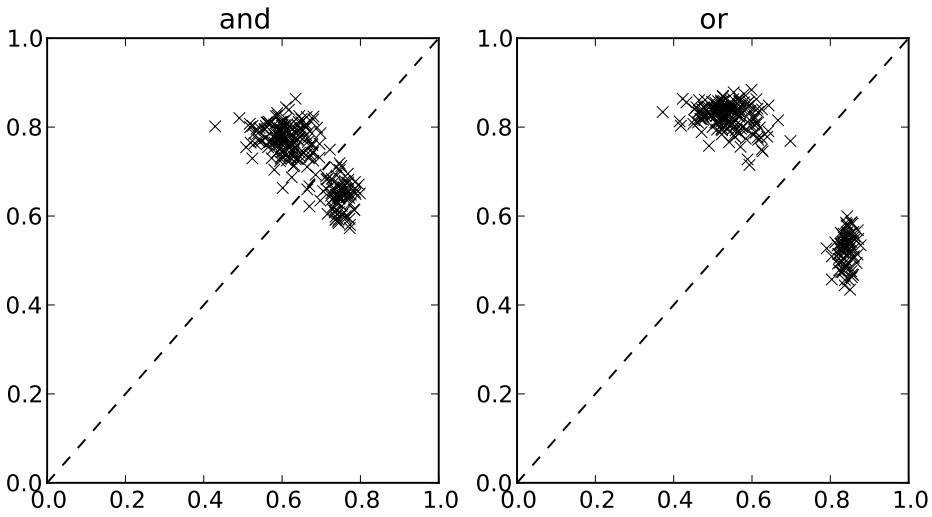

Abbildung 4.2: Kalibrierung verschiedener Paare von Verteilungen

wird. Bei einem Fehler von 0 sind zwei Verteilung kalibriert, ein positiver Winkel kennzeichnet einen Kalibrierungsfehler zugunsten der ersten Verteilung, ein negativer Winkel kennzeichnet einen Kalibrierungsfehler zugunsten der zweiten Verteilung. Der Kalibrierungsfehler soll immer im Bereich zwischen -1 und 1 liegen.

$$\mathrm{kal}_\circ(A, B) = 1 - \frac{4}{\pi} \arccos\left(\frac{\varrho_{A,\mathrm{eval}(A \circ B)}}{\sqrt{\varrho^2_{A,\mathrm{eval}(A \circ B)} + \varrho^2_{B,\mathrm{eval}(A \circ B)}}} \right) \tag{4.7}$$

Gleichung 4.7 erfüllt für die Konjunktion und Disjunktion diesen Zweck. Dabei wird erst der Winkel zwischen der horizontalen Achse und dem Ortsvektor der Kalibrierung bestimmt. Dieser liegt zwischen 0 und $\frac{\pi}{2}$. Die Kalibrierungslinie und die horizontalen Achse schließen einen Winkel von $\frac{\pi}{4}$ ein.

Der Kalibrierungsfehler für das Beispiel liegt bei $\mathrm{kal}_\vee(a, b) = -0.720$. Es an diesem negativen Wert erkennbar, dass a deutlich weniger in das Auswertungsergebnis eingeht.

4.6 Untersuchung der CQQL-Operatoren

Da eine häufige genannte Anforderungen an Normalisierungsverfahren die Angleichung der Verteilungen ist, soll hier getestet werden, ob Unterschiede in den Verteilungen einen großen Einfluss auf die Kalibrierung haben.

Dafür wird eine Stichprobe von Ähnlichkeitswert-Tupeln mit der entsprechenden Auswertung eines Operators darauf erzeugt. Dann kann die Pearson-Korrelation der Zufallsvariablen zu dem Ergebnis bestimmt werden.

Die folgenden Eigenschaften sollen dabei untersucht werden:

- die Form der Verteilung
- der Erwartungswert als Lageparameter der Verteilung
- die Standardabweichung als Skalierungparameter der Verteilung

Folgende drei Verteilungsformen erscheinen dabei besonders interessant: Die Exponentialverteilung, da sie die im Information-Retrieval auftretende Normal-Exponential-Verteilung approximiert [MS02]. Die Normalverteilung, da diese oft anzutreffende Verteilung nicht nur häufig als Modell für die Verteilung der relevanten Ergebnisse einer Anfrage auftritt [AK09], sondern auch bei Anfragen auf entsprechend verteilte Attributwerte erhalten werden kann. Eine Gleichverteilung kann leicht durch die Nutzung von normierten Rängen als Ähnlichkeitswerte entstehen und ist damit die dritte zu untersuchende Verteilungsform.

Die Standardabweichung und der Erwartungswert werden so festgelegt, dass wenige Ähnlichkeitswerte außerhalb des $[0, 1]$-Intervalls auftreten. Treten solche Werte auf, werden sie auf die jeweilige Grenze abgebildet, was zu Verzerrungen der Verteilung führen kann. Betrachtet wird damit ein Erwartungswert von $0.4, 0.5$ und 0.7 und eine Standartabweichung von $0.05, 0.10$ und 0.15.

Da die Tests auf einer Stichprobe von jeweils 25000 Wertepaaren ausgeführt werden, kann es zu Abweichungen vom erwarteten Ergebnis kommen. Deshalb werden auch der Kalibrierungsfehler von gleichen Verteilungen ermittelt, um einen Anhaltspunkt zur Einschätzung des Fehlers zu erhalten.

4.6.1 Abweichende Form der Verteilung

Um die Kalibrierung der verschiedenen Verteilungsformen zu testen, wird für beiden Verteilungen ein Erwartungswert von 0.5 und eine Standardabweichung von 0.10 benutzt.

Dann werden drei Kombinationen von Verteilungen untersucht:

- Normalverteilung und Gleichverteilung
- Normalverteilung und Exponentialverteilung
- Gleichverteilung und Exponentialverteilung

Exakt gleiche Verteilungen korrelieren nach Definition des Pearson-Korrelationskoeffizienten gleich stark mit der Auswertung, somit sind sie immer kalibriert. Deshalb sind sind sie nicht in der grafischen Darstellung vertreten.

In Abbildung 4.3 kann man zwei Beobachtungen machen: Normal- und Exponentialverteilung sind annähernd kalibriert. Die Gleichverteilung ist mit Normal- und Exponentialverteilung zu Ungunsten der Gleichverteilung nicht kalibriert. Im Vergleich zu Abbildung 4.4 und 4.5 ist zu beobachten, das die Stärke des Kalibrierungsproblem variieren kann, wenn für beide Verteilungen andere Parameter festgelegt werden.

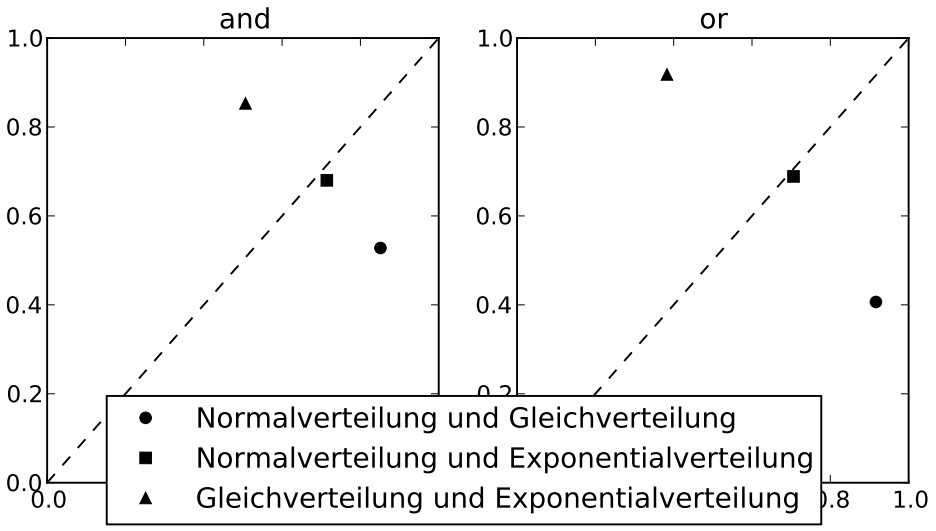

Abbildung 4.3: Kalibriertheit verschiedener Verteilungen mit $\mu = 0.5$ und $\sigma = 0.10$

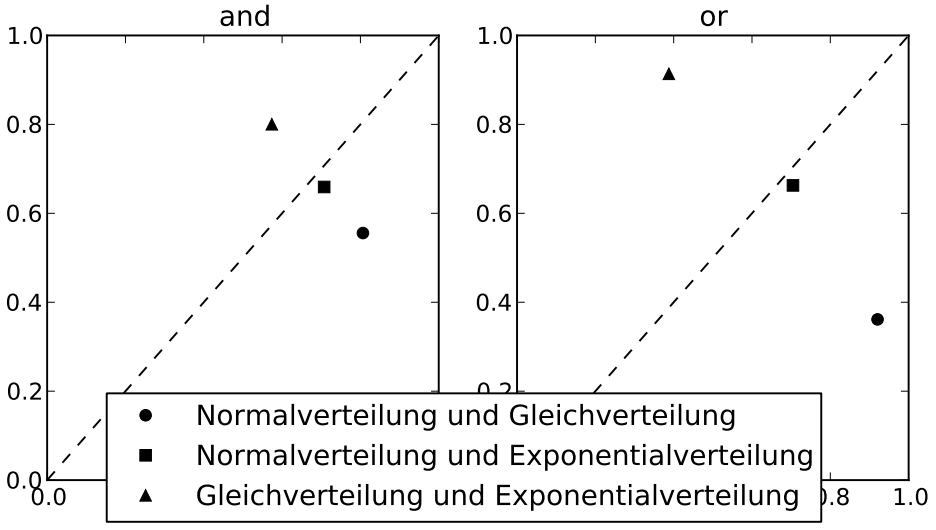

Abbildung 4.4: Kalibriertheit verschiedener Verteilungen mit $\mu = 0.5$ und $\sigma = 0.15$

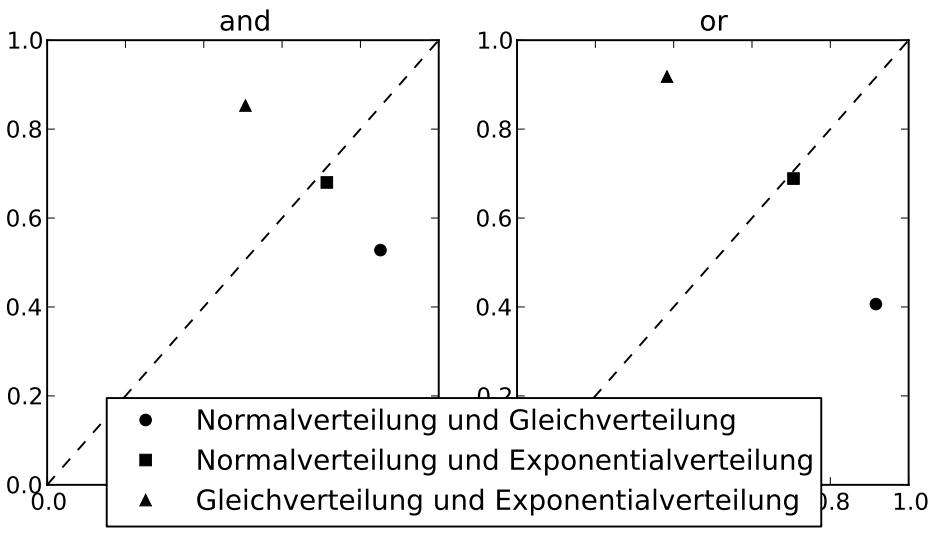

Abbildung 4.5: Kalibriertheit verschiedener Verteilungen mit $\mu = 0.4$ und $\sigma = 0.10$

Der Test bestätigt, dass bei gleichem Erwartungswert und Standardabweichung unterschiedliche Verteilungsformen zu Kalibrierungsfehlern führen können. Es es aber auch sichtbar, dass verschiedene Verteilungsformen nicht zwangsläufig zu Kalibrierungsfehlern führen müssen. Die Wahl der für beide Verteilungen identischen Parameter Erwartungswert und Standardabweichung kann die Stärke dieses Effektes verändern.

4.6.2 Abweichender Erwartungswert

Hier wird der Kalibrierungsfehler zwischen Verteilungen mit gleicher Form und Standardabweichung, aber einem unterschiedlichen Erwartungswert untersucht.

Aufgrund der Menge der verschiedenen Kombinationen von drei Erwartungswerte für alle Verteilungen und beide Operatoren wird hier auf eine grafische Darstellung verzichtet. Anstelle dessen wird der Kalibrierungsfehler in Tabelle 4.5 dargestellt. Es wird immer der Kalibrierungsfehler der Verteilungen der Zeile zur Verteilung der Spalte angegeben. Die Zahlen sind auf 3 Nachkommastellen gerundet.

Die angegebenen gemessenen Kalibrierungsfehler für exakt gleiche Verteilungen liegen nicht genau bei 0 und liefern einen Anhaltspunkt für die Genauigkeit der Messung mit der genutzten Stichprobengröße. Die größte Abweichung von 0.011 tritt dabei beim $\vee$-Operator bei Verwendung zweier Normalverteilungen mit $\mu = 0.6$ auf.

Insgesamt bestätigt sich, dass Unterschiede im Erwartungswert einen Kalibrierungsfehler hervorrufen können. Größere Abweichungen des Erwartungswerts führen bei allen untersuchten Verteilungsformen und beiden Operatoren zu größeren absoluten Kalibrierungsfehlern.

Konjunktion und Disjunktion unterscheiden sich aber in der Richtung des Fehlers. Bei Benutzung des Konjunktion ist die Verteilung mit dem geringenen Erwartungswert über-

$\sigma = 0.1$		Konjunktion			Disjunktion		
	μ	0.4	0.5	0.6	0.4	0.5	0.6
Normalverteilung	0.4	0.007	0.136	0.251	0.004	-0.122	-0.248
	0.5	-0.142	0.005	0.113	0.119	0.005	-0.134
	0.6	-0.248	-0.115	-0.002	0.255	0.142	-0.011
Gleichverteilung	0.4	-0.005	0.176	0.303	-0.003	-0.097	-0.218
	0.5	-0.169	0.003	0.133	0.102	0.005	-0.124
	0.6	-0.307	-0.136	-0.006	0.213	0.118	-0.005
Exponentialverteilung	0.4	0.007	0.148	0.262	0.007	-0.109	-0.234
	0.5	-0.134	0.005	0.11	0.116	0.003	-0.148
	0.6	-0.266	-0.125	-0.002	0.223	0.133	0.007

Tabelle 4.5: Ergebnisse des Tests mit abweichendem Erwartungswert

bewertet, die Benutzung der Disjunktion führt zu einem Kalibrierungsfehler zugunsten der Verteilung mit dem größeren Erwartungswert.

4.6.3 Abweichende Standardabweichung

Hier wird der Kalibrierungsfehler zwischen Verteilungen mit gleicher Form und gleichem Erwartungswert, aber unterschiedlicher Standardabweichung betrachtet.

Wie auch im vorherigen Test entspricht der angegebene Wert dem Kalibrierungsfehler von der Verteilung aus der Zeile zu der Verteilung aus der Spalte und ist auf 3 Nachkommastellen gerundet.

Auch hier sind die gemessenen Werte nicht genau, der größte zwischen gleichen Verteilungen gemessene Fehler liegt bei 0.13 für zwei Exponentialverteilungen mit $\sigma = 0.10$.

Der Test bestätigt, dass eine unterschiedliche Standardabweichung bei gleichen Verteilungsformen und Erwartungswert zu Kalibrierungsfehlern führen kann. Es ist deutlich erkennbar, dass eine Verteilung mit größerer Standardabweichung bei allen Verteilungen und bei beiden Operatoren stärkeren Einfluss hat und das der absolute Kalibrierungsfehler mit dem Unterschied in der Standardabweichung ansteigt.

4.6.4 Auswertung

Die durchgeführten Test haben bestätigt, dass sowohl unterschiedliche Verteilungsformen als auch Unterschiede in Erwartungswert und Standardabweichung zu Kalibrierungsfehlern führen können.

Ersteres bedeutet vor allem, dass das Kalibrierungsproblem nicht durch einfaches Normieren von Durchschnitt und Standardabweichung zu lösen ist, wenn unterschiedliche Formen von Verteilungen auftreten. Bei gleichen Verteilungsformen genügt ein solches Verfahren, um das Kalibrierungsproblem in seiner hier verwendeten Definition zu beseitigen. Die in Kapitel 2. vorgestellten Normalisierungsverfahren gehen meist von dieser Voraussetzung aus.

$\mu = 0.5$		Konjunktion			Disjunktion		
	σ	0.05	0.10	0.05	0.05	0.10	0.15
Normalverteilung	0.05	-0.0	-0.415	-0.6	-0.002	-0.415	-0.6
	0.1	0.412	-0.003	-0.258	0.412	-0.001	-0.259
	0.15	0.574	0.246	-0.001	0.576	0.25	-0.002
Gleichverteilung	0.05	-0.001	-0.472	-0.674	-0.003	-0.381	-0.541
	0.1	0.461	-0.004	-0.313	0.371	-0.005	-0.207
	0.15	0.656	0.317	0.001	0.524	0.21	0.001
Exponentialverteilung	0.05	0.003	-0.402	-0.55	0.004	-0.397	-0.547
	0.1	0.402	-0.009	-0.211	0.403	-0.013	-0.216
	0.15	0.575	0.238	0.005	0.572	0.235	0.006

Tabelle 4.6: Ergebnisse des Tests mit abweichender Standardabweichung

Das zweite Ergebnis dieser Betrachtung ist die Tatsache, dass eine größere Standardabweichung bei Konjunktion und Disjunktion zu einer größeren Wichtung einer Variablen führt, ein größerer Erfahrungswert aber zu einer einer kleineren Wichtung bei einer Konjunktion und einer größeren Wichtung bei einer Disjunktion führt. Eine Begründung hierfür kann man in der Semantik der Operatoren suchen, wobei die Disjunktion tendenziell einen größeren Operanden stärker gewichtet und die Konjunktion einen größeren Operanden.

5 Zusammenfassung und Ausblick

In dieser Arbeit wurden zuerst verschiedene Ansätze zur Normalisierung sowie darauf aufbauende Normalisierungsverfahren aus der Literatur vorgestellt. Danach wurde kurz auf die Anpassung dieser Verfahren zur Nutzung in der CQQL-Anfrageverarbeitung eingegangen und festgestellt, dass es nicht trivial ist, Kalibrierungsverfahren zu finden, welche die formal geforderte Eigenschaften für Ähnlichwerte für die CQQL-Auswertung erhalten.

Dann wurde das Kalibrierungsproblem als Folge der unterschiedlich starken Korrelation verschieden verteilter Zufallsvariablen mit der Auswertung von Disjunktion und Konjunktion auf diesen Variablen vorgestellt und eine Möglichkeit erarbeitet, einen Kalibrierungsfehler zu messen.

Mithilfe des angesprochenen Maßes wurden verschiedene Kombinationen von Verteilungen untersucht und festgestellt, dass allein durch Angleichung von Erwartungswert und Standardabweichung der Verteilungen der Kalibrierungsfehler nicht zu beheben ist.

Es bleibt daher zu untersuchen, ob Kalibrierung mittels linearer Verfahren überhaupt möglich ist. Vorstellbar wäre zum Beispiel, die Minimierung des Kalibrierungsfehlers durch die Wahl verschiedener Verteilungsparameter.

Da die verwendete Definition des Kalibrierungsproblems sich allein auf die Beobachtung der Einflussstärke einzelner Ähnlichkeitswerte und nicht auf deren semantische Bedeutung beruht, wäre zu überprüfen, ob die Nutzung von kalibrierten Verteilungen überhaupt die Qualität von Anfrageergebnisse verbessert.

Da die Nutzung kalibrierter Verteilungen eine ungewollte Überbewertung einer Variablen vorbeugen soll, bietet sich das Maß für den Kalibrierungsfehler möglicherweise auch für die Untersuchung der Kalibrierung gewichteter CQQL-Operatoren an.

Literaturverzeichnis

[AK09] Arampatzis, A. ; Kamps, J.: A signal-to-noise approach to score normalization. In: *Proceedings of the 18th ACM conference on Information and knowledge management*. New York, NY, USA : ACM, 2009 (CIKM '09). – ISBN 978–1–60558–512–3, 797–806

[ARK09] Arampatzis, A. ; Robertson, S. ; Kamps, J.: Score distributions in information retrieval. In: *Advances in Information Retrieval Theory* (2009), S. 139–151

[FVC06] Fernández, M. ; Vallet, D. ; Castells, P.: Probabilistic Score Normalization for Rank Aggregation. Version: 2006. http://dx.doi.org/10.1007/11735106_63. In: Lalmas, Mounia (Hrsg.) ; MacFarlane, Andy (Hrsg.) ; Rüger, Stefan (Hrsg.) ; Tombros, Anastasios (Hrsg.) ; Tsikrika, Theodora (Hrsg.) ; Yavlinsky, Alexei (Hrsg.): *Advances in Information Retrieval* Bd. 3936. Springer Berlin / Heidelberg, 2006. – ISBN 978–3–540–33347–0, 553-556

[Ham80] Hampel, F.: Robuste Schätzungen: Ein anwendungsorientierter Überblick. In: *Biometrical Journal* 22 (1980), Nr. 1, 1–21. http://dx.doi.org/10.1002/bimj.4710220102. – DOI 10.1002/bimj.4710220102. – ISSN 1521–4036

[Hen08] Henrich, A.: *Information Retrieval 1: Grundlagen, Modelle und Anwendungen*. Jan. 2008. – Version 1.2

[JNR05] Jain, A. ; Nandakumar, K. ; Ross, A.: Score normalization in multimodal biometric systems. In: *Pattern Recognition* 38 (2005), Nr. 12, 2270 - 2285. http://dx.doi.org/10.1016/j.patcog.2005.01.012. – DOI 10.1016/j.patcog.2005.01.012. – ISSN 0031–3203

[MA01] Montague, M. ; Aslam, J. A.: Relevance score normalization for metasearch. In: *Proceedings of the tenth international conference on Information and knowledge management*. New York, NY, USA : ACM, 2001 (CIKM '01). – ISBN 1–58113–436–3, 427–433

[MRF01] Manmatha, R. ; Rath, T. ; Feng, F.: Modeling score distributions for combining the outputs of search engines. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA : ACM, 2001 (SIGIR '01). – ISBN 1–58113–331–6, 267–275

[MS02] Manmatha, R. ; Sever, H.: A formal approach to score normalization for meta-search. In: *Proceedings of the second international conference on Human Language Technology Research*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2002 (HLT '02), 98–103

[OC03] Ogilvie, P. ; Callan, J.: Combining document representations for known-item search. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA : ACM, 2003 (SIGIR '03). – ISBN 1–58113–646–3, 143–150

[Sch08] Schmitt, I.: QQL: A DB&IR Query Language. In: *The VLDB Journal* 17 (2008), Januar, Nr. 1, 39–56. http://dx.doi.org/10.1007/s00778-007-0070-1. – DOI 10.1007/s00778–007–0070–1. – ISSN 1066–8888

[SSK06] Shalabi, L.A. ; Shaaban, Z. ; Kasasbeh, B.: Data mining: A preprocessing engine. In: *Journal of Computer Science* 2 (2006), Nr. 9, S. 735–739

[WCB06] Wu, S. ; Crestani, F. ; Bi, Y.: Evaluating Score Normalization Methods in Data Fusion. Version: 2006. http://dx.doi.org/10.1007/11880592_57. In: Ng, Hwee (Hrsg.) ; Leong, Mun-Kew (Hrsg.) ; Kan, Min-Yen (Hrsg.) ; Ji, Donghong (Hrsg.): *Information Retrieval Technology* Bd. 4182. Springer Berlin / Heidelberg, 2006. – ISBN 978–3–540–45780–0, 642-648

[wika] *Effizienz (Statistik)*. Wikipedia. https://de.wikipedia.org/w/index.php?title=Effizienz_(Statistik)&oldid=105392532. – Version vom 10. Jul. 2012, 10:31

[wikb] *Korrelationskoeffizient*. Wikipedia. https://de.wikipedia.org/w/index.php?title=Korrelationskoeffizient&oldid=107965499. – Version vom 12. Sep. 2012, 10:16

[wikc] *Sigmoidfunktion*. Wikipedia. https://de.wikipedia.org/w/index.php?title=Sigmoidfunktion&oldid=108698435. – Version vom 30. Sep. 2012, 12:15

[wikd] *Tschebyscheff-Ungleichung*. Wikipedia. https://de.wikipedia.org/w/index.php?title=Tschebyscheff-Ungleichung&oldid=105528371. – Version vom 13. Jul. 2012, 15:00

Selbstständigkeitserklärung

Eidesstattliche Erklärung

Der Verfasser erklärt an Eides statt, dass er die vorliegende Arbeit selbständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt hat. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Ort, Datum Unterschrift des Verfassers