BAUHAUS-UNIVERSITÄT WEIMAR

Fakultät Medien

Studiengang Mediensysteme

Web-Technologie und Content Management Systeme

Verfahren zur Modellbildung für das Dokumenten-Clustering

Diplomarbeit

Tim Gollub

Betreuer: Dr. SvenMeyer zu Eissen

1.Gutachter:Prof.Dr. Benno Maria Stein

Datum der Abgabe: 30.04.2008

Erklärung

Hiermit versichere ich, dass ich diese Diplomarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Stellen meiner Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich unter Angabe der Quelle als Entlehnung kenntlich gemacht. Dasselbe gilt sinngemäß für Tabellen und Abbildungen. Diese Arbeit hat in dieser oder einer ähnlichen Form noch nicht im Rahmen einer anderen Prüfung vorgelegen.

*(*Ort, Datum*)* *(Unterschrift)*

Danksagung

Nun da die Arbeit ihr Ende gefunden hat soll ihr Anfang um einige dankende Worte erweitert werden.

Ich bedanke mich bei Sven Meyer zu Eißen für eine kompetente Betreuung in einem überaus lehrreichen halben Jahr.

Danke an Prof. Benno Stein, Martin Potthast, Maik Anderka und Nedim Lipka. Ihr hattet stehts ein offenes Ohr für meine Fragen - Eure Antworten haben mich vorangebracht; vielen Dank.

Außerdem möchte ich mich bei Michael Bies für die professionelle Korrektur in windeseile bedanken, bei Robert Pötig für die Gesellschaft in der Nacht.

Schließlich ein großes Dankeschön an meine Familie, die mir viel Kraft gegeben und stets den Rücken freigehalten hat. Katrin, du bist die Beste!

Inhaltsverzeichnis

Vorwort **12**

1 Einleitung **13**

1.1 Information Retrieval 13

1.2 Dokumenten-Clustering als Erweiterung des IR-Systems 14

1.3 Experimentelle Evaluierung 16

2 Retrieval-Modelle **17**

2.1 Indexierung 19

2.2 Vektorraummodell 21

2.2.1 Gewichtung der Dokumentterme 22

2.2.2 Berechnung der Retrieval-Werte 28

2.3 2-Poisson-Modell 28

2.4 Divergence From Randomness 29

2.4.1 Informationsgehalt 29

2.4.2 Risiko 32

2.4.3 Normalisierung der Termhäufigkeiten 33

2.4.4 Berechnung der Retrievalwerte 34

2.5 Best Match-Modell 34

2.6 Statistisches Sprachmodell 37

2.7 Explizit Semantisches Indexierungsmodell 39

3 Dokumenten-Clustering **42**

3.1 Cluster-Hypothese 42

3.2 Definition und Ähnlichkeitsmodell 43

3.3 Fusionierungsalgorithmen 45

3.3.1 Hierarchisch agglomerative Fusionierung 46

3.3.2 Exemplarbasierte, iterative Fusionierung 47

3.3.3 Dichtebasierte Fusionierung mit MajorClust 49

3.4 Emittlung der besten Clusterings 49
3.4.1 Elbow-Kriterium 50
3.4.2 Gap-Statistik 51
3.5 Modellvereinfachung 51
3.5.1 k-Nearest-Neighbour-Graph 56
3.5.2 Expected Similarity 56
3.5.3 Local Average 60
3.5.4 Relevance Ratio 60
3.5.5 Major Expected Density 61

4 Qualitätsindizes 63
4.1 Externe Qualitätsindizes 63
4.1.1 F-Measure 63
4.2 Interne Qualitätsindizes 65
4.2.1 Dunn-Index 66
4.2.2 Davis-Boldin-Index 66
4.2.3 Expected Density 67

5 Experimente 68
5.1 Allgemeines Korrelationsexperiment 70
5.2 Spezielles Korrelationsexperiment 72
5.3 Vergleich der Retrieval-Modelle 74
5.4 Vergleich der Ansätze zur Modellvereinfachung 80
5.4.1 Vereinfachung mit Local Average 81
5.4.2 Vereinfachung mit dem Relevance-Ratio-Verfahren 83
5.4.3 Vereinfachung mit Major Expected Density 84
5.4.4 Vereinfachung mit dem Expected-Similarity-Verfahren . . . 85
5.4.5 Fazit 86

6 Zusammenfassung und Ausblick 88

Literatur 89

Abbildungsverzeichnis

1 Herausforderung im IR. Aus der *linguistischen Theorie* muss sich ein Prinzip formalisieren lassen, das die Relevanzbestimmung zwischen Anfrage und Dokument auf semantischer Ebene leisten kann. 14

2 Dokumenten-Clustering. Die eingehende Dokumentmenge wird in Gruppen thematisch verwandter Dokumente aufgeteilt. 15

3 Konzeptueller Aufbau eines Retrieval-Modells (vgl. [Stein08]). Um die Dokumente $d \in D$ der Relevanzberechnung $\rho_{\mathcal{R}}$ zugänglich zu machen, müssen sie auf eine adäquate Datenstruktur $\mathbf{d} \in \mathbf{D}$ abgebildet werden ($\alpha_{\mathcal{R}}$). Die äquivalente Abbildung $\mathbf{q}$ des Informationsbedarfs q geschieht, in existierenden IR-Systemen, mit Formulierung der Anfrage durch den Nutzer. 17

4 Repräsentation der Dokumente d_1 und d_2 als Vektor über die Menge der Terme T. 18

5 Taxonomie von Retrieval-Modellen (in Anlehnung an [Stein04e]). Die Modelle lassen sich hinsichtlich der Merkmale, die von der unterliegenden linguistischen Theorie verwendet werden, einteilen. Blau hervorgehobene Modelle sind Gegenstand dieses Kapitels. . . 19

6 Techniken zur Indexierung einer Dokumentkollektion (vgl. [Stock07]) 20

7 Vektorraum. 21

8 Normierung des Dokumentvektors auf die Einheitslänge. 24

9 Pivoted Unique Normalization (Quelle: [Singhal96]). Für Wortschatzgrößen, bei denen die Wahrscheinlichkeit für ein Dokument gefunden zu werden ($P(Retrieval)$) größer ist als die Wahrscheinlichkeit relevant zu sein ($P(Relevance)$), wird der Normalisierungsfaktor erhöht. Gilt das Gegenteil, wird der Faktor gesenkt. 25

10 Relevanz als Kosinus des Winkels φ zwischen Anfrage und Dokument. 28

11 ESA-Szenario. Über die Gewichte der Artikelrepräsentationen $\mathbf{c} \in \mathbf{C}$ und der Dokumentrepräsentationen $\mathbf{d} \in \mathbf{D}$ wird jedes Dokument auf einen gewichteten Vektor von Wikipedia-Artikeln $\mathbf{k} \in \mathbf{K}$ abgebildet. (Quelle: [Gabrilovich07]) 40

12 Ähnlichkeit zwischen Dokumentpaaren. Die Retrieval-Modelle aus Kapitel 2 können für die Berechnung der Ähnlichkeiten zwischen den Dokumenten der Kollektion verwendet werden. 42

13 Auszug aus einem Ähnlichkeitsgraphen G(V,E,w). Die Dokumentmenge D wird bijektiv auf die Menge der Knoten V abgebildet, die Ähnlichkeit $\varphi(\mathbf{d}_i, \mathbf{d}_j)$ entspricht dem Gewicht $w(\{u, v\})$ der Kante $\{u, v\} \in E$ zwischen den mit $\mathbf{d}_i$ und $\mathbf{d}_j$ assoziierten Knoten u und v [Stein04a]. 45

14 Taxonomie der Fusionierungsalgorithmen (Quelle: [Stein04a]) . . . 45

15 Dendrogramm einer beendeten, agglomerativen Fusionierung (Quelle: [Stein04b]). 47

16 Iterationsschritt des k-Means-Algorithmus. Nachdem die Objekte zugeordnet sind, werden die Centroiden (blau) dem Zielkriterium gemäß verschoben. (Quelle: [Stein04c]) 48

17 Elbow-Kriterium. Beim Auftragen der Heterogenität gegen die Anzahl der Cluster bildet sich an der Stelle der optimalen Clusteranzahl ein „Ellenbogen“ (Quelle: [MeyerZuEissen07]). 50

18 Masse-Verteilung im Ähnlichkeitsgraphen. Die kleinsten Ähnlichkeitswerte steuern aufgrund ihrer Vielzahl den größten Teil zur Gesamtmasse bei (rote Balken). Die meisten Werte stammen dabei von Dokumentpaaren $(\mathbf{d_i}, \mathbf{d_j})$ aus unterschiedlichen Clustern $\mathbf{d}_i \in \mathbf{C}_i$,$\mathbf{d}_j \in \mathbf{C}_{j \neq i}$ (grüne Balken). Mit Zunahme des Ähnlichkeitswertes steigt der Anteil informationstragender Ähnlichkeiten (blaue Balken). 52

19 Falsche Mehrheitsentscheidung, verursacht durch Rauschen. Die vielen kleinen Ähnlichkeiten aus der Dokumentgruppe links erzeugen in Summe eine größere Anziehungskraft als die beiden schweren Kanten rechts. Die Fusionierung endet mit einem einzigen, gemeinsamen Cluster. 53

20 Verteilung der Ähnlichkeitsmasse unter Verwendung des Tf-Idf-Modells. Im Vergleich zur reinen Tf-Gewichtung setzen sich Ähnlichkeiten innerhalb der Gruppen (blau) früher von den Ähnlichkeiten zwischen Gruppen (grün) ab. Ihr Anteil an der Gesamtmasse bleibt dabei nahezu unverändert. 54

21 Ähnlichkeitsmasse im Intervall [0, 0.4] des Tf-Idf-Modells. Bei genauererer Betrachtung der Ähnlichkeitsverteilung in den kleinen Wertebereichen ergibt sich die Verteilungscharakteristik des Tf-Modells. 55

22 Erwartetes Termgewicht $w(t, \overline{\mathbf{d}})$. Während für Term t_1 das erwartete Termgewicht $w(t_1, \overline{\mathbf{d}})$ weit unter den Gewichten $w(t_1, \mathbf{d}_{1,\ldots,4})$ der, den Term t_1 enthaltenden, Dokumente 1 bis 4 liegt (rote Elemente), streuen die Termgewichte $w(t_2, \mathbf{d}_{1,\ldots,13})$ von Term t_2 (grün) um das erwartete Gewicht $w(t_2, \overline{\mathbf{d}})$ 58

23 Termgewichte des Referenzdokuments $\mathbf{d}^*$. Das Referenzgewicht $w(t_1, \mathbf{d}^*)$ für Term 1 ist niedriger als die Gewichte in den Dokumenten 1 bis 4 (rot). Im Gegensatz dazu schaffen es nur zwei Dokumente, für Term 2 das Referenzgewicht $w(t_2, \mathbf{d}^*)$ zu überschreiten (grün). . . 59

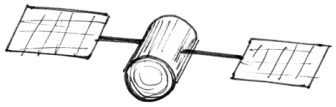
24 Bestimmung des „stärksten“ Partners (*best match*). Die Entscheidung, über welches Clustering iteriert wird, beeinflusst bei ungleicher Zahl der Cluster den Wert des F-Measures. 65

25 Teil der Themenhierarchie in der Reuters-Sammlung RCV1 (Quelle: [MeyerZuEissen07]) 68

26 Korrelation zwischen F-Measure und Davis-Boldin-Index (oben links), Dunn-Index (oben rechts) und Expected Density (unten links).Da ein kleinerer Wert für den Davis-Boldin-Index ein besseres Clustering beschreibt, wurde zur besseren Vergleichbarkeit das Vorzeichen der Indexwerte geändert. 71

27 Streudiagramm für das spezielle Korrelationsexperiment. Auf der linken Seite der Abbildung befinden sich die Kollektionen RC6E600 (oben) und RC5E500 (unten) aus der Reuters-Sammlung, auf der rechten Seite sind die entsprechenden Kollektionen aus der LA Times-Sammlung zu sehen. 73

28 Schematischer Ablauf des Experiments. Dargestellt für eine Testkollektion und ein Retrieval-Modell. 76



Vorwort

Nachdem die Sowjetunion am 4. Oktober 1957 als erste Nation der Welt einen Satelliten ins All geschossen hatte, saß der Schock in den USA tief. Man war im technologischen Wettlauf um die Krone der Weltraumforschung ins Hintertreffen geraten. Und auch informationswissenschaftlich war das Ereignis von historischem Belang. In den USA beschäftigte man sich ein halbes Jahr mit der Dekodierung der Signale, die vom Satelliten zur Erde gesendet wurden - ein halbes Jahr, das man sich hätte sparen können, da die Bedeutung und Kodierung des Signals bereits vor Start der Rakete in einem sowjetischen Fachblatt veröffentlicht, und sogar von amerikanischen Stellen übersetzt worden waren[Rauch88]. Die Lösung lag also längst in den Bibliotheken des Landes. In aller Deutlichkeit offenbarte sich, dass das Vorhandensein von Information in Anbetracht eines explodierenden Datenvolumens allein nicht mehr ausreichte, um einen Anstieg an Informiertheit zu garantieren. Vielmehr benötigte man zudem Informationssysteme, die den *Transfer* der Information vom Wissensproduzenten zum Nachfragendem erfassen und unterstützen: Systeme, die sich der *Semantik* der Dokumente annehmen; Systeme, die den *Informationsbedarf* des Fragenden analysieren; kurz Systeme, die Information Retrieval (IR) betreiben.

1 Einleitung

Diese Arbeit fällt in das Themengebiet des Information Retrievals (IR). Sie befasst sich im Schwerpunkt mit dem Einsatz von Retrieval-Modellen in der textbasierten Cluster-Analyse.

1.1 Information Retrieval

Information Retrieval beschäftigt sich mit dem Prozess des Wissenstransfers vom Wissensproduzenten zum Informationsnachfragenden [Fuhr96]. Das Ziel ist es, einem Menschen in seiner individuellen Situation die Informationen zukommen zu lassen, die er für die Bewältigung seines Anliegens benötigt.

Der Informationsnachfragende (i.w. Nutzer) hat die Möglichkeit, seinen Informationsbedarf in Form einer Anfrage (engl. *query*) an das System zu richten. Im Gegensatz zu formalen Datenbanksystemen kann die Anfrage bei einem IR-System vage und in natürlicher Sprache formuliert werden.

Zur Beantwortung der Anfrage hat ein IR-System Zugriff auf eine oder mehrere Datenquellen, die zur Informationsbeschaffung genutzt werden. In der Regel liegen die Daten dort in Form von Textdokumenten vor, es sind (in dieser Arbeit nicht betrachtet) aber auch alle Bild- und Ton- Formate denkbar. Unter der Systemantwort stellt man sich idealerweise eine Synthese aller relevanten Informationen zu einer geschlossenen Präsentation vor. Üblich ist derzeit eine nach Relevanz sortierte Liste der informationstragenden Dokumente (sog. *relevance ranking*).

Die Aufgabe eine IR-Systems ist eine große Herausforderung. Auf Grundlage der vagen Anfrage müssen der Informationsbedarf des Nutzers abgeleitet und die Dokumente im Speicher auf Nützlichkeit zur Beantwortung der Anfrage geprüft werden. Die Bearbeitung verlangt nach einem formalen Prinzip, mit dem auf semantischer Ebene die Relevanz eines Dokuments für eine Anfrage festgestellt werden kann (Abbildung 1).

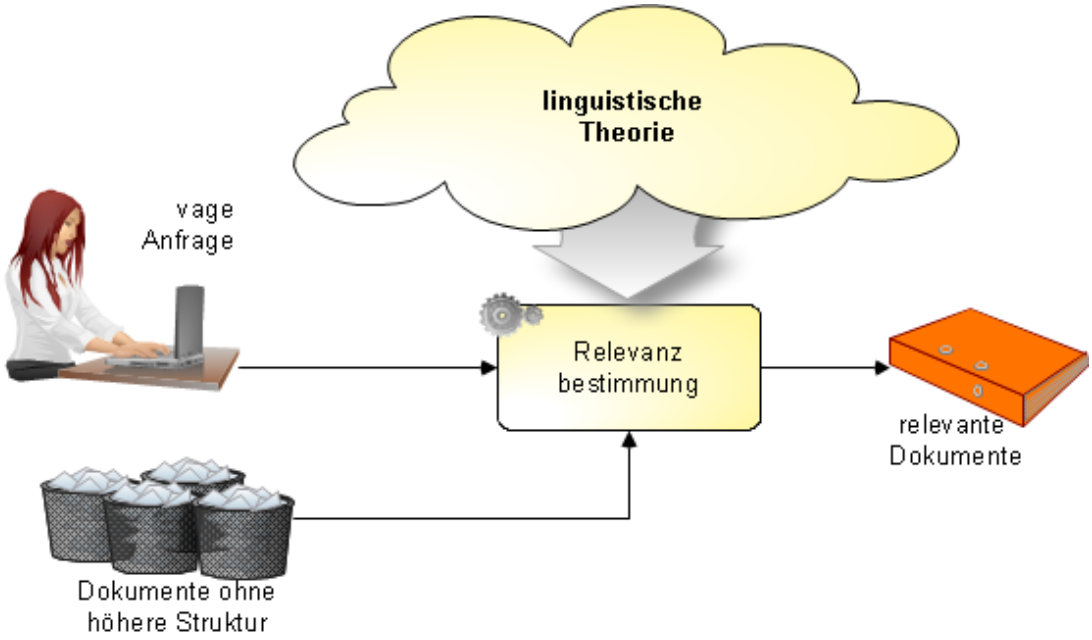


Abbildung 1: Herausforderung im IR. Aus der *linguistischen Theorie* muss sich ein Prinzip formalisieren lassen, das die Relevanzbestimmung zwischen Anfrage und Dokument auf semantischer Ebene leisten kann.

Die Größe dieser Herausforderung lässt sich veranschaulichen, wenn man sich menschliche Bearbeiter anstelle des Computers vorstellt. Trotz ihres aus Sicht des Computers enormen Verständnisses für die Welt und ihre Probleme, werden die Vorstellungen von einer korrekten Antwort im Einzelfall differieren. Es wird verschiedene Meinungen darüber geben, wie die Anfrage des Nutzers zu deuten ist und ob eine bestimmte Information von Belang ist oder nicht. In Anbetracht dessen ist es nur schwer vorstelbar, dass ein Computer auf Grundlage formaler Berechnungen dem vielfältig verschleierten, oft assoziativen Prozess der Interpretation menschlicher Kommunikation in einer nur annähernd adäquaten Weise entgegentreten kann.

Und trotzdem gibt es sie, die Retrieval-Systeme, und wir Internetnutzer befragen sie täglich. Im zweiten Kapitel dieser Arbeit werden die klassischen Modelle vorgestellt, auf deren Grundlage der Retrieval-Prozess stattfindet. Darauf aufbauend wird dann das Dokumenten-Clustering mit in den Retrieval-Prozess eingebunden.

1.2 Dokumenten-Clustering als Erweiterung des IR-Systems

Im Bestreben, die Qualität der Retrieval-Ergebnisse zu verbessern, wurde vorgeschlagen, die Cluster-Analyse in den Retrieval-Prozess zu integrieren [vanRijsbergen79].

Ausgangspunkt dafür ist die Menge der als relevant eingestuften Dokumente (Abbildung 2).

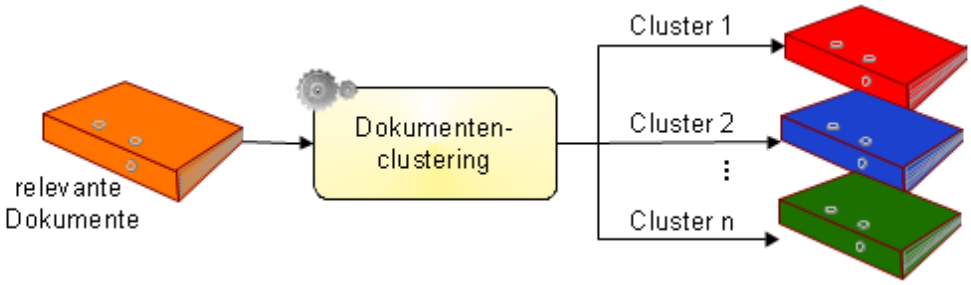


Abbildung 2: Dokumenten-Clustering. Die eingehende Dokumentmenge wird in Gruppen thematisch verwandter Dokumente aufgeteilt.

Die Cluster-Analyse teilt eine heterogene Menge von Objekten (hier die Dokumente) in homogene Teilmengen (sog. *Cluster*). Man kalkuliert, dass mit der Cluster-Analyse fälschlicherweise als relevant eingestufte Dokumente von der Menge der tatsächlich relevanten Dokumente (bzgl. einer Anfrage) getrennt werden können. Damit kann gerechnet werden, wenn sich die Merkmale, auf deren Grundlage die Cluster bestimmt werden, auf die Semantik der Dokumente beziehen. Nur eine thematische Aufteilung der Dokumente gibt Grund zur Annahme, dass Dokumente aus verschiedenen Clustern Antworten auf verschiedene Informationsbedürfnisse liefern können. Somit besteht auch im Dokumenten-Clustering eine Aufgabe darin, die Dokumente in ihrer Bedeutung zu erfassen und zu repräsentieren. In dem der textbasierten Cluster-Analyse gewidmeten dritten Kapitel wird gezeigt, wie Retrieval-Modelle im Dokumenten-Clustering eingesetzt werden, und wie Fusionierungsalgorithmen bei der Zerlegung der Objektmenge vorgehen. Außerdem werden Ansätze zur Modellvereinfachung vorgestellt, die im Rahmen dieser Arbeit entwickelt wurden und die Qualität der Fusionierung verbessern können.

Im vierten Kapitel werden Indizes thematisiert, die die Qualität eines Clusterings messen. Zwei Arten werden hierbei unterschieden. Interne Qualitätsindizes bewerten ein Clustering auf Grundlage seiner strukturellen Eigenschaften. Sie können als integraler Bestandteil im Dokumenten-Clustering eingesetzt werden, um aus einer Menge erzeugter Clusterings das wohlgeformteste zu nominieren. Externe Qualitätsindizes hingegen vergleichen ein Clustering mit einer von außen kommenden Referenzlösung. Sie werden in Experimenten eingesetzt, um eine objektive Aussage über die Cluster-Qualität zu erhalten.

1.3 Experimentelle Evaluierung

Die Erstellung der Arbeit wird von einer Reihe an Experimenten begleitet. Drei Fragestellungen sollen untersucht werden:

1. Wie zuverlässig sind interne Qualitätsindizes bei der Einschätzung der Cluster-Qualität?
2. Wie gut lassen sich Retrieval-Modelle zur Ähnlichkeitsbestimmung im Dokumenten-Clustering einsetzen?
3. Können die entwickelten Verfahren zur Modellvereinfachung die Qualität der Fusionierung signifikant steigern?

Die Ergebnisse der Experimente werden in Kapitel 5 vorgestellt.

2 Retrieval-Modelle

Sei $D = \{d_1, d_2, ..., d_n\}$ die Menge der Dokumente in einem IR-System, $Q = \{q_1, q_2, ..., q_m\}$ die Menge aller Informationsbedürfnisse. Sei weiter die Formulierung einer Anfrage durch den Nutzer, die Abbildung seines Informationsbedarfs $q_i \in Q$ auf eine formalisierte Anfrage $\mathbf{q}_i \in \mathbf{Q}$.

Um dem Informationsbedarf $q_i \in Q$ eines Nutzers nachzukommen ist es notwendig, auf Grundlage von Anfrage $\mathbf{q}_i$ die Dokumentmenge $D^* \subseteq D$ zu bestimmen, die relevante Informationen hinsichtlich q enthält. Eine linguistische Theorie die die Ermittlung von D^* ermöglicht heißt formalisiert *Retrieval-Modell* oder *Retrieval-Strategie* $\mathcal{R}$. Ein Retrieval-Modell $\mathcal{R}$ definiert eine Vorschrift, um die Relevanz jedes Dokuments $d \in D$ bezüglich eines Informationsbedarfs q zu ermitteln. Abbildung 3 zeigt den konzeptuellen Aufbau eines Retrieval-Modells.

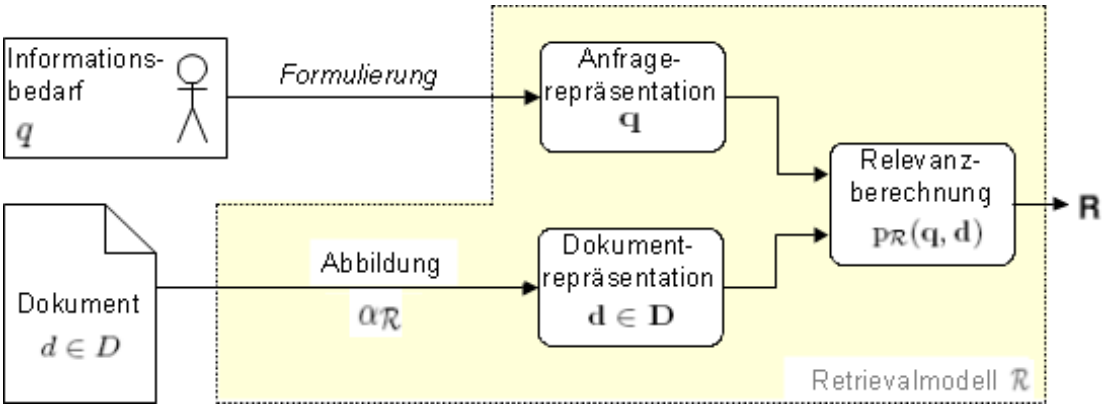


Abbildung 3: Konzeptueller Aufbau eines Retrieval-Modells (vgl. [Stein08]). Um die Dokumente $d \in D$ der Relevanzberechnung $\rho_{\mathcal{R}}$ zugänglich zu machen, müssen sie auf eine adäquate Datenstruktur $\mathbf{d} \in \mathbf{D}$ abgebildet werden ($\alpha_{\mathcal{R}}$). Die äquivalente Abbildung $\mathbf{q}$ des Informationsbedarfs q geschieht, in existierenden IR-Systemen, mit Formulierung der Anfrage durch den Nutzer.

Zentrales Element im Retrieval-Modell ist $\alpha_{\mathcal{R}}$, die Abbildung der Dokumente $d \in D$ auf abstrakte Dokumentrepräsentationen $\mathbf{d} \in \mathbf{D}$. Mit $\alpha_{\mathcal{R}}$ wird ein Dokument auf quantifizierbare Merkmale reduziert, die auf semantischer Ebene mit dem Dokument verknüpft und auf die Verrechnung mit Anfragen $\mathbf{q} \in \mathbf{Q}$ zugeschnitten sind (vgl. [Stein08]). Abbildung $\alpha_{\mathcal{R}}$ beinhaltet die *Indexierung* der Dokumentmenge D sowie die *Gewichtung* der Dokumentrepräsentationen $\mathbf{d} \in \mathbf{D}$.

Die in diesem Kapitel vorgestellten Retrieval-Modelle abstrahieren mit der Indexierung die Dokumentmenge D auf eine unstrukturierte Menge von Termen $T = \{t_1, t_2, \ldots, t_k\}$ [Stein04f]. Auf Techniken zur Konstruktion von T wird in

Abschnitt 2.1 näher eingegangen. Die Termmenge T wird entweder unmittelbar von den Dokumentrepräsentationen $\mathbf{d} \in \mathbf{D}$ referenziert (Abbildung 4), oder als Grundlage für eine weitere Abstraktion der Dokumente auf *Konzepte* $C = \{c_1, c_2, \ldots, c_l\}$ verwendet. Falls nicht anders hervorgehoben, wird im Folgenden die unmittelbare Verwendung der Terme angenommen.

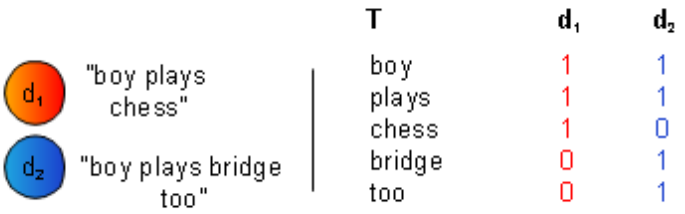


Abbildung 4: Repräsentation der Dokumente d_1 und d_2 als Vektor über die Menge der Terme T.

Während die Indexierung für alle hier vorgestellten Retrieval-Modelle gleichermaßen gültig ist, definiert jedes der Retrieval-Modelle eigene Funktionen $w(t, \mathbf{d})$ bzw. $w(c, \mathbf{d})$ zur Gewichtung der Dokumentrepräsentationen $\mathbf{d} \in \mathbf{D}$. Die Gewichtungsfunktion w quantifiziert die dokumentspezifische Bedeutung eines Merkmals der Dokumentrepräsentationen $\mathbf{d} \in \mathbf{D}$,

$$w : T \times \mathbf{D} \rightarrow \mathrm{R}\,, \text{bzw. } w : C \times \mathbf{D} \rightarrow \mathrm{R}\,.$$

Es sind die Gewichte, in denen sich die Semantik eines Dokuments manifestiert.

Neben $\alpha_{\mathcal{R}}$ ist die *Retrieval-Funktion* $\rho_{\mathcal{R}}(\mathbf{q}, \mathbf{d})$ definiert, die jedem Anfrage-Dokument-Paar $(q, d) \in Q \times D$ einen *Retrieval-Wert* zuordnet. Der Retrieval-Wert kann als Quantifizierung der Relevanz zwischen q und d interpretiert werden [Stein04e],

$$\rho_{\mathcal{R}} \colon \mathbf{Q} \times \mathbf{D} \rightarrow \mathrm{R}\,.$$

Mit dem Retrieval-Wert ist entscheidbar, welche Dokumente $d \in D$ in die Menge relevanter Dokumente D^* aufgenommen werden.

Abbildung 5 zeigt eine Taxonomie existierender Retrieval-Modelle.

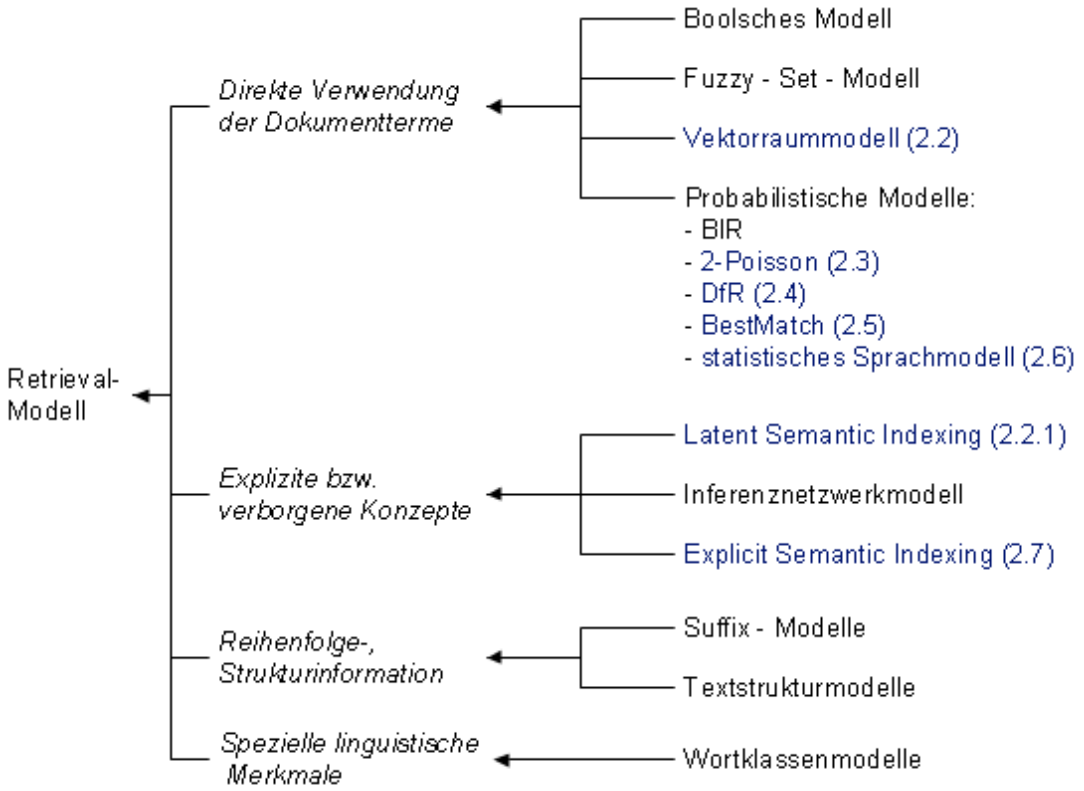


Abbildung 5: Taxonomie von Retrieval-Modellen (in Anlehnung an [Stein04e]). Die Modelle lassen sich hinsichtlich der Merkmale, die von der unterliegenden linguistischen Theorie verwendet werden, einteilen. Blau hervorgehobene Modelle sind Gegenstand dieses Kapitels.

2.1 Indexierung

Mit der Indexierung werden die Dokumente d einer Dokumentkollektion D auf eine abstrakte Datenstruktur $\mathbf{d} \in \mathbf{D}$ abgebildet, die mit Blick auf die weitere Verwendung effektiv ist. Die in dieser Arbeit vorgestellten Retrieval-Modelle abstrahieren die Dokumente auf eine Menge von Termen $T = \{t_1, t_2, \ldots, t_k\}$. Die für die Extraktion der Terme einsetzbaren Verfahren sind in Form einer Verarbeitungskette in Abbildung 6 illustriert.

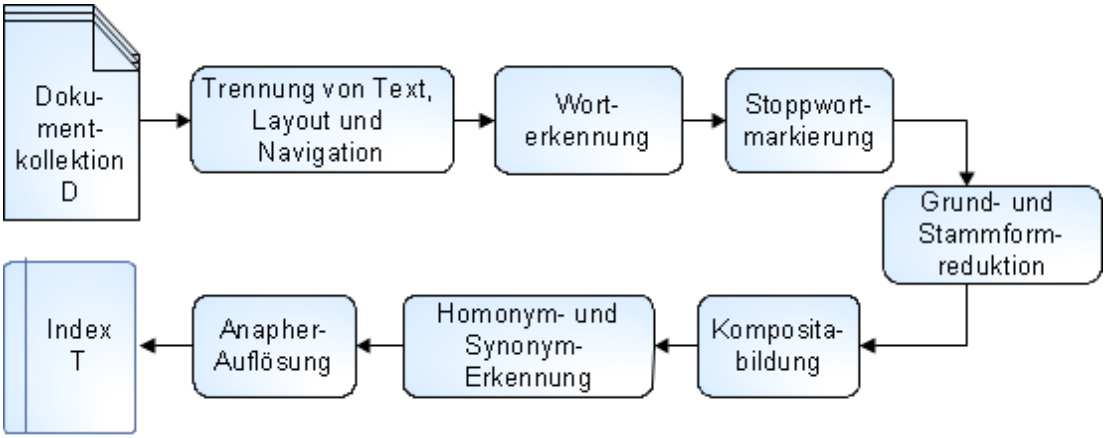


Abbildung 6: Techniken zur Indexierung einer Dokumentkollektion (vgl. [Stock07]) .

Im Ausgangspunkt der Indexierung liegen die Dokumente der Dokumentkollektion in Form von Zeichenketten vor. Diese werden zunächst auf das Vorhandensein von Strukturinformationen wie Layout- oder Navigationsbefehle (z.B. die Tags in Auszeichnungssprachen wie HTML oder XML) hin überprüft. Im Falle solcher werden sie vom Text getrennt. Während Strukturinformationen für die meisten Retrieval-Modelle (unter ihnen die hier beschriebenen) keine Rolle spielen, werden sie im sogenannten „Structured Information Retrieval“ als wichtige Hinweisgeber wahrgenommen. So leitet man beispielsweise aus Auszeichnungen wie „Titel“ oder „Überschrift“ eine größere Bedeutung der markierten Zeichenfolgen für die inhaltliche Beschreibung eines Dokuments ab.

Bei der Worterkennung werden die Zeichenketten anhand von Leer- und Interpunktionszeichen tokenisiert. Daran anschließend werden Wortformen wie Artikel, Konjunktionen oder Partikeln, die häufig und gleichverteilt in den Texten vorkommen, anhand von Stopwortlisten entfernt. Man argumentiert, dass diese Wörter nie etwas zur inhaltlichen Beschreibung eines Dokuments beitragen, aber rund die Hälfte eines Textes ausmachen.

Im weiteren Verarbeitungprozess können computerlinguistische Verfahren zur Behandlung sprachlicher Phänomene eingesetzt werden. Vor allem graphematische Ansätze (Grund- und Stammformreduktion) erreichen in weniger flektierten Sprachen wie dem Englischen eine hohe Genauigkeit und werden im Rahmen des Information Retrievals oftmals verwendet [Fuhr06]. Bei der Grundformreduktion werden Wörter in ihren Flexionsformen auf einen gemeinsamen Bezeichner zurückgeführt. Ohne Grundformreduktion würden beispielweise die Wörter „spielen“, „spielt“, „gespielt“ und „spielend“ jeweils einen eigenständigen Term bilden.

Zur Auflösung von Derivationsformen (unterschiedliche Wortformen eines Wortstamms) wird die Stammformreduktion eingesetzt. Komposita, Homonyme und Synonyme können mit lexikalischen Verfahren, basierend auf einem Wörterbuch, identifiziert werden. Zur Auflösung von Anaphern (z.B. Pronomen) werden meist Heuristiken eingesetzt [Feldmann03].

Die aus dem Verarbeitungsprozess hervorgegangenen Terme bilden die Elemente der Termmenge T. Die im Folgenden vorgestellten Retrieval-Modelle verwenden, unabhängig von der konkret gewählten Verarbeitungstiefe, die Termmenge T zur Repräsentation der Dokumente.

2.2 Vektorraummodell

Das Vektorraummodell ist einer der grundlegenden theoretischen Ansätze im Information Retrieval und wurde in den 60er Jahren mit dem SMART-System in der Arbeitsgruppe um Gerard Salton entwickelt [Stock07]. Es fasst die Terme $t \in T$ als Dimensionen eines n-dimensionalen Raumes auf. Dokumentrepräsentationen $\mathbf{d}$ sowie Anfragen $\mathbf{q}$ sind als Vektoren in diesem Raum definiert (Abbildung 7).

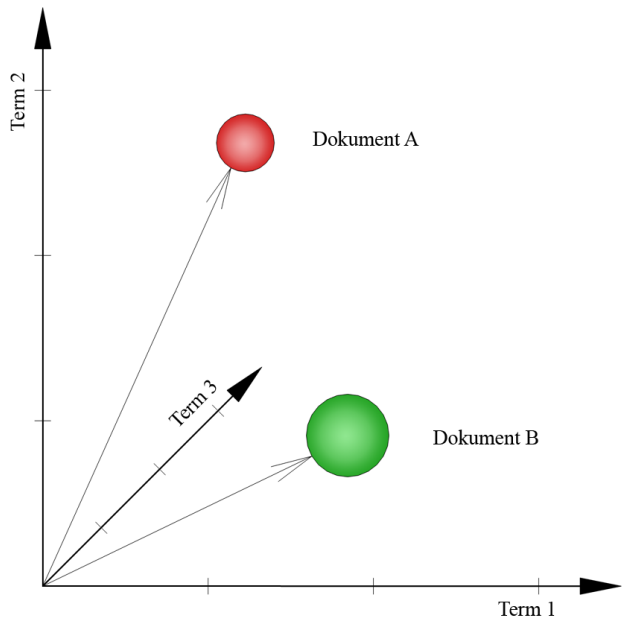


Abbildung 7: Vektorraum.

2.2.1 Gewichtung der Dokumentterme

Zur Gewichtung $w(t, \mathbf{d})$ der Terme einer Dokumentrepräsentation existieren verschiedener Ansätze. Ziel ist es, die Bedeutung eines Terms t zur inhaltlichen Beschreibung eines Dokuments $d \in D$ mit einem Wert zu quantifizieren.

Absolute Termhäufigkeit

Die Behauptung, die Häufigkeit eines Terms in einem Text sei ein Maß für seine Relevanz, ist bekannt als die „These von Luhn“. Sie besagt, dass sehr häufig, sowie äußerst selten vorkommende Terme gar keine oder wenig Bedeutung für ein Dokument haben. Dazwischen erstreckt sich die Bedeutung eines Wortes für einen Text in Form einer Gauß-Funktion [Stock07]. Werden die sehr häufigen, inhaltslosen Worte durch eine Stoppwortliste eliminiert, gibt die absolute Häufigkeit $tf(t, d)$ die Bedeutung eines Terms t für ein Dokument d :

$$w_{tf}(\mathrm{t}, \mathbf{d}) = tf(\mathrm{t}, d) \tag{1}$$

Donna Harman stellte fest, dass die Bedeutung eines Terms in einem Dokument nicht linear mit der Häufigkeit steigt. Um den Wertebereich der Termgewichte zu stauchen, logarithmierte sie die Berechnungsvorschrift [Harman86],

$$w_{tf_{\log}}(t, \mathbf{d}) = \log_2 \ tf(t, d) + 1$$

Relative Termhäufigkeit

Mit Blick auf die gesamte Dokumentkollektion D, gerät die Gewichtung mit der absoluten Häufigkeit unter Druck. Sie bevorzugt lange Dokumente. Allein auf Grund des großen Termvolumens erreichen diese höhere Häufigkeiten und damit höhere Termgewichte als kurze Dokumente.

Gerard Salton schlägt als Lösung zunächst eine Normierung über die Dokumentlänge $l(d)$ (= Summe aller Termhäufigkeiten tf in d) vor [Salton68]:

$$w_{tf_l}(t, \mathbf{d}) = \frac{tf(t, d)}{l(d)}$$

Weitere Ansätze benutzen die maximale Termfrequenz max_tf,

$$w_{tf_{max}}(t, \mathbf{d}) = \frac{\log\ tf(t,d) + 1}{\log\ \max_\mathrm{tf}(d)}\ ,$$

bzw. die durchschnittliche Termfrequenz avg_tf,

$$w_{tf_{avg}}(t, \mathbf{d}) = \frac{\log\ tf(t,d) + 1}{\log\ \mathrm{avg_tf}(d)}\ , \tag{2}$$

zur Normierung von Dokumentrepräsentation **d** [Singhal96].

Neben den höheren absoluten Termfrequenzen verfügen lange Dokumente naturgemäß auch über einen größeren Wortschatz. Dadurch steigt die Zahl potentieller Übereinstimmungen mit den Termen einer Anfrage, und damit unweigerlich die Chance, gefunden zu werden. Eine Normierung, die auch diese zweite Bevorzugung relativiert, ist

$$w_{\mathrm{tf}_{\cos}}(t, \mathbf{d}) = \frac{\log\ tf(t,d) + 1}{\sqrt{\sum_{i \in \mathbf{d}} (\log\ tf(t_i,d) + 1)^2}}\ . \tag{3}$$

Der Nenner steigt sowohl mit Schwere als auch mit Anzahl der Termgewichte. Im Vokabular des SMART-Systems trägt Formel 3 zur Gewichtung der Dokumentrepräsentationen **d** den Namen „lnc“.

Geometrisch betrachtet ist Formel 3 die Normierung des Dokumentvektors auf die Einheitslänge (Abbildung 8). Zwei Dokumente sind identisch, wenn die Terme im gleichen Verhältnis verwendet werden. Die räumliche Distanz wird unerheblich.

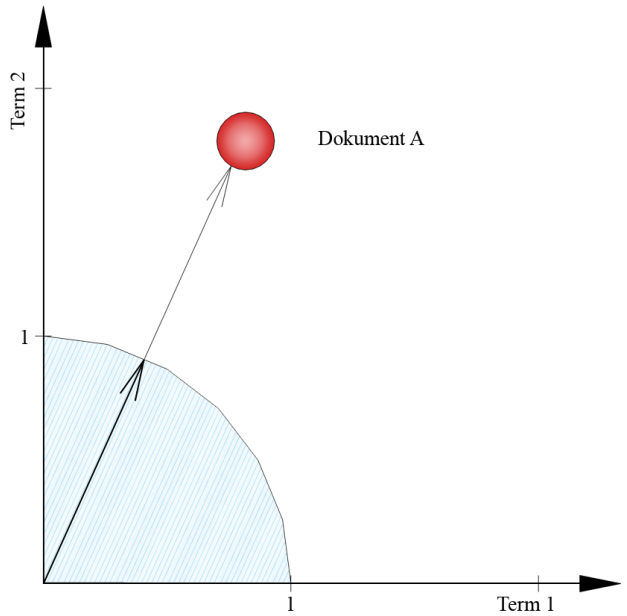


Abbildung 8: Normierung des Dokumentvektors auf die Einheitslänge.

Ein anderer Ansatz wird im SMART-System durch das Kürzel „Lnu“ beschrieben, bei dem die Korrektur der beiden Längenprobleme „höhere tf-Werte“ und „größerer Wortschatz“ separat stattfindet. Für das erste Problem wird die Normalisierung mittels durchschnittlicher Termfrequenz (2) eingesetzt, der größere Wortschatz wird mit der „Pivoted Unique Normalization“ (kurz *u-Normalisierung*) von Singhal und Buckley [Singhal96] kompensiert:

$$w_{Lnu}(t, \mathbf{d}) = \frac{\frac{\log\ tf(t,d)+1}{\log\ \mathrm{avg_tf}(d)}}{(1.0 - slope) \cdot pivot + slope \cdot |W_d|} \tag{4}$$

$|W_d = \{t \,|\, tf(t,d) > 0,\ d \in D\}|$ ist die Anzahl unterschiedlicher Terme t in einem Dokument und wird als Wortschatz von d bezeichnet. Mit den Variablen *pivot* („Drehachse“) und *slope* („Steigung“) wird der Verlauf des Normalisierungsfaktors über die Wortschatzgröße gesteuert. Es wird angestrebt, die Normalisierungsgerade so zu legen, dass die Wahrscheinlichkeit eines Dokumentes $d_i \in D$ mit Wortschatzgröße $|W_{d_i}|$ gefunden (als relevant eingestuft) zu werden, nahe bei der Wahrscheinlichkeit liegt, dass ein relevantes Dokument die Wortschatzgröße $|W_{d_i}|$ besitzt (Abbildung 9).

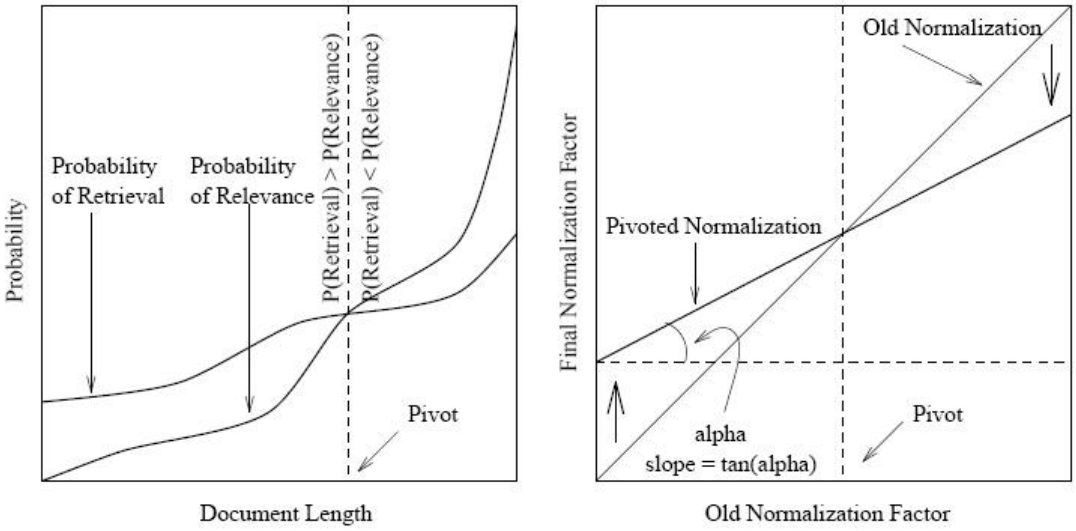


Abbildung 9: Pivoted Unique Normalization (Quelle: [Singhal96]). Für Wortschatzgrößen, bei denen die Wahrscheinlichkeit für ein Dokument gefunden zu werden ($P(Retrieval)$) größer ist als die Wahrscheinlichkeit relevant zu sein ($P(Relevance)$), wird der Normalisierungsfaktor erhöht. Gilt das Gegenteil, wird der Faktor gesenkt.

Eine Angabe der zweiten Wahrscheinlichkeit ist im Anwendungsfall nicht möglich, weshalb empirisch ermittelte Werte für die Parameter benutzt werden müssen. In [Buckley98] zur vierten Text Retrieval Konferenz (TREC-4) wird der pivot-Wert als „durchschnittliche Wortschatzgröße in der Kollektion D“ fixiert, der slope-Wert auf 0,2 gesetzt.

Inverse Dokumenthäufigkeit

Der Ansatz der inversen Dokumenthäufigkeit, der von Stephen Robertson und Karen Sparck Jones geprägt wurde, bestimmt das Gewicht eines Terms nach seiner Diskriminanzkraft fest [Stock07]. Je weniger Dokumente einen Term enthalten, desto trennschärfer ist er, und desto höher ist sein Gewicht zu bemessen:

$$w_{Idf}(t) = ld\left(\frac{\mathrm{N}}{\mathrm{n}_t}\right) \ ,$$

mit N = Gesamtzahl der Dokumente, n_t = Zahl der Dokumente die Term t enthalten.

Eine alternative Formulierung der inversen Dokumenthäufigkeit ergibt sich aus der Robertson/Sparck-Jones-Formel [Robertson76], die das Termgewicht an die Wahrscheinlichkeiten der bedingten Ereignisse $\mathrm{Rel_d}$ =„Dokument d ist relevant“

und $\mathrm{Ent_{d,t}}$ =„Dokument d enthält Term t“ knüpft. Eine große Wahrscheinlichkeit für das Eintreffen oder Ausbleiben *beider* Ereignisse führt zu einem hohen Gewicht des Terms t für Dokument d. Ist die Wahrscheinlichkeit groß, dass nur eines der Ereignisse auftritt, bekommt Term t ein kleines Gewicht,

$$w_{RSJ}(\mathrm{t}, \mathbf{d}) = \log \frac{P(\mathrm{Ent_{d,t}}|\mathrm{Rel_{\mathbf{d}}})\ \left(1 - P(\mathrm{Ent_{d,t}}|\overline{\mathrm{Rel_{\mathbf{d}}}})\right)}{P(\mathrm{Ent_{d,t}}|\overline{\mathrm{Rel_{\mathbf{d}}}})\ (1 - P(\mathrm{Ent_{d,t}}|\mathrm{Rel_{\mathbf{d}}}))}$$

Robertson und Sparck-Jones approximieren die unbekannten Wahrscheinlichkeiten der Formel,

$$w_{RSJ}(\mathrm{t}, \mathbf{d}) \approx \log \frac{(r + 0.5)/(R - r + 0.5)}{(\mathrm{N} - r + 0.5)/(\mathrm{N} - \mathrm{n}_t - R + r + 0.5)}\ .$$

Die Parameter r und R der Approximation stehen für gesicherte Informationen, wobei R als „Anzahl der als relevant bekannten Dokumente“ und r als „Zahl der Dokumente aus R die Term t enthalten“ definiert sind. Liegen keine Informationen dieser Art vor (was in dieser Arbeit stets angenommen wird), reduziert sich die Formel zu einer IDF-Variante :

$$w_{IDF_{RSJ}} = \log \left(\frac{\mathrm{N} - \mathrm{n}_t + 0.5}{\mathrm{n}_t + 0.5} \right) \tag{5}$$

Es fällt auf, dass die Häufigkeit eines Terms innerhalb eines Dokuments keine Rolle bei der Berechnung spielt, was unweigerlich zu der Idee führt, den IDF-Wert mit der relativen Häufigkeit zu kombinieren. So wird sowohl die Information in einem Dokument, als auch die Information innerhalb der Dokumentkollektion zur Gewichtung genutzt.

TfIdf

Die Kombination aus relativer Termfrequenz und inverser Dokumenthäufigkeit gelingt durch Multiplikation der beiden Werte:

$$w_{tfIdf}(t, \mathbf{d}) = w_{tf}(t, \mathbf{d}) \cdot w_{IdF}(t) \tag{6}$$

Die Gewichtung nach TfIdf hat sich als äußerst robust herausgestellt. In verschiedenen Varianten findet sie in fast allen heutigen Termgewichtungsverfahren Verwendung [Robertson04].

Im SMART-System werden normalisierte TfIdf-Funktionen eingesetzt, um Anfra-

gen $\mathbf{q}$ zu gewichten: zum einen „ltc“ mit der Kosinusnormalisierung

$$w_{ltc}(t, \mathbf{q}) = \frac{w_{tf_{\log}}(t, \mathbf{q}) \cdot w_{IDF}(t)}{\sqrt{\sum_{i \in \mathbf{q}} (w_{tf_{\log}}(t, \mathbf{q}) \cdot w_{IDF}(t))^2}} \ ,$$

zum anderen „Ltu“

$$w_{Ltu} = \frac{\frac{\log\ tf(t,d)+1}{\log\ \mathrm{avg_tf}(d)} \cdot w_{IDF}(t)}{(1.0 - slope) \cdot pivot + slope \cdot |W_d|} \ , \tag{7}$$

mit der durchschnittlichen Termfrequenz (2) im Zähler und der „Pivoted Unique Normalization “ im Nenner.

Latent Semantic Indexing (LSI)

Der Hauptkritikpunkt am klassischen Vektorraummodell ist die Unabhängigkeitsannahme der Terme. Jeder Term steht senkrecht zu den übrigen, hat nichts mit den anderen gemein. In der Realität verbindet sie jedoch eine Vielzahl syntaktischer und semantischer Zusammenhänge. Ausgereifte linguistische Verfahren können das Problem bei der Erzeugung der Dokumentrepräsentation zwar dämpfen. Beseitigen können sie es jedoch nicht.

Ein Verfahren, das ganz ohne Vorverarbeitung im Sinne linguistischer Verfahren den Anspruch erhebt, semantische Konzepte aufzudecken, ist das „latent semantische Indexierungsmodell“ (LSI bzw. LSA für „Analyse“).

Dem Verfahren liegt die Vorstellung zu Grunde, dass Linearkombinationen in der Term-Dokumentmatrix eine Entsprechung in den erörterten linguistischen Phänomenen (Flexion, Synonyme, Nominalphrasen, ...) haben. Die Term-Dokument-Matrix ist dabei die Zusammenfassung aller Dokumentrepräsentationen zu einer Matrix. Mittels Singulärwertzerlegung wird die Term-Dokumentmatrix in einen Unterraum mit niedrigerer Dimensionalität projiziert, die Achsen stellen Linearkombinationen von Termvektoren (sog. Konzeptvektoren) dar [Dumais04]. Damit sind Termvektoren, wie die Dokumentrepräsentationen, im neuen Konzeptraum definiert und nicht mehr zwangsläufig orthogonal.

Auch die Anfragen werden in den Konzeptraum projiziert. Die Relevanz eines Dokuments bezüglich einer Anfrage ergibt sich so auf Basis der Konzepte mit der bemerkenswerten Folge, dass Dokumente als relevant eingestuft werden, obwohl sie die Terme der Query nicht enthalten.

2.2.2 Berechnung der Retrieval-Werte

Mit der Interpretation von Dokument- und Anfragerepräsentation als Vektoren im Raum bietet sich ein geometrisches Maß zur Bestimmung der Relevanz an. Der Retrieval-Wert eines Dokuments d bezüglich einer Anfrage q ergibt sich aus dem Skalarprodukt zwischen den gewichteten Termvektoren,

$$\rho_{\mathcal{R}_{\mathrm{VR}}}(\mathbf{q}, \mathbf{d}) = \sum_{t \in \mathbf{q}} \left([\mathbf{q}]_t \cdot [\mathbf{d}]_t \right) \ .$$

Im Falle von auf die Einheitslänge normierten Dokumentvektoren entspricht das Skalarprodukt dem Kosinus des Winkels zwischen Dokument und Anfrage (Abbildung 10),

$$\rho_{\mathcal{R}_{\mathrm{VR}}}(\mathbf{q}, \mathbf{d}) = \varphi_{\cos}(\mathbf{q}, \mathbf{d}) = \frac{\sum_{t \in \mathbf{q}} \left([\mathbf{q}]_t \cdot [\mathbf{d}]_t \right)}{||\mathbf{q}|| \cdot ||\mathbf{d}||} \ . \tag{8}$$

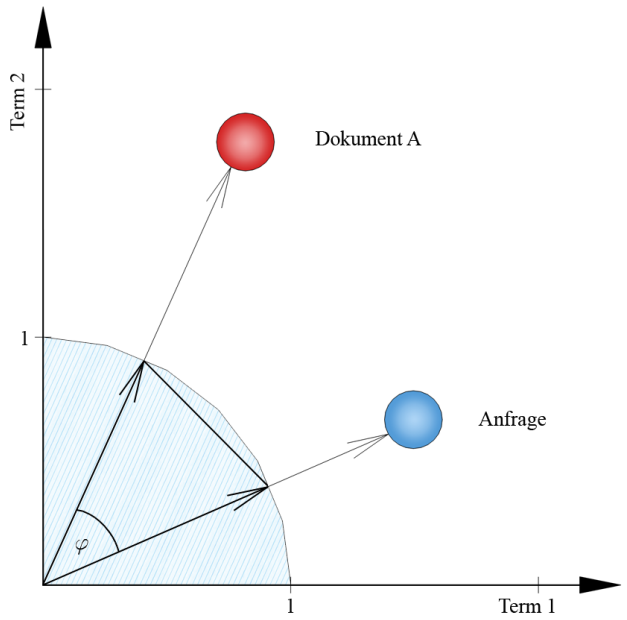


Figure 10: Relevanz als Kosinus des Winkels φ zwischen Anfrage und Dokument.

2.3 2-Poisson-Modell

Das 2-Poisson-Modell gehört zu den frühen probabilistischen Retrievalmodellen und wurde ursprüglich nur für das Auffinden guter Indexterme genutzt. Das Po-

tential des Ansatzes für die Berechnung von Termgewichten wurde erst später in Verfahren von Robertson und Walker [Robertson94] oder Amati und Rijsbergen [Amati02] genutzt. Es gründet auf Beobachtungen von Bookstein und Swanson [Swanson74] sowie Harter [Harter75]. Diese hatten festgestellt, dass die Signifikanz eines Terms über eine Hypothese bezüglich seiner Verteilung in der Dokumentkollektion beurteilt werden kann. Während unbedeutende Worte (v.a. Stoppwörter) sich einer Poisson-Verteilung folgend zufällig über die Dokumente verteilen, konzentrieren sich die von Harter „specialty words“ getauften, informativen Terme (wie z.B. Fachbegriffe) der „Poisson-Hypothese“ entgegen in einer kleinen Dokumentmenge.

Harter nimmt weiter an, dass sich ein informatives Wort (also „specialty word“) innerhalb einer kleinen Menge von „erlesenen“[1] Dokumenten dann wiederum gemäß einer (zweiten) Poisson-Verteilung verhält. Das Modell von Harter wird daher „2-Poisson-Modell“ genannt. Die folgenden Retrievalmodelle „Divergence From Randomness“ und „ Best Match“ machen von den Ergebnissen Harters Gebrauch.

2.4 Divergence From Randomness

In der von Amati und Rijsbergen [Amati02] veröffentlichen Arbeit werden die beschriebenen Ideen von Harter (2.3) aufgegriffen, verallgemeinert und einer parameterlosen Termgewichtung zugänglich gemacht.

Das Gewicht $w(t, \mathbf{d})$ eines Terms t errechnet sich nach Amati und Rijsbergen aus einer abfallenden Funktion über zwei Wahrscheinlichkeiten:

$$w_{DfR}(t, \mathbf{d}) \;=\; (-\log_2 \; Prob_1) \cdot (1 - Prob_2) \tag{9}$$

Der erste Faktor der Funktion gibt Auskunft über den *Informationsgehalt* eines Terms, der zweite Faktor ist Audruck des *Risikos,* Term t als guten Deskriptor für ein Dokument $d \in D$ zu wählen.

2.4.1 Informationsgehalt

Die Argumentation für die erste Wahrscheinlichkeit $Prob_1$ in der Gewichtungsfunktion folgt den Ausführungen von Harter zur zufälligen Verteilung unbedeutender Terme über die *gesamte* Dokumentkollektion. Die Poisson-Verteilung ist dabei

[1] Harter führt in seiner Arbeit den Begriff der „elite documents“ für diese Untermenge ein. Sie enthält für ihn alle Dokumente die vom semantischen Konzept hinter Term t handeln.

nur eines von mehreren möglichen Modellen. Neben ihr stehen weitere Approximationen einer Bernoulli-Verteilung, sowie Modelle mit Bose-Einstein-Statistik und Tf-Idf-Derivaten zur Auswahl. Insgesamt sind so sieben Basismodelle vorhanden, die die Wahrscheinlichkeit schätzen, dass die Termhäufigkeit $tf_{t,d}$ in Dokument d ein Produkt des Zufalls ist. Je kleiner diese Wahrscheinlichkeit, desto höher ist der Informationsgehalt des Terms, und desto wichtiger ist er zur inhaltlichen Beschreibung des Dokuments. Im Folgenden sind einige der Basismodelle beschrieben.

Bernoulli Modell

In den Bernoulli - Modellen wird angenommen, dass sich unbedeutende Terme binomial über die N Dokumente der Dokumentkollektion verteilen. Die Wahrscheinlichkeit, einen Term t ganze $tf_{t,d}$ Mal in einem Dokument anzutreffen, ergibt sich aus

$$Prob_1(tf_{t,d}) = B(N, F_t, tf_{t,d}) = \binom{F_t}{tf_{t,d}} p^{tf_{t,d}} q^{F_t - tf_{t,d}} ,$$

mit $p = 1/\mathrm{N}$, $q = (\mathrm{N}-1)/\mathrm{N}$ und F_t als Häufigkeit des untersuchten Terms t in D.

Unter Annahme einer großen Dokumentkollektion und einer konstanten Ereignisrate $\lambda = p \cdot F = const.$ lässt sich der Bernoulli-Prozess dann unter anderem mit der Poisson-Verteilung approximieren:

$$Prob_{1_P}(tf_{t,d}) = B(N, F_t, tf_{t,d}) \approx \frac{e^{-\lambda} \lambda^{tf_{t,d}}}{tf_{t,d}!} . \tag{10}$$

Bose-Einstein-Modell

In der Bose-Einstein-Statistik ist die Wahrscheinlichkeit, das ein beliebiges Dokument $d_k \in D$ exakt $tf_{t,d}$ Instanzen (engl. token) eines Terms t enthält, durch das Verhältnis s_2/s_1 gegeben, wobei s_1 die Anzahl der N-Tupel repräsentiert, welche die Gleichung

$$tf_{t,1} + tf_{t,2} + \cdots + tf_{t,N} = F_t$$

erfüllen (mit $tf_{t,k}$= Termhäufigkeit in Dokument d_k , N = Gesamtzahl der Dokumente und F_t = Anzahl der Token des Terms t in D). Die Menge ergibt sich aus dem Binomialkoeffizienten

$$s_1 = \binom{\mathrm{N} + \mathrm{F}_t - 1}{\mathrm{F}_t} = \frac{(\mathrm{N} + \mathrm{F}_t - 1)!}{(\mathrm{N} - 1)!\mathrm{F}_t!} \ .$$

Der Zähler s_2 des Verhältnisses geht von der Annahme aus, für Dokument d_k eine Beobachtung bzgl. seiner Termhäufigkeit $tf_{t,d}$ gemacht zu haben. Die Anzahl der (N-1)-Tupel, welche die Gleichung für die restlichen $\mathrm{F}_t - tf_{t,d}$ Token erfüllen, reduziert sich dadurch zu

$$s_2 = \binom{\mathrm{N} - 1 + (\mathrm{F}_t - tf_{t,d}) - 1}{\mathrm{F}_t - tf_{t,d}} = \frac{(\mathrm{N} + \mathrm{F}_t - tf_{t,d} - 2)!}{(\mathrm{N} - 2)!(\mathrm{F}_t - tf_{t,d})!} \ .$$

Mit der Annahme, dass $N \ggg tf$ und $\lambda = \mathrm{F}/\mathrm{N}$, lässt sich das Verhältnis s_2/s_1 z.B. über die geometrische Verteilung approximieren:

$$Prob_{1_{BE}}(tf_{t,d}) = \frac{s_2}{s_1} \approx \left(\frac{1}{1+\lambda}\right) \cdot \left(\frac{\lambda}{1+\lambda}\right)^{tf_{t,d}}$$

Tf-Idf-Modell

Der dritte Ansatz zur Berechnung von Wahrscheinlichkeit $Prob_1$ der Gewichtungsformel 9 läuft über den Satz von Bayes. Dabei wird zunächst nach der Wahrscheinlichkeit p gesucht, mit der ein zufällig gezogenes Dokument $d \in D$ den Term t enthält. Die a priori-Wahrscheinlichkeit dieses Ereignisses $P(X = p|N)$ ist unbekannt, a posteriori lässt sie sich, mit n_t als Anzahl der Dokumente die Term t enthalten, gemäß des Satzes von Bayes durch

$$P(x = p|\mathrm{n}_t, \mathrm{N}) = \frac{P(X = p|\mathrm{N})\ P(\mathrm{n}_t|\mathrm{N}, p)}{\sum_p P(X = p|\mathrm{N})\ P(\mathrm{n}_t|\mathrm{N}, p)}$$

angeben. Je größer die Anzahl der Dokumente N wird, desto näher kommt die a posteriori-Wahrscheinlichkeit, ungeachtet der wahren a priori-Wahrscheinlichkeit, der relativen Häufigkeit n_t/N . Nimmt man für die a priori-Verteilung der Token des Terms t in D eine beta-Form an (Dichtefunktion der Verteilung proportional zu $p^\alpha q^\beta$, $\alpha, \beta > -1$), ist die a posteriori-Wahrscheinlichkeit

$$P(x = p|\mathrm{n}_t, \mathrm{N}) = \frac{\mathrm{n}_t + 1 + \alpha}{\mathrm{N} + 2 + \alpha + \beta} \ .$$

Nimmt man weiter die Unabhängigkeit aller Token aller Terme voneinander und $\alpha = \beta = -0.5$ an, kann die Wahrscheinlichkeit, dass $tf_{t,d}$ Token des Terms t in Dokument d auftreten, mit

$$Prob_{1_{Idf}}(tf_{t,d}) = \left(\frac{\mathrm{n}_t + 1 + \alpha}{\mathrm{N} + 2 + \alpha + \beta}\right)^{tf_{t,d}} = \left(\frac{\mathrm{n}_t + 0.5}{\mathrm{N} + 1}\right)^{tf_{t,d}}$$

angegeben werden. Der erste Faktor der Gewichtungsfunktion 9 bekommt damit die Gestalt einer TfIdf-Gewichtungsvorschrift (vgl. Formel 6):

$$-\log_2 \; Prob_1 = tf_{t,d} \cdot \log_2 \frac{\mathrm{N} + 1}{\mathrm{n}_t + 0.5} \tag{11}$$

Eine Abwandlung zu Formel 11 schätzt die Zahl n_t der Dokumente in D mit Term t über eine Binomialverteilung:

$$\mathrm{n}_e = \mathrm{N} \cdot Prob(tf \neq 0) = \mathrm{N} \cdot (1 - B(\mathrm{N}, \mathrm{F}_t, 0)) = \mathrm{N} \cdot \left(1 - \left(\frac{\mathrm{N} - 1}{\mathrm{N}}\right)^{\mathrm{F_t}}\right) \tag{12}$$

2.4.2 Risiko

Der zweite Koeffizient $(1 - Prob_2)$ in der Gewichtungsformel ist Ausdruck des *Risikos,* einen Term t fälschlicherweise als guten Deskriptor für ein Dokument $d \in D$ anzunehmen. Es gilt: Je höher das Risiko, desto höher ist auch der Informationsgewinn, der durch den Term erzielt werden kann, sollte sich die Befürchtung als falsch herausstellen. Zur Berechnung gibt es zwei Alternativen. Die erste führt über Laplace's „law of succession“,

$$\frac{tf_{t,d}}{tf_{t,d} + 1}\,, \tag{13}$$

und betrachtet lediglich die Häufigkeit eines Terms t im Dokument d. Je seltener t, desto höher das Risiko. Der zweite Rechenweg bezieht sich auf die von Harter beschriebene „erlesene“ Dokumentmenge, die hier einfach als Menge der Dokumente $d \in D$ definiert ist, die den Term t enthalten[2]. Das Risiko ergibt sich aus dem Verhältnis zweier Bernoulli-Experimente. Das Experiment im Zähler gibt die Wahrscheinlichkeit an, dass bei Hinzunahme eines weiteren Tokens des Terms t in die „erlesenen“ Dokumente n_t die Termhäufigkeit $tf_{t,d}$ im betrachteten Dokument d um eins steigt (die neue Instanz also in Dokument d gelandet ist). Diese Wahrscheinlichkeit wird ins Verhältnis zur Ausgangssituation gesetzt:

[2] Durch diese Auslegung des „elite sets“ gelingt es, die Ereignisrate μ der Poisson-Verteilung parameterlos zu bestimmen.

$$Prob_2 = 1 - \alpha = 1 - \frac{B(\mathrm{n_t}, \mathrm{F_t} + 1, tf_{t,d} + 1)}{B(\mathrm{n_t}, \mathrm{F_t}, tf_{t,d})} = 1 - \frac{\mathrm{F_t} + 1}{\mathrm{n_t} \cdot (\mathrm{tf_{t,d}} + 1)} \tag{14}$$

$\mathrm{F_t}$ bezeichnet dabei die Anzahl der Token eines Terms t in der gesamten Dokumentkollektion. Bei einer (im Sinne der Bernoulli-Verteilung) unwahrscheinlich hohen Ausgangstermhäufigkeit $tf_{t,d}$ wird das Verhältnis kleiner eins werden, während bei einer kleinen Ausgangstermhäufigkeit der Zähler wahrscheinlicher wird. Damit ist das Verhältnis α direketer Audruck des Risikos und muss nicht noch, wie in der Gewichtsfunktion 9 angedacht, negiert werden ($1 - Prob_2 = 1 - (1 - \alpha) = \alpha$).

Das Risiko positiv in der Gewichtsfunktion zu etablieren, erscheint kontraproduktiv. Ein klar als signifikant erkennbarer Term trägt ein kleines Riskio, wohingegen Terme, die im Verhältnis zu anderen Dokumenten selten im Dokument d auftreten, großes Risiko bergen. In Kombination mit dem Informationsgehalt ergeben sich aber zwei konkurrierende Pole, die zusammen auch „sensible“ Signifikanzen erkennen sollen.

2.4.3 Normalisierung der Termhäufigkeiten

Die Normalisierung der Termhäufigkeiten bezüglich der Länge eines Dokumentes (= Anzahl der Token) wurde bereits im Abschnitt über das Vektorraummodell (2.2) thematisiert. Amati und Rijsbergen ziehen dafür zwei Hypothesen in Betracht:

H1 Die Terme in einem Dokument sind gleichverteilt. Die normalisierte Termfrequenz $tfn_{t,d}$ ergibt sich aus

$$tfn_{t,d} = tf_{t,d} \cdot \frac{avg_l}{l(d)} \ ,$$

mit avg_l als mittlere Dokumentlänge in D und $l(d)$ als Länge des Dokuments d.

H2 In einem kurzen Dokument ($l(d) < avg_l$) sind die Terme „dichter“ als in einem langen Dokument ($l(d) \geqq avg_l$). Sie werden nicht linear, sondern logarithmisch skaliert:

$$tfn_{t,d} = tf_{t,d} \cdot \log_2 \left(1 + \frac{avg_l}{l(d)}\right) \tag{15}$$

Die normalisierten Termhäufigkeiten $tfn_{t,d}$ ersetzen alle Vorkommen von $tf_{t,d}$ zur Bestimmung der Termgewichte.

2.4.4 Berechnung der Retrievalwerte

Die Retrievalwerte $\rho_{\mathcal{R}_{\mathrm{DfR}}}(\mathbf{q},\mathbf{d})$ zwischen eine Anfrage $\mathbf{q}$ und einer Dokumentrepräsentation $\mathbf{d}$ ergeben sich schließlich aus der Summe der Termgewichte $w_{DfR}(t,\mathbf{d})$, deren Terme t auch in der Anfrage $\mathbf{q}$ stehen.

$$\rho_{\mathcal{R}_{\mathrm{DfR}}}(\mathbf{q},\mathbf{d}) = \sum_{t\in q} qtf_{t,q} \cdot w_{DfR}(t,\mathbf{d}) \qquad = \\ = \sum_{t\in q} qtf_{t,q} \cdot (1 - Prob_2(tfn_{t,d})) \cdot (-log_2 Prob_1(tfn_{t,d}))\ ,$$

mit $qtf_{t,q}$ als Anzahl der Token des Terms t in Anfrage q.

2.5 Best Match-Modell

Das von Robertson und Walker [Robertson94] im Retrieval-System „Okapi“ eingesetzte Retieval-Modell umfasst eine Reihe an Funktionen, die unter dem Namen Best Match (BM) zusammengefasst sind. Sie sind von unterschiedlicher Komplexität und beinhalten gleich beide Komponenten, Termgewichtung $w(t,\mathbf{d})$ und Retrievalwertberechnung $\rho_{\mathcal{R}}(\mathbf{q},\mathbf{d})$. Im folgenden werden die Funktionen BM11, BM15 sowie BM25 hergeleitet.

Robertson/Sparck-Jones Formel

Ausgangspunkt der Betrachtungen ist die Robertson/Sparck-Jones-Formel (5):

$$w_{RSJ}(t,\mathbf{d}) \;=\; \log \frac{P(\mathrm{Ent_d}|\mathrm{Rel_{\mathbf{d}}})\ \left(1 - P(\mathrm{Ent_d}|\overline{\mathrm{Rel_{\mathbf{d}}}})\right)}{P(\mathrm{Ent_d}|\overline{\mathrm{Rel_{\mathbf{d}}}})\ (1 - P(\mathrm{Ent_d}|\mathrm{Rel_{\mathbf{d}}}))} \approx \\ \approx w_{IDF_{RSJ}} = \log \frac{N - n + 0.5}{n + 0.5}$$

mit $\mathrm{Rel_d}$ =„Dokument d ist relevant“ und $\mathrm{Ent_d}$ =„Dokument d enthält Term t“. Die Funktion w_{RSJ} ist Basis aller Best-Match-Formeln.

Hinzunahme der Termhäufigkeit

Zur Integration der Termhäufigkeit $tf_{t,d}$ wird die Robertson/Sparck-Jones-Formel als

$$w_{RSJ_{tf}}(\mathrm{t}, \mathbf{d}) = \log \frac{P(\mathrm{Ent_{tf,d}}|\mathrm{Rel_{\mathbf{d}}})\ P(\mathrm{Ent_{0,d}}|\overline{\mathrm{Rel_{\mathbf{d}}}})}{P(\mathrm{Ent_{tf,d}}|\overline{\mathrm{Rel_{\mathbf{d}}}})\ P(\mathrm{Ent_{0,d}}|\mathrm{Rel_{\mathbf{d}}})} \tag{16}$$

mit den Ereignissen $\mathrm{Rel_d}$ =„Dokument d ist relevant“ und $\mathrm{Ent_{n,d}}$ =„Dokument d enthält n Mal Term t“ umgeschrieben. Eine Abschätzung dieser Formel kann über die Annahmen des 2-Poisson-Modells von Harter gemacht werden. Mit μ als Ereignisrate der Poissonverteilung im „elite set“[3] und λ der Ereignisrate der Gesamtkollektion, $p' = P(\mathbf{d} \in elite\, set_{\mathrm{t}}|Rel_{\mathbf{d}})$ und $q' = P(\mathbf{d} \in elite\, set_{\mathrm{t}}|\overline{Rel_{\mathbf{d}}})$, ergibt die Kombination aus 2-Poisson-Modell und Formel 16,

$$w_{RSJ_{Poisson}}(t, \mathbf{d}) = \log \frac{(p'\lambda^{tf}e^{-\lambda} + (1-p')\mu^{tf}e^{-\mu})(q'e^{-\lambda} + (1-q')e^{-\mu})}{(q'\lambda^{tf}e^{-\lambda} + (1-q')\mu^{tf}e^{-\mu})(p'e^{-\lambda} + (1-p')e^{-\mu})}\ . \tag{17}$$

Formel 17 enthält für jeden Term vier Parameter, für die es keine direkten Anhaltspunkte gibt. Daher geht man dazu über, eine einfachere Funktion zu suchen, die ähnliche Eigenschaften besitzt.

Funktion 17

(a) wird Null für $tf_{t,d} = 0$,

(b) steigt monoton mit $tf_{t,d}$,

(c) bis zu einem asymptotischen Maximum,

(d) approximiert durch Formel 16.

Eine Funktion, die Eigenschaften (a) - (c) erfüllt, ist $tf_{t,d}/(const. + tf_{t,d})$, mit einem Maximum von 1. Da Formel 16 nicht direkt approximiert werden kann, wird Forderung (d) durch die gewöhnliche Robertson/Sparck-Jones-Formel (5) angenähert. Die Gewichtungsformel ergibt sich demnach zu

$$w_{BM15}(t, \mathbf{d}) = \frac{tf_{t,d}}{(k_1 + tf_{t,d})}\, w_{RSJ}\ , k_1\text{=const.}, \tag{18}$$

dem Gewichtungsterm der Best-Match-Formel BM15.

Normalisierung der Dokumentlänge

Momentan wird die Termhäufigkeit $tf_{t,d}$ nicht in Bezug auf die Länge des Do-

[3] Das „elite set“ wird hier nicht wie bei Amati auf die Menge der Dokumente d, die den Term t enthalten, ausgeweitet, sondern birgt einen „versteckten“ Parameter.

kumentes d relativiert. Dies kann mit

$$w_{BM11}(\mathrm{t}, \mathbf{d}) = \frac{tf_{t,d}}{\frac{k_1 \cdot l_d}{avg_l} + tf_{t,d}} \, w_{RSJ}$$

geändert werden (l_d = Anzahl Token in d, avg_l = mittlere Tokenzahl in der Kollektion).

Dies gibt den Gewichtungsterm bei BM11. In [Robertson95] heißt es, die Normalisierung bei BM11 würde „übertreiben“, wodurch man sich entschlossen hat, für BM25 einen Mix der Formel mit und ohne Normalisiserung einzusetzen:

$$w_{BM25}(t, \mathbf{d}) = \frac{tf_{t,d}}{(K + tf_{t,d})} \, w_{RSJ} \ ,$$

mit $K = k_1 \left((1-b) + b \frac{l_d}{avg_l} \right)$. Mit $b = 1$ ergibt sich BM11, mit $b = 0$ BM15.

Die hier vorgenommene Normalisierung bezog sich auf die Approximation 18 des 2-Poisson-Modells (17). Legt man das ursprüngliche Modell für die Normalisierung zugrunde entsteht ein weiterer, globaler Normalisierungsfaktor, der durch

$$correction = k_2 \cdot nq \frac{avg_l - l_d}{avg_l + l_d}$$

approximiert wird. nq ist dabei die Anzahl der Queryterme, k_2 ist eine weitere Unbekannte, die, wie k_1, per „trial and error“ bestimmt werden muss.

Querytermhäufigkeiten

Analog der Dokumentterme werden auch die Termhäufigkeiten qtf einer Anfrage $\mathbf{q}$ behandelt. Nach Formel 18 ist das Gewicht eines Anfrageterms t_q

$$w_{BM}(t, \mathbf{q}) = \frac{qtf}{(k_3 + qtf)} \cdot w_{RSJ} \ , k_3 = const.$$

Zusammenfassung der Faktoren

Insgesamt ergeben sich schließlich die finalen Best-Match-Formeln:

(BM15) $$\rho_{\mathcal{R}_{\mathrm{BM15}}}(\mathbf{q}, \mathbf{d}) = \sum_{t \in \mathbf{q}} \left(\frac{tf_{t,d}}{(k_1 + tf_{t,d})} \cdot \log \frac{N-n+0.5}{n+0.5} \cdot \frac{qtf}{(k_3+qtf)} + k_2 \cdot nq \frac{(avgl - l_d)}{(avgl + l_d)} \right)$$

(BM11) $$\rho_{\mathcal{R}_{\mathrm{BM11}}}(\mathbf{q}, \mathbf{d}) = \sum_{t \in \mathbf{q}} \left(\frac{tf_{t,d}}{(\frac{k_1 \cdot l_{\mathbf{d}}}{avgl} + tf_{t,d})} \cdot \log \frac{N-n+0.5}{n+0.5} \cdot \frac{qtf}{(k_3+qtf)} + k_2 \cdot nq \frac{(avgl - l_d)}{(avgl + l_d)} \right)$$

(BM25) $$\rho_{\mathcal{R}_{\mathrm{BM25}}}(\mathbf{q}, \mathbf{d}) = \sum_{\mathbf{t} \in \mathbf{q}} \left(\frac{tf_{t,d}}{(K + tf_{t,d})} \cdot \log \frac{N-n+0.5}{n+0.5} \cdot \frac{qtf}{(k_3+qtf)} + k_2 \cdot nq \frac{(avgl - l_d)}{(avgl + l_d)} \right)$$

Für die Konstanten wurden in den Experimenten bei TREC-4 [Robertson96] Werte von $k_1 = 1.0 - 2.0$, $k_3 = 8$ und $b = 0.6 - 0.75$ verwendet. Der Wert für k_2 war durchweg 0.

2.6 Statistisches Sprachmodell

Auch das statistische Sprachmodell von Ponte und Croft [Ponte98] gehört zu den probabilistischen Retrieval-Modellen. Im Gegensatz zu den bereits vorgestellten Ansätzen macht das statistische Sprachmodell aber keine parametrischen Annahmen über die generelle Verteilung von Termen in einer Dokumentkollektion (bei Harter z.B. poissonverteilt), sondern bezieht sich allein auf die konkret vorgefundenen Termhäufigkeiten.

Der Ausdruck „Sprachmodell“ kommt aus dem Feld der Spracherkennung und steht für eine Wahrscheinlichkeitsverteilung, die statistische Regelmäßigkeiten bei der Spracherzeugung erfasst [Yamron97]. Für den Gebrauch im Information Retrieval kann ein Sprachmodell $\mathcal{S}$ als ein Termgenerator verstanden werden, der einen Term $t \in T$ mit der Wahrscheinlichkeit $p(t|\mathcal{S})$ produziert. Jedes Dokument $d_i \in D$ wird als Produkt eines eigenständigen Sprachmodells $\mathcal{S}_{d_i}$ verstanden.

Der Retrievalwert $\rho_{\mathcal{R}_{\text{Ponte}}}(\mathbf{q}, \mathbf{d})$ wird mit der Wahrscheinlichkeit $p(\mathbf{q}|\mathcal{S}_{\text{d}})$ gleichgesetzt, mit der eine Anfrage $\mathbf{q}$ vom Sprachmodell hinter Dokument d erzeugt wird. Zur Vereinfachung wird angenommen, dass die Terme unabhängig voneinander sind. So lässt sich die Wahrscheinlichkeit $p(\mathbf{q}|\mathcal{S}_{\text{d}})$ als Multiplikation der Einzelwahrscheinlichkeiten für die Erzeugung eines Terms $t \in T$ formulieren,

$$\rho_{\mathcal{R}_{\text{Ponte}}}(\mathbf{q}, \mathbf{d}) = p(\mathbf{q}|\mathcal{S}_{\text{d}}) = \prod_{t \in q} p(t|\mathcal{S}_{\text{d}}) \cdot \prod_{t \notin q} 1 - p(t|\mathcal{S}_{\text{d}}) \ . \tag{19}$$

Im Folgenden wird eine Abschätzung für die Wahrscheinlichkeiten $p(t_i \in T|\mathcal{S}_{\text{d}})$ gesucht.

Für ein Sprachmodell $\mathcal{S}_d$ ist bekannt, dass es das Dokument d hervorgebracht hat. Ausgehend von der Termverteilung in d ist die Maximum-Likelihood-Abschätzung für die Erzeugung eines Terms $t \in T$ von Sprachmodell $\mathcal{S}_d$

$$p_{ml}(t|\mathcal{S}_{\text{d}}) = \frac{tf_{t,d}}{l_d} \ ,$$

mit l_d als Anzahl der Token in Dokument d. Diese Abschätzung ist für Ponte und Croft schwach, da sie sich allein auf die Termverteilung in Dokument d der

bestimmten Länge l_d stützt. Deshalb wird eine zusätzliche, robustere Abschätzung $p_{avg}(t|\mathcal{S}_\mathrm{d})$ eingeführt.

Sei dafür D_t die Menge der Dokumente, für die gilt:

$$d \in D_t \Leftrightarrow t \in d$$

Die mittlere Wahrscheinlichkeit $p_{avg}(t)$, mit der ein Term t von den Sprachmodellen $\mathcal{S}_d : d \in D_t$ erzeugt wird, ergibt sich aus

$$p_{avg}(t|\mathcal{S}_\mathrm{d}) = \frac{\left(\sum_{d \in D_t} p_{ml}(t|\mathcal{S}_\mathrm{d})\right)}{\mathrm{n}_t} ,$$

mit n_t als Anzahl der Dokumente, die den Term t enthalten[4]. Diese Abschätzung ist robust in dem Sinne, dass sie auf eine größere Menge an Dokumenten gründet. Allerdings kann im Allgemeinen nicht davon ausgegangen werden, dass die Sprachmodelle den unterstellten Zusammenhang tatsächlich aufweisen. Der Gebrauch dieser Abschätzung ist daher risikant. Das Risiko ist umso höher, je weiter die Termhäufigkeit $tf_{t,d}$ im Dokument d eines Sprachmodells $\mathcal{S}_d : d \in D_t$ von der durchschnittlichen Häufigkeit $\bar{f}_t = p_{avg}(t|\mathcal{S}_d) \cdot l_d$ des Terms t abweicht. Zur Modellierung des Risikos $R_{t,d}$ verwenden Ponte und Croft die geometrische Verteilung:

$$R_{t,d} = \left(\frac{1}{(1+\bar{f}_t)}\right) \cdot \left(\frac{\bar{f}_t}{(1+\bar{f}_t)}\right)^{tf_{t,d}}$$

Ergibt sich ein kleines Risiko $R_{t,d}$ für die Abschätzung $p_{avg}(t|\mathcal{S}_\mathrm{d})$ über den Durchschnitt, wird sie zur Berechnung der Termwahrscheinlichkeit $p(t|\mathcal{S}_\mathrm{d})$ benutzt. Bei einem hohem Risiko $R_{t,d}$ wird auf die Maximum-Likelihood-Abschätzung $p_{ml}(t|\mathcal{S}_\mathrm{d})$ zurückgegriffen. Dazwischen werden die beiden Abschätzungen kombiniert:

$$p(t|\mathcal{S}_\mathrm{d}) = p_{ml}(t|\mathcal{S}_\mathrm{d})^{(1-R_{t,d})} \cdot p_{avg}(t|\mathcal{S}_\mathrm{d})^{R_{t,d}} \tag{20}$$

Betrachtet man noch einmal die Berechnung des Retrievalwertes in Formel 19, dann fällt auf, dass sich der Wert durch *Multiplikation* der Einzelwahrscheinlichkeiten $p(t|\mathcal{S}_\mathrm{d})$ ergibt. Ein Term, der in einem Dokument d nicht vorkommt, wird nach der Maximum-Likelihood-Abschätzung auch nie vom Sprachmodell $\mathcal{S}_d$

[4] In Analogie zur Definition des „elite sets“ bei Amati.

erzeugt. Durch die multiplikative Verknüpfung erzeugt jedes Sprachmodell, das auch nur einen Term der Query nicht enthält, die Query mit einer Wahrscheinlichkeit von Null. Um diesen sicher zu harten Ausschluss zu vermeiden, wird die Berechnung der Termwahrscheinlichkeiten $p(t|\mathcal{S}_\mathrm{d})$ nochmals modifiziert. Mit F_t als Anzahl der Token eines Terms t in der Dokumentkollektion D und F_T als Gesamtzahl aller Token in D lautet die endgültige Fassung

$$p(t|\mathcal{S}_\mathrm{d}) = \begin{cases} p_{ml}(t|\mathcal{S}_\mathrm{d})^{(1-R_{t,d})} \cdot p_{avg}(t|\mathcal{S}_\mathrm{d})^{R_{t,d}} & , falls\, tf_{t,d} > 0 \\ \frac{\mathrm{F}_t}{\mathrm{F}_\mathrm{T}} & sonst. \end{cases} . \tag{21}$$

Durch die Fallunterscheidung erzeugt ein Sprachmodell $\mathcal{S}_d$ auch dokumentfremde Terme, was prinzipiell eine mächtige Eigenschaft ist. Auf der anderen Seite sorgt die Fallunterscheidung auch dafür, dass *jedes* Dokument $d \in D$ einen Retrieval-Wert größer Null bezüglich einer Anfrage erreicht. Im Zusammenhang mit der (später folgenden) Verwendung der Retrieval-Modelle im Dokumenten-Clustering ist dies eine unangenehme Feststellung (siehe dazu Kapitel 3.5). Daher hat sich der Autor entschlossen, für die Verwendung des statistischen Sprachmodells im Dokument-Clustering folgende Vorschrift für die Ermittlung der Retrieval-Werte einzusetzen:

$$p(t|\mathcal{S}_\mathrm{d}) = p_{ml}(t|\mathcal{S}_\mathrm{d})^{(1-R_{t,d})} \cdot p_{avg}(t|\mathcal{S}_\mathrm{d})^{R_{t,d}} ,$$

$$\rho_{\mathcal{R}_\mathrm{Ponte2}}(\mathbf{d}_i, \mathbf{d}_j) = \sum_{t \in T} p(t|\mathcal{S}_\mathrm{d_i}) \cdot p(t|\mathcal{S}_\mathrm{d_j}) . \tag{22}$$

Analog zu den anderen vorgestellten Retrieval-Modellen ergibt sich der Retrieval-Wert aus dem Skalarprodukt zwischen $\mathbf{d}_i, \mathbf{d}_j \in \mathbf{D}$.

2.7 Explizit Semantisches Indexierungsmodell

Das Retrieval-Modell von Gabrilovich und Markovitch [Gabrilovich07] verwendet zur Relevanz-Bestimmung eine externe Wissensquelle C, die freie Online-Enzyklopädie *Wikipedia*.[5] Das Interessante an der Enzyklopädie ist, dass sich die rund 250.000 Artikel[6] jeweils einem bestimmten Thema widmen und umfassend

[5] www.wikipedia.org, englischsprachige Version.

[6] Nach Entfernung sehr kleiner Artikel und solcher mit wenigen eingehenden und ausgehenden Verweisen.

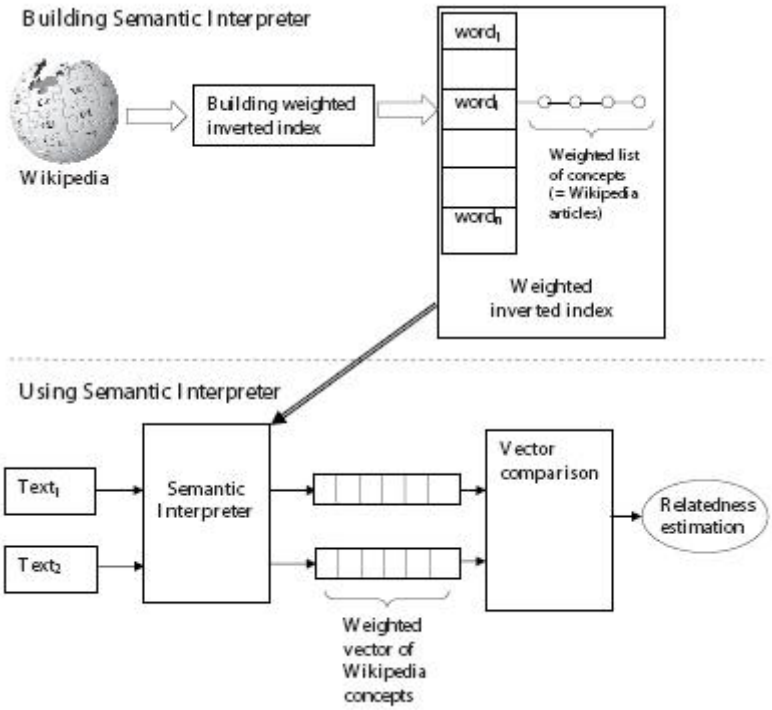


Abbildung 11: ESA-Szenario. Über die Gewichte der Artikelrepräsentationen $\mathbf{c} \in \mathbf{C}$ und der Dokumentrepräsentationen $\mathbf{d} \in \mathbf{D}$ wird jedes Dokument auf einen gewichteten Vektor von Wikipedia-Artikeln $\mathbf{k} \in \mathbf{K}$ abgebildet. (Quelle: [Gabrilovich07])

erörtert wird, was im Allgemeinen darunter zu verstehen ist. Dadurch ist das in einem Artikel verwendete Vokabular sehr speziell und kann als Indikator für eine semantische Beziehung zwischen einem Dokument und einer Anfrage eingesetzt werden.

Folgendes Szenario ist mit dem Modell von Garilovich und Markovitch möglich:

Sei Term t_x Element einer Anfrage q, $d \in D$ ein Dokument mit dem Term t_y. Weiterhin sei ein Artikel $c \in C$, der sowohl t_x als auch t_y enthält. Dokument d wird vom Retrieval-System als relevant erkannt, obwohl Dokument d den Term t_x nicht enthält.

Die Herleitung der Verfahrensvorschrift beginnt mit der Abbildung der Dokumente $d \in D$ und Artikel $c \in C$ auf Repräsentationen $\mathbf{d} \in \mathbf{D}$ bzw. $\mathbf{c} \in \mathbf{C}$. Für die Abbildung wird gemäß dem Vektorraummodell mit TfIdf-Gewichtung (Formel 6) verfahren. Zu beachten ist, dass sich der Idf-Wert der Gewichtungsvorschrift für die Dokumente $d \in D$ auf die Termverteilung in der Artikelmenge C bezieht.

Sei nun $\mathbf{k} \in \mathbf{K}$ ein Vektor der Dimension $|\mathbf{C}|$. Sei $\mathbf{k}$ *Konzeptvektor* genannt. Über die Vorschrift

$$\forall \mathbf{c}_i \in \mathbf{C} \,:\, [\mathbf{k}]_i = \sum_{t \in c_i} [\mathbf{d}]_t \cdot [\mathbf{c}_i]_t \tag{23}$$

wird jede Dokumentrepräsenation $\mathbf{d} \in \mathbf{D}$ auf einen Konzeptvektor $\mathbf{k} \in \mathbf{K}$ abgebildet. Ein Dokument d ist damit nicht länger über seinen Termvektor $\mathbf{d}$, sondern über einen gewichteten Vektor $\mathbf{k}$ von Wikipedia-Artikeln c repräsentiert (Abbildung 11).

Zur Relevanzberechnung wird eine Anfrage $\mathbf{q}$ mit Vorschrift 23 auf einen Konzeptvektor $\mathbf{l}$ abgebildet. Der Retrievalwert zwischen q und Dokument d ergibt sich aus der Kosinusähnlichkeit zwischen den Konzeptvektoren,

$$\rho_{\mathcal{RESA}}(\mathbf{l}, \mathbf{k}) = \varphi_{\cos}(\mathbf{l}, \mathbf{k}) = \frac{\sum_{c \in \mathbf{l}} (\mathbf{l}_c \cdot \mathbf{k}_c)}{||\mathbf{l}|| \cdot ||\mathbf{k}||} \,.$$

3 Dokumenten-Clustering

Mit den vorgestellten Retrieval-Modellen lassen sich die Terme (bzw. Konzepte) der Dokumentrepräsentationen *gewichten* und die Dokumente nach Relevanz bezüglich einer Anfrage *sortieren*. Eine Sortierung wird dann erfolgreich sein, wenn der Nutzer seinen Informationsbedarf adäquat und eindeutig mit der Anfrage zum Ausdruck gebracht hat. Leicht kann es jedoch passieren, dass sich die entstandene (in der Regel sehr kurze) Formulierung im Kontext der Dokumentkollektion als mehrdeutig erweist und unter verschiedenen Gesichtspunkten Relevanz festgestellt wird. Das Ergebnis ist dann eine thematisch inhomogene, für den Nutzer unbefriedigende Dokumentsammlung. Nach der Cluster-Hypothese von Rijsbergen eröffnet das Dokumenten-Clustering eine Möglichkeit zur Verbesserung.

3.1 Cluster-Hypothese

Die Integration der Cluster-Analyse in den Retrieval-Prozess wird mit der sog. *Cluster-Hypothese* gerechtfertigt. Sie besagt, dass thematisch eng verbundene Dokumente dazu neigen, für denselben Informationsbedarf relevant zu sein [vanRijsbergen79]. In Kapitel 2 wurde gezeigt, wie Retrieval-Modelle auf Basis von Ähnlichkeiten in der Termverteilung den thematischen Zusammenhang zwischen Anfrage und Dokument ermitteln. Mit der Anfrage als ein zweites Dokument aus der Dokumentkollektion sollte es möglich sein, eine semantische Aussage auch über Dokumentpaare machen zu können (Abbildung 12).

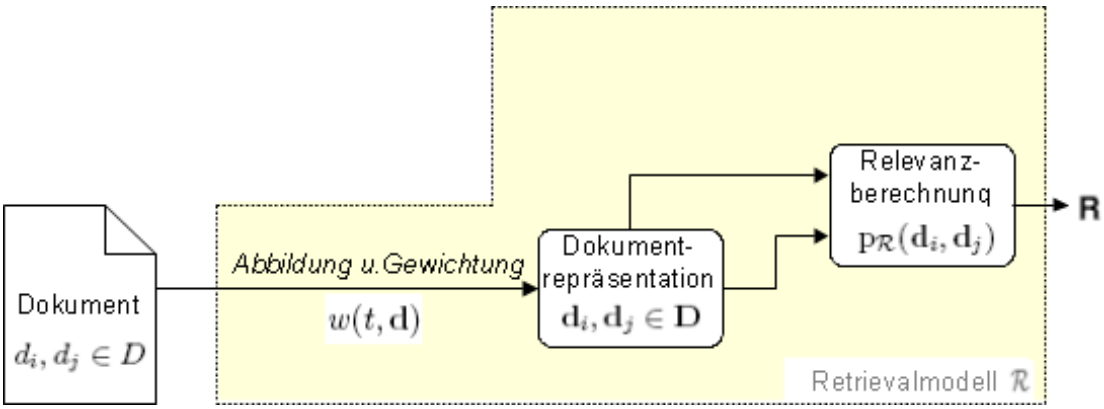


Abbildung 12: Ähnlichkeit zwischen Dokumentpaaren. Die Retrieval-Modelle aus Kapitel 2 können für die Berechnung der Ähnlichkeiten zwischen den Dokumenten der Kollektion verwendet werden.

Die Cluster-Hypothese lässt sich somit auch wie folgt formulieren. Die *Ähnlichkei-*

ten unter den relevanten Dokumenten (bzgl. eines Informationsbedarfs) sind größer als die zwischen relevanten und irrelevanten Dokumenten [vanRijsbergen79].

Die Cluster-Analyse ermöglicht es, die Dokumentkollektion a priori so zu strukturieren, dass sich ähnliche Dokumente jeweils zusammen in einer Klasse befinden. Strenggenommen würde sich der Retrieval-Prozess nach der Hypothese so auf das Auffinden passender Klassen reduzieren. Diese enthielten dann alle relevanten Dokumente. Leider erwies sich diese Kalkulation jedoch als nicht perfomant (gezeigt z.B. in [Vorhees85]), weshalb die Cluster-Hypothese in jüngerer Zeit häufig abgeschwächt wird. Nach Hearst und Pedersen [Hearst96] sollte nicht unterstellt werden, dass zwei Dokumente, die für eine Anfrage q_1 relevant oder irrelevant sind, dies ebenso für alle anderen Anfragen $q_{2,\ldots}$ sind. Vielmehr kann sich angesichts der hohen Dimensionalität der Dokumentvektoren Ähnlichkeit aus verschiedenen Teilmengen der Merkmale herausdefinieren. Infolgedessen bietet es sich an, das Dokumenten-Clustering erst a posteriori, also auf die Menge D^* der als bereits relevant eingestuften Dokumente, anzuwenden. Durch den Bezug aller Dokumente zur Anfrage konzentrieren sich die signifikanten Unterschiede dort auf einen Bruchteil der Terme, was die Effektivität des Verfahrens entscheidend verbessert.

3.2 Definition und Ähnlichkeitsmodell

Die Clusteranalyse zählt zu den strukturentdeckenden[7] Verfahren der multivariaten Statistik und wird zur Zerlegung einer heterogenen Objektmenge in homogene Teilmengen eingesetzt.

Ein (exklusives) *Clustering* C der Objektmenge D,

$$\mathcal{C} = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mathrm{k}\},\ \mathrm{C}_\mathrm{i} \subset \mathrm{D}\ ,$$

ist die Zerlegung von D in paarweise disjunkte Mengen C_i mit $\bigcup_{\mathrm{C}_\mathrm{i} \in \mathcal{C}} \mathrm{C}_\mathrm{i} = \mathrm{D}$ [Stein04a]. Maxime zur Aufteilung ist die Minimierung der Ähnlichkeit zwischen den Gruppen C_i bei gleichzeitiger Maximierung der Ähnlichkeiten innerhalb einer Gruppe.

Ausgangspunkt der Analyse bildet die *Rohdatenmatrix,* in der die Eigenschaften der Objekte eingetragen sind. Im Kontext des Dokumentenclusterings bildet sich die Matrix aus den Dokumentrepräsentationen $\mathbf{d} \in \mathbf{D}$ und deren Termen t. Die

[7] In Abgrenzung zu strukturprüfenden Verfahren wie der Varianzanalyse oder logistischen Regression.

Felder der Matrix enthalten die berechneten Termgewichte $w(\mathbf{d}, \mathrm{t})$.

Rohdatenmatrix	Term 1	Term 2	.....	Term |T|
Dokument 1	$w(\mathbf{d_1}, \mathbf{t_1})$	$w(\mathbf{d_1}, \mathbf{t_2})$	$\cdots$	$w(\mathbf{d_1}, \mathbf{t_{\|T\|}})$
Dokument 2	$w(\mathbf{d_2}, \mathbf{t_1})$	$w(\mathbf{d_2}, \mathbf{t_2})$		$\vdots$
$\vdots$				$\vdots$
Dokument N	$w(\mathbf{d_N}, \mathbf{t_1})$	...	...	$w(\mathbf{d_N}, \mathbf{t_{\|T\|}})$

Im ersten von drei Analyseschritten wird die Rohdatenmatrix in ein *Ähnlichkeitsmodell* eingeführt. Dieses enthält die quantifizierten Ähnlichkeiten zwischen jedem Dokumentenpaar $(\mathbf{d_i}, \mathbf{d_j}) \in \mathbf{D} \times \mathbf{D}$ in der Ähnlichkeitsmatrix M.

Ähnlichkeitsmatrix	Dokument 1	Dokument 2	$\cdots$	Dokument N
Dokument 1	1	$\varphi(\mathbf{d_1}, \mathbf{d_2})$		$\varphi(\mathbf{d_1}, \mathbf{d_N})$
Dokument 2	$\varphi(\mathbf{d_2}, \mathbf{d_1})$	1		$\varphi(\mathbf{d_2}, \mathbf{d_N})$
$\vdots$			$\ddots$	$\vdots$
Dokument N		$\cdots$	$\cdots$	1

Ein Maß φ zur Bestimmung der Ähnlichkeitswerte ist beispielsweise der Q-Korrelationskoeffizient,

$$\varphi_Q(\mathbf{d}_i, \mathbf{d}_j) = \frac{\sum_{t \in d}(tf_{t,d_i} - \overline{tf_{d_i}}) \cdot (tf_{t,d_j} - \overline{tf_{d_j}})}{\sqrt{\sum_{t \in d}(tf_{t,d_i} - \overline{tf_{d_i}})^2 \cdot \sum_{t \in d}(tf_{t,d_j} - \overline{tf_{d_j}})^2}} \ ,$$

mit $\overline{tf_d}$ als Durchschnittswert aller Termhäufigkeiten in Dokument d. Im betrachten Fall wird die Retrieval-Funktion $\rho_{\mathcal{R}}$ des zur Termgewichtsberechnung eingesetzten Retrieval-Modells zur Ähnlichkeitsbestimmung verwendet.

Für die nachfolgenden Betrachtungen ist es mitunter hilfreich, sich die Ähnlichkeitsmatrix als Adjazenzmatrix eines gewichteten, ungerichteten Graphen G(V,E,w) vorzustellen (Abbildung 13). Die Dokumente werden als Knoten V modelliert, die Ähnlichkeiten bilden sich im Gewicht $w(e)$ einer Kante $e \in$ E ab.

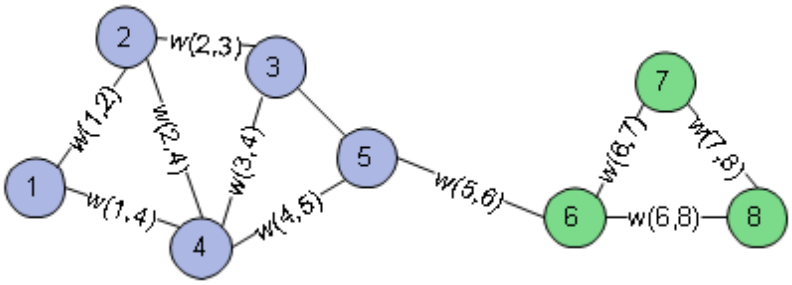


Abbildung 13: Auszug aus einem Ähnlichkeitsgraphen G(V,E,w). Die Dokumentmenge D wird bijektiv auf die Menge der Knoten V abgebildet, die Ähnlichkeit $\varphi(\mathbf{d}_i, \mathbf{d}_j)$ entspricht dem Gewicht $w(\{u, v\})$ der Kante $\{u, v\} \in E$ zwischen den mit $\mathbf{d}_i$ und $\mathbf{d}_j$ assoziierten Knoten u und v [Stein04a].

3.3 Fusionierungsalgorithmen

Anhand des Ähnlichkeitsmodells werden im zweiten Analyseschritt Anzahl und Zusammensetzung der Cluster bestimmt. Es gilt, unter den $\sum_{k=1}^{N} {}^{k^N}/_{k!}$ möglichen Clusterings [Jain90] dasjenige zu bestimmen, das dem Informationsbedarf des Anfragestellers am besten nachkommt. Über welche Kriterien das gesuchte Clustering gefunden werden kann, ist dabei zunächst offen und gibt Platz für ein breites Spektrum an Algorithmen. Diese sind, angesichts der kombinatorischen Komplexität, oft Heuristiken zur Annäherung an ein statistisches Zielkriterium. Abbildung 14 zeigt eine Taxonomie der bekannten Fusionierungsalgorithmen.

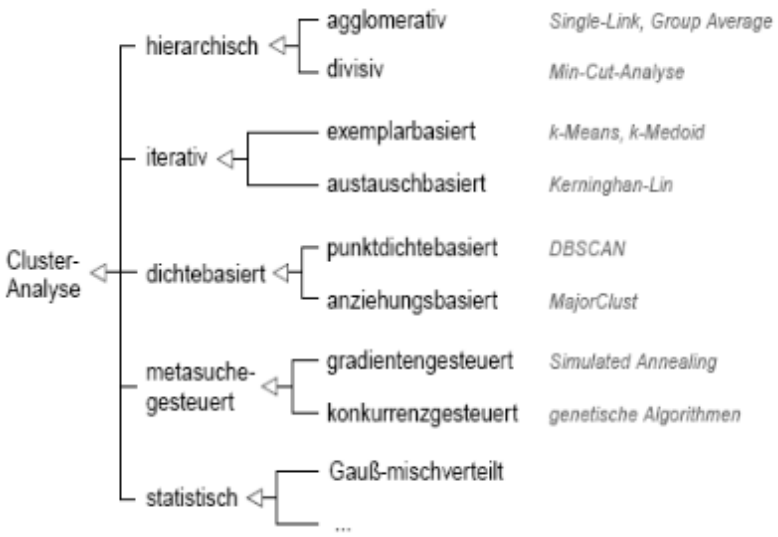


Abbildung 14: Taxonomie der Fusionierungsalgorithmen (Quelle: [Stein04a])

Im folgenden wird auf die in den Experimenten (Kapitel 5) eingesetzten Verfahren

Bezug genommen.

3.3.1 Hierarchisch agglomerative Fusionierung

Bei den hierarchischen Verfahren unterscheidet man weiter zwischen divisiven und agglomerativen Algorithmen. Während ein divisiver Ansatz von einem einzigen, gemeinsamen Cluster ausgeht, der im Laufe der Gruppierung zerlegt wird, steckt im Ausgangsstadium eines agglomerativen Verfahrens jedes Objekt in einem eigenen Cluster. In jedem Durchlauf der Fusionierung wird dann die Ähnlichkeit zwischen allen Clusterpaaren $(C, C') \in \mathcal{C}\text{x}\mathcal{C}$ bestimmt. Sie ergibt sich, nach einer Ähnlichkeitsfunktion φ_C, aus den Ähnlichkeiten zwischen den in den Clustern enthaltenen Objekten.

Verbreitete Ähnlichkeitsfunktionen sind:

Single-Link	$\varphi_C(C, C') = \max\limits_{\substack{\mathbf{d}_i \in C, \\ \mathbf{d}_j \in C'}} \varphi(\mathbf{d}_i, \mathbf{d}_j)$
Complete-Link	$\varphi_C(C, C') = \min\limits_{\substack{\mathbf{d}_i \in C, \\ \mathbf{d}_j \in C'}} \varphi(\mathbf{d}_i, \mathbf{d}_j)$
Group-Average-Link	$\varphi_C(C, C') = \frac{1}{C \cdot C'} \sum\limits_{\substack{\mathbf{d}_i \in C, \\ \mathbf{d}_j \in C'}} \varphi(\mathbf{d}_i, \mathbf{d}_j)$
Ward	$\varphi_C(C, C') = \sqrt{\frac{2 \cdot \lvert C \rvert \cdot \lvert C' \rvert}{\lvert C \rvert + \lvert C' \rvert}} \, \lvert\lvert \overline{\mathbf{d}}_i(C) - \overline{\mathbf{d}}_j(C') \rvert\rvert$

Die beiden Cluster mit der größten Ähnlichkeit φ_C zueinander[8] werden zusammengefasst und die Ähnlichkeitswerte $\varphi_C(C \cup C', C_i)$ für den nächsten Durchlauf ermittelt.

Eine effiziente Berechnung dieser ermöglicht die Lance-Williams Formel

$$\varphi_C(C \cup C', C_i) =$$
$$a \cdot \varphi_C C, C') + b \cdot \varphi_C(C, C_i) + c \cdot \varphi_C(C', C_i) + d \cdot \lvert \varphi_C(C, C_i) - \varphi_C(C', C_i) \rvert \, ,$$

wobei die Konstanten a, b, c und d je nach verwendeter Ähnlichkeitsfunktion variieren.

[8] Beim Ward-Verfahren sind dies die Cluster welche die Heterogenität (Varianz) am wenigsten vergrößern.

Verfahren	a	b	c	d
Single-Link	0.5	0.5	0	-0.5
Complete-Link	0.5	0.5	0	0.5
Average-Link	0.5	0.5	0	0
Ward	$\frac{\lvert C\rvert+\lvert C'\rvert}{\lvert C\rvert+\lvert C'\rvert+\lvert C_i\rvert}$	$\frac{\lvert C\rvert+\lvert C_i\rvert}{\lvert C\rvert+\lvert C'\rvert+\lvert C_i\rvert}$	$-\frac{\lvert C\rvert}{\lvert C\rvert+\lvert C'\rvert+\lvert C_i\rvert}$	0

mit $|C|, |C'|, |C_i|$ = Zahl der Objekte in C, C', C_i [Backhaus06].

Der Fusionierungsalgorithmus endet, wenn alle Cluster zu einem einzigen zusammengefasst sind. Die Stelle, an der die beste Einteilung der Objekte erreicht wurde, wird dabei nicht wahrgenommen. Erst in einem Folgeschritt kann sie aus dem Vergleich der im Prozess aufgetretenen Konfigurationen ermittelt werden. Automatisierte Verfahren zur Emittlung der Clusterzahl werden in Abschnitt 3.4 vorgestellt. Veranschaulichen lässt sich der Agglomerationsprozess anhand eines Dendrogramms, beispielhaft rechts in Abbildung 15 zu sehen. Intuitiv ist zu erkennen, dass die richtige Zahl an Clustern k* = 2 beträgt.

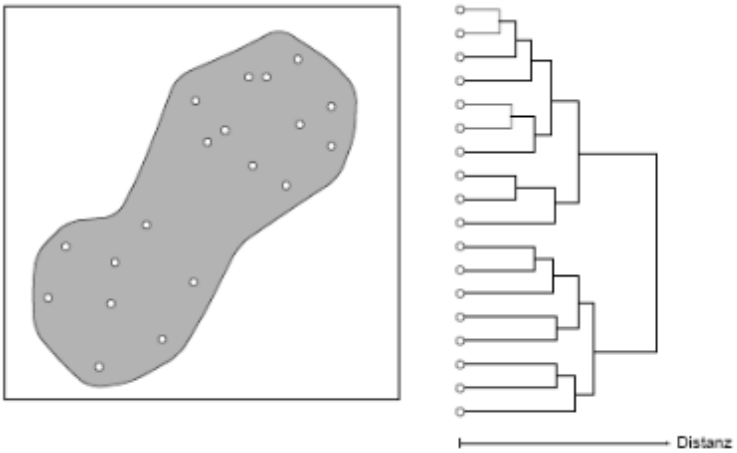


Abbildung 15: Dendrogramm einer beendeten, agglomerativen Fusionierung (Quelle: [Stein04b]).

3.3.2 Exemplarbasierte, iterative Fusionierung

Die iterativen Fusionierungsmethoden starten mit einer festgelegten Anzahl k an Clustern. Sie ändern die Zuordnung der Objekte zu den Clustern so lange, bis ein vorgegebenes Zielkriterium erreicht ist. Exemplarbasierte Verfahren repräsentieren die Cluster dabei durch sog. Centroiden (bzw. Medoiden), die im Merkmalsraum der Objekte definiert und zunächst zufällig initialisiert sind. Die Clusterzugehörigkeit x eines Objektes $\mathbf{d}$ ergibt sich in jeder Iteration aus der

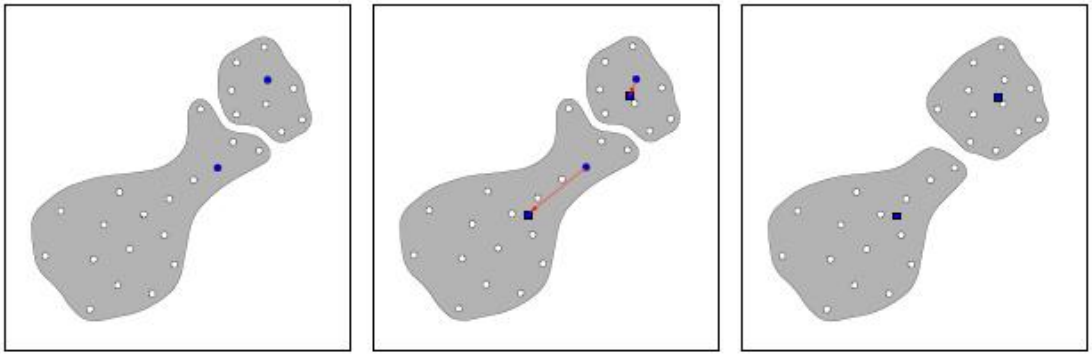


Abbildung 16: Iterationsschritt des k-Means-Algorithmus. Nachdem die Objekte zugeordnet sind, werden die Centroiden (blau) dem Zielkriterium gemäß verschoben. (Quelle: [Stein04c])

Ähnlichkeit mit den Centroiden $\mathbf{c}_i$:

$$x = \arg\max_{i=1,\ldots,k} \varphi(\mathbf{c}_i, \mathbf{d}) \,,\; C_x = C_x \cup \mathbf{d}$$

Die Lage der Centroiden im Merkmalsraum wird durch ein Zielkriterium gesteuert. Ein Centroid wird so versetzt, dass das Kriterium in Bezug auf die ihm zugeordneten Objekte optimiert wird. Das gebräuchlichste Zielkriterium ist das Varianzkriterium, das die Minimierung der clusterspezifischen Fehlerquadratsummen anstrebt (k-Means). Für einen Centroiden $\mathbf{c}_i$ ergibt sich die neue Position aus dem arithmetischen Mittelwert $\overline{\mathbf{d}_{C_i}}$ aller enthaltenen Objekte $\mathbf{d}_{C_i}$.

Eine interessante Variante ist das Fuzzy-k-Means-Verfahren. Ein Objekt $\mathbf{d}$ gehört hier mit einem individuellen Zugehörigkeitswert $\mu_{\mathbf{d}_i} \in [0, 1]$ zu allen Clustern. Die aktuelle Position eines Centroiden $\mathbf{c}_i$ ergibt sich aus

$$\mathbf{c}_i = \frac{\sum_{\mathbf{d}\in\mathbf{D}} {\mu_{\mathbf{d}_i}}^2 \cdot \mathbf{d}}{\sum_{\mathbf{d}\in\mathbf{D}} {\mu_{\mathbf{d}_i}}^2} \,,\; \mu_{\mathbf{d}_i} = \frac{1}{\sum\limits_{j\in C} \left(\frac{\varphi(\mathbf{d},\mathbf{c}_i)}{\varphi(\mathbf{d},\mathbf{c_j})}\right)^2} \,.$$

Der Einsatz der unscharfen Objektzuordnung ist immer dann nützlich, wenn sich nicht alle Objekte klar einer Gruppe zuordnen lassen [Stein04c].

Durch die Verschiebung der Centroiden ergibt sich die Ausgangslage für die nächste Iteration (Abbildng 16). Die Schritte Objektzuordnung und Centroidverschiebung werden so lange wiederholt, bis die Verschiebung einen Grenzwert ε unterschreitet.

Im Gegensatz zu den hierarchischen Verfahren haben iterativen Algorithmen den Vorteil, dass gebildete Gruppen im Laufe der Fusionierung wieder aufgelöst wer-

den können. Gemein haben sie jedoch dass Problem, die richtige Anzahl der Cluster separat ermitteln zu müssen. Hinzu kommt, dass das Ergebnis der Fusionierung abhängig von der Reihenfolge ist, in der die Objekte betrachtet werden.

3.3.3 Dichtebasierte Fusionierung mit MajorClust

Dichtebasierte Clusteralgorithmen verstehen eine homogene Teilmenge des Objektvorrats als Bereich mit gleicher Dichte im Graphen G = (V,E,w) der Ähnlichkeitsmatrix. Die Zuordnung eines Knotens (bzw. Objektes) zu einem Cluster erfolgt über ein Prinzip zur Dichteschätzung. Der MajorClust - Algorithmus interpretiert die Kantengewichte zwischen den Knoten (bzw. Ähnlichkeiten zwischen Objekten) als Kräfte, wobei Knoten desselben Clusters ihre Kraft bündeln [Stein04d]. In der Ausgangskonfiguration des Algorithmus steckt jeder Knoten $\mathbf{d}_i$ allein in einem Cluster C_i. In einem Fusionierungsdurchgang wird für jeden Knoten $\mathbf{d}_i$ der Cluster C_x ermittelt, der die stärkste Anziehungskraft auf ihn auswirkt.

$$x = \arg \max_{j=1,\ldots,|V|} f(C_j, \mathbf{d}_i)\,, \quad C_x = C_x \cup \mathbf{d}_i$$

Die Anziehungskraft eines Clusters C_j ergibt sich aus der Summe aller Kantengewichte die adjazent zu $\mathbf{d}_i$ sind, und zu einem Knoten $\mathbf{d}_k$ aus C_j führen.

$$f(C_j, \mathbf{d}) = \sum_{\mathbf{d}_k : \mathbf{d}_k \in C_j} w(\mathbf{d}_i, \mathbf{d}_k)$$

Die Fusionierung wird so lange fortgeführt, bis sich in einem Durchgang keine Änderungen in den Cluster-Zugehörigkeiten mehr ergeben. Das Major-Clust-Verfahren hat den großen Vorteil, dass die Anzahl der Cluster selbständig ermittelt wird. Nachteilig erweist sich, dass das Verfahren abhängig von der Reihenfolge, in der die Knoten besucht werden, unterschiedliche Ergebnisse erzeugt.

3.4 Emittlung der besten Clusterings

Hierarchische wie iterative Fusionierungsalgorithmen stellen keine Ermittlungen hinsichtlich der angemessenen Clusteranzahl an. Während hierarchische Verfahren in jedem Fusionierungsdurchlauf ein Clustering bilden, können iterative Prinzipien zur Erzeugung verschiedener Clusterlösungen mit unterschiedlichen Vorgaben für die Clusteranzahl k gestartet werden. Im dritten Schritt der Clusteranalyse

besteht die Aufgabe nun darin, das „beste“ Clustering unter den Lösungsvorschlägen zu nominieren. Das für diese Zwecke am häufigsten vorgestellte Verfahren ist das Ellbow-Kriterium, ein weiteres Verfahren ist die Gap-Statistik. Theoretisch stehen auch die allgemeineren internen Qualitätsmaße (Abschnitt 4.2) zur Nominierung des wohlgeformtesten Clusterings zur Verfügung.

3.4.1 Elbow-Kriterium

Ein übliches Qualitätsmerkmal eines Clusterings $\mathcal{C}_k$ ist die Heterogenität. Sie beschreibt, wie die Ähnlichkeiten zwischen den Objekten $\mathbf{d}_j \in C_i$ in einem Cluster $C_i \in \mathcal{C}_k$ verteilt sind. Ein Maß für Heterogenität ist beipielsweise die Fehlerquadratsumme

$$\mathrm{v}_{C_i} = \sum_{\mathbf{d} \in C_i} \sum_{t \in \mathbf{d}} ([\mathbf{d}]_t - [\overline{\mathbf{d}_{C_i}}]_t)^2 \ ,$$

mit $\overline{\mathbf{d}_{C_i}}$ als Mittelwert der Merkmale $t \in d$ in den Objekten aus Cluster C_i. Die Gesamtfehlersumme eines Clusterings $v_{\mathbf{C}_k}$ ist die Summe über die Einzelfehler v_{C_i}. Typischerweise nimmt die Heterogenität in einer Serie an Clusterings $\mathbf{C} = \{\mathcal{C}_1, ..., \mathcal{C}_k, ..., \mathcal{C}_n\}$ automatisch mit Zunahme der Anzahl k an Clusterinstanzen ab. Das Elbow-Kriterium nominiert als bestes k jenes, bei dem der Abfall der Heterogenität bezüglich seines Vorgängers am stärksten ist. Trägt man die Heterogenitätsentwicklung gegen die Clusteranzahl in einem Diagramm ab, bildet sich an der Stelle des stärksten Abfalls der namensgebende „Ellenbogen“ (Abbildung 17).

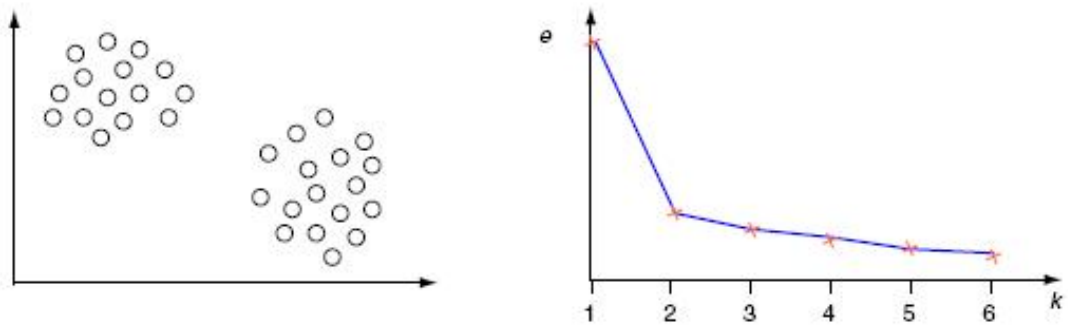


Abbildung 17: Elbow-Kriterium. Beim Auftragen der Heterogenität gegen die Anzahl der Cluster bildet sich an der Stelle der optimalen Clusteranzahl ein „Ellenbogen“ (Quelle: [MeyerZuEissen07]).

3.4.2 Gap-Statistik

Wie das Elbow-Kriterium betrachtet auch die Gap - Statistik die Heterogenitätsentwicklung einer Serie an Clusterings mit unterschiedlicher Clusteranzahl. Zusätzlich wird jedoch ein Referenzverlauf herangezogen, der die Entwicklung einer im Wertebereich des tatsächlichen Clusterings gleichverteilten Objektmenge beschreibt. Der Referenzverlauf schätzt den Heterogenitätsverlust, der allein durch Zunahme einer weiteren Clusterinstanz zu erwarten ist. Die Differenz

$$\mathrm{Gap}(k) = (\mathrm{E}^* \cdot \log(\mathrm{V}_{\mathcal{C}_k})) - log(\mathrm{V}_{\mathcal{C}_k})$$

beschreibt, mit $\mathrm{E}^* \cdot \log(\mathrm{V}_{\mathcal{C}_k})$ als Erwartungswert unter dem Referenzverlauf[9] und $\mathrm{V}_{\mathcal{C}_k}$ als Maß der Heterogenität (Fehlerquadratsumme), den Teil des Heterogenitätsverlustes, der durch einen Qualitätszuwachs im Clustering $\mathcal{C}_k$ verursacht wird. Die optimale Anzahl an Clusterinstanzen k^* maximiert den Wert für $\mathrm{Gap}(k)$. Bemerkenswert ist, dass die Gap-Statistik im Gegensatz zu anderen Verfahren auch einen Wert für das triviale Clustering mit $k = 1$ produziert. So ist eine Aussage darüber möglich, ob eine Objektmenge überhaupt in Teilmengen zerlegt werden sollte.

3.5 Modellvereinfachung

Die zur Erstellung der Ähnlichkeitsmatrix verwendeten Retrieval-Modelle erzeugen für ein Dokumentpaar $(\mathbf{d_i}, \mathbf{d_j}) \in \mathbf{D} \times \mathbf{D}$ dann, und nur dann, einen Ähnlichkeitswert von exakt Null, wenn die beiden Dokumentrepräsentationen ein komplett unterschiedliches Vokabular verwenden, bzw. voneinander verschiedene Konzepte bedienen.[10] Beim statistischen Sprachmodell von Ponte und Croft ist die „absolute Unähnlichkeit“ durch die Fallunterscheidung in Formel 21 ganz ausgeschlossen.

Ist disjunktes Vokabular bei der Relevanzberechnung zu einer Nutzeranfrage mit wenigen Termen noch wahrscheinlich, sinkt diese Erwartung beim Vergleich der weitaus längeren Dokumente aus D untereinander erheblich. Sind zwei Dokumente thematisch nicht verwandt, verbindet sie daher meist dennoch eine insignifikante Überschneidung im Termgebrauch. Die daraus resultierenden, kleinen Ähnlich-

[9] Einzelheiten zur Generierung der Referenzfunktion finden sich in [Tibshirani00].

[10] Modelle mit einer IDF-Komponente lassen noch die gemeinsame Verwendung eines Terms zu, der auch in allen anderen Dokumenten verwendet wird.

keitswerte finden auf semantischer Ebene keine sinnträchtige Entsprechung und werden deshalb im Folgenden als Rauschen interpretiert.

Naturgemäß entsteht Rauschen zwischen Dokumenten aus unterschiedlichen Themenbereichen. Trägt man bei bekannter Referenzkategorisierung die Ähnlichkeitswerte, die zwischen den Themenbereichen bestehen, neben den Ähnlichkeiten innerhalb der Gruppen auf, lässt sich ein Eindruck von der Beschaffenheit des Rauschanteils gewinnen.[11] Abbildung 18 zeigt für eine Dokumentkollektion aus zehn Gruppen zu jeweils 1000 Nachrichtenmeldungen die aufsummierten Ähnlichkeitswerte in zehn Intervallen zwischen 0 und 1. Zur Berechnung der Ähnlichkeiten wurde, nach Stoppwortentfernung und Einsatz des Porter-Stemming-Algorithmus [Porter80], das Vektorraummodell mit Tf-Gewichtung (Formel 3) verwendet.

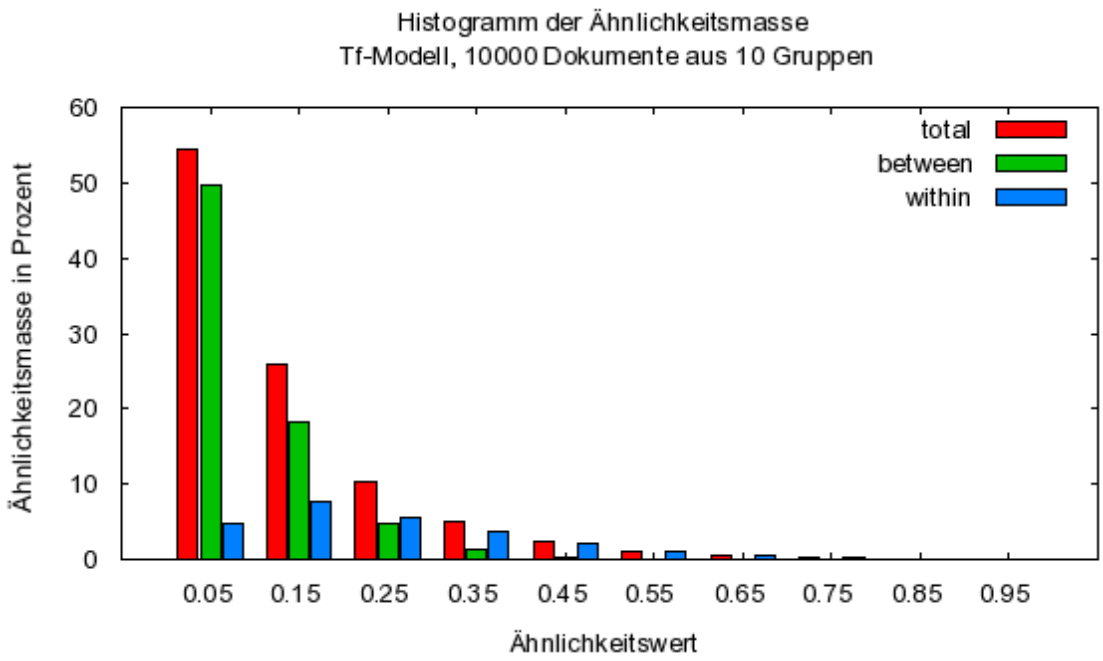


Abbildung 18: Masse-Verteilung im Ähnlichkeitsgraphen. Die kleinsten Ähnlichkeitswerte steuern aufgrund ihrer Vielzahl den größten Teil zur Gesamtmasse bei (rote Balken). Die meisten Werte stammen dabei von Dokumentpaaren $(\mathbf{d_i}, \mathbf{d_j})$ aus unterschiedlichen Clustern $\mathbf{d}_i \in \mathbf{C}_i$,$\mathbf{d}_j \in \mathbf{C}_{j \neq i}$ (grüne Balken). Mit Zunahme des Ähnlichkeitswertes steigt der Anteil informationstragender Ähnlichkeiten (blaue Balken).

Es ist deutlich zu erkennen, dass Rauschen (grüne Balken) trotz minimaler Einzelwerte einen beachtlichen Anteil der Gesamtmasse (74,5%) ausmacht. Für Fusionierungsprinzipien, die auf Akkumulation von Ähnlichkeitswerten beruhen (z.B.

[11] Rauschen ist dabei stets nur eine (wenn auch sehr große) Teilmenge der Ähnlichkeitswerte zwischen Clustern. Im Hinblick auf die Fusionierung können die Ausdrücke jedoch synonym verwendet werden, da es sich in jedem Fall um störende Einflüsse handelt.

Major Clust), erwächst daraus die Gefahr, falsche Mehrheitsentscheidungen zu treffen (Abbildung 19).

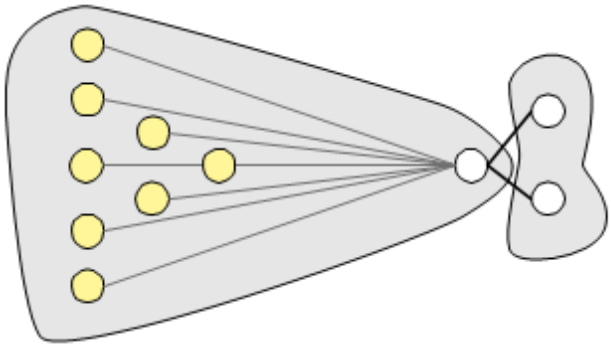


Abbildung 19: Falsche Mehrheitsentscheidung, verursacht durch Rauschen. Die vielen kleinen Ähnlichkeiten aus der Dokumentgruppe links erzeugen in Summe eine größere Anziehungskraft als die beiden schweren Kanten rechts. Die Fusionierung endet mit einem einzigen, gemeinsamen Cluster.

Retrieval-Modelle, die eine IDF-Komponente zur Gewichtsberechnung verwenden, bewerten gemeinhin benutzte Terme weniger stark. Gerade Rauschen, hervorgerufen durch zufällige Termübereinstimmungen, sollte von dieser Abwertung betroffen sein. Abbildung 20 zeigt die obige Zusammensetzung der Ähnlichkeitsmasse generiert mit Tf-Idf-Gewichtung (Formel 6).

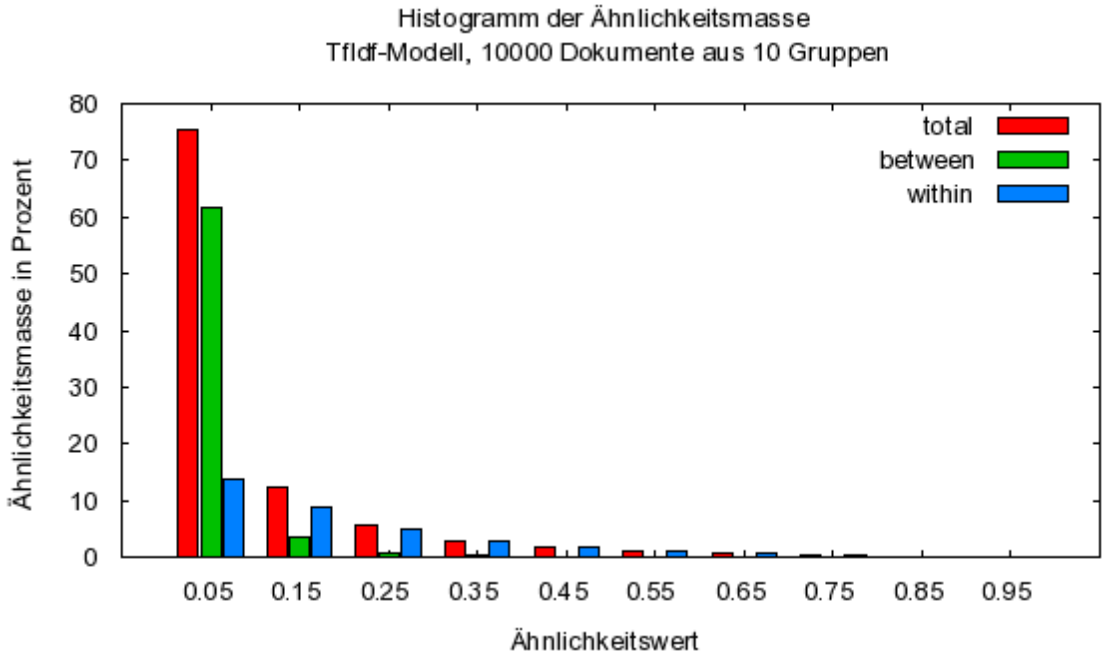


Abbildung 20: Verteilung der Ähnlichkeitsmasse unter Verwendung des Tf-Idf-Modells. Im Vergleich zur reinen Tf-Gewichtung setzen sich Ähnlichkeiten innerhalb der Gruppen (blau) früher von den Ähnlichkeiten zwischen Gruppen (grün) ab. Ihr Anteil an der Gesamtmasse bleibt dabei nahezu unverändert.

Durch die IDF-Komponente konzentriert sich das Rauschen (grün) sichtbar in den kleinsten Ähnlichkeitswerten. Bemerkenswert ist, dass der Anteil des Rauschens an der Gesamtmasse dabei jedoch nur um 8% abnimmt. Wirft man einen genaueren Blick auf die Verteilung innerhalb der kleinen Ähnlichkeitswerte im Intervall [0,0.4], fällt auf, dass sich die Masse der informationstragenden Ähnlichkeitswerte wie bei Verwendung der Tf-Gewichtung (Abbildung 18) mit den rauschenden Ähnlichkeitswerten überschneidet (Abbildung 21). Eine Entschärfung der Fusionierungsproblematik ist unter diesem Gesichtspunkt daher nicht zu erwarten.

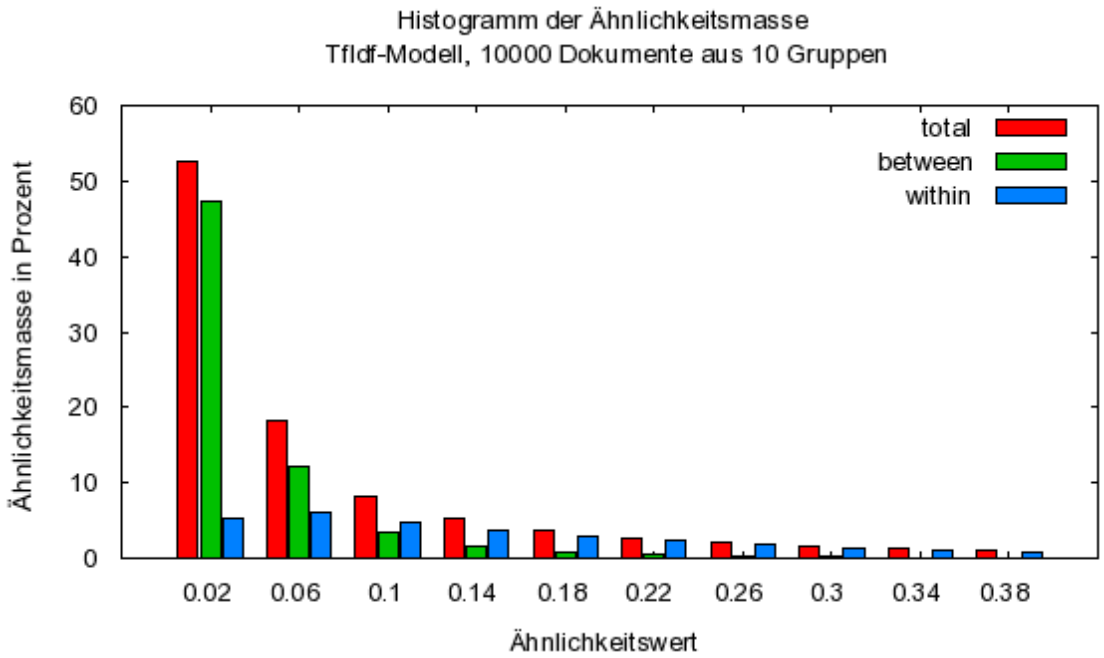


Abbildung 21: Ähnlichkeitsmasse im Intervall [0, 0.4] des Tf-Idf-Modells. Bei genauererer Betrachtung der Ähnlichkeitsverteilung in den kleinen Wertebereichen ergibt sich die Verteilungscharakteristik des Tf-Modells.

Tatsächlich erreicht man eine Modellverbesserung mit der expliziten Eliminierung kleinster Ähnlichkeitswerte. Im obigen Fall mit IDF-Komponente ließe sich durch Entfernen aller Ähnlichkeiten kleiner 0,1 zum Beispiel der Großteil des Rauschens beseitigen. Dieser Maßnahme würden allerdings auch gut 40% der informationstragenden Werte zum Opfer fallen. Die im weiteren beschriebenen Verfahren zur Modellvereinfachung gehen daher dokumentspezifisch vor. Die These ist, dass sich die Überlappung von Rauschen und Information bei kleinen Ähnlichkeitswerten (im Beispiel mit TfIdf-Gewichtung im Intervall [0, 0.3]) durch die dokumentspezifische Betrachtung besser auflöst. Dokumente, die innerhalb des eigenen Clusters mit kleinen Ähnlichkeitswerten angebunden sind, haben mit cluster-fremden Dokumenten entsprechend noch kleinere Ähnlichkeit.

Ein Standardverfahren, mit dem sich dokumentspezifisch Ähnlichkeitswerte eliminieren lassen, ist die Verwendung eines k-Nearest-Neighbour-Graphen. Darüberhinaus wurden im Rahmen dieser Arbeit neue, parameterlose Ansätze zur Modellvereinfachung entwickelt, die ebenfalls im Folgenden vorgestellt werden.

3.5.1 k-Nearest-Neighbour-Graph

Im k-Nearest-Neighbour-Graphen (kNN-Graph) $\mathrm{G}_{knn}(\mathrm{V}, \mathrm{E}_{knn}, w)$ eines Graphen $\mathrm{G}(\mathrm{V}, \mathrm{E}, w)$ bleiben jeweils die k schwersten Kanten erhalten, die inzident mit einem Knoten $v \in \mathrm{V}$ sind. Eine Variante des Algorithmus, der *gegenseitige kNN-Graph*, entfernt zusätzlich alle Kanten $\{u, v\} \in \mathrm{E}$, die nicht unter den k schwersten Kanten *beider* definierenden Knoten u und v aus V sind.

Liegen in einem Clustering Cluster gleicher Größe vor, lässt sich mit einem Wert für k gleich der Cluster-Größe eine gute Modellvereinfachung mit dem gegenseitigen k-Nearest-Neighbour-Graphen erreichen. Dies liegt daran, dass unter den k schwersten Kanten eines Knotes $v \in \mathrm{V}$ tendenziell die $k-1$ Kanten sind, die aus dem Cluster des Knotens v stammen. Kanten die zu Knoten aus anderen Clustern führen werden dann vom Algorithmus eliminiert. Die Größe der Cluster ist im Vorfeld natürlich nicht bekannt, ebenso kann die uniforme Verteilung der Dokumente nicht angenommen werden. Es bleibt für die direkte Verwendung des kNN-Graphen somit nur das Ausprobieren mehrerer Werte für k, mit anschließender Ermittlung des besten Ergebnisses (Ausführungen dazu in Kapitel 4.2).

3.5.2 Expected Similarity

Das unter dem Namen „Expected Similarity“ (*erwartete Ähnlichkeit*) entwicklelte Konzept zur Modellvereinfachung schätzt auf Grundlage der Termgewichtsverteilung in der Dokumentkollektion für jede Dokumentrepräsentation $\mathbf{d}_i \in \mathbf{D}$, einen zu *erwartenden Ähnlichkeitswert* $\overline{\rho}(\mathbf{d}_i)$. Einem Dokumentpaar $(\mathbf{d}_i, \mathbf{d}_j) \in \mathbf{D} \times \mathbf{D}$ mit zugehörigem Retrieval-Wert $\rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j)$ wird dann eine thematische Verwandtschaft zugesprochen, wenn der Retrievalwert einen signifikant höheren Wert annimmt als die erwarteten Ähnlichkeiten $\overline{\rho}(\mathbf{d}_i)$ und $\overline{\rho}(\mathbf{d}_j)$. Liegt der Wert unter oder in der Nähe einer der erwarteten Ähnlichkeiten, wird Rauschen unterstellt und der Ähnlichkeitswert $\varphi(\mathbf{d}_i, \mathbf{d}_j)$ zwischen den beiden Dokumenten in der Ähnlichkeitsmatrix M auf Null gesetzt.

Zur Feststellung der erwarteten Ähnlichkeit $\overline{\rho_{\mathcal{R}}} : \mathbf{D} \times \overline{\mathbf{D}} \rightarrow \mathcal{R}$ wird zunächst das *mittlere Referenzdokument* $\overline{\mathbf{d}} \in \overline{\mathbf{D}}$ erzeugt, das sich aus den arithmetischen Mittelwerten der Termgewichte $w(t, \mathbf{d})$ in den Dokumenten aus $\mathbf{D}$ zusammensetzt,

$$[\overline{\mathbf{d}}]_t = \frac{\sum_{i=1}^{\mathrm{N}} w(t, \mathbf{d}_i)}{\mathrm{N}} \ . \tag{24}$$

Für jedes Dokument $\mathbf{d}_i \in \mathbf{D}$ ergibt sich die *erwartete Ähnlichkeit* $\overline{\rho_{\mathbf{d}}}(\mathbf{d}_i)$ aus dem

Retrieval-Wert mit dem Referenzdokument $\overline{\mathbf{d}}$,

$$\overline{\rho_{\mathbf{d}}}(\mathbf{d}_i) = \rho_{\mathcal{R}}(\mathbf{d}_i, \overline{\mathbf{d}}) \ .$$

Sei für die folgenden Betrachtungen ein Term t_1 angenommen, der in nur wenigen Dokumentrepräsentationen $\mathbf{D}_1 \subset \mathbf{D}$ ein Gewicht verschieden von Null annimmt. Der Argumentation für die IDF-Gewichtung (2.2.1 auf Seite 25) folgend, ist Term t_1 zur inhaltlichen Beschreibung der Dokumente aus $\mathbf{D}_1$ wichtig. Das *erwartete Gewicht* des Terms, $w(t_1, \overline{\mathbf{d}})$, wird aufgrund der wenigen Summanden größer Null im Zähler von Formel 24 deutlich unter den Gewichten $w(t_1, \mathbf{d}|\mathbf{d} \in \mathbf{D}_1)$ der Dokumente aus $\mathbf{D}_1$ liegen (Abbildung 22, rote Elemente). Für Dokumentrepräsentationen $\mathbf{d}_i, \mathbf{d}_j \in \mathbf{D}_1$ gilt somit die Ungleichung

$$[\mathbf{d}_i]_{t_1} \cdot [\mathbf{d}_j]_{t_1} > \max_{k \in \{i,j\}} \left([\mathbf{d}_k]_{t_1} \cdot [\overline{\mathbf{d}}]_{t_1} \right) \ . \tag{25}$$

Nimmt man, ohne Verlust der Allgemeingültigkeit, für die folgenden Betrachtungen das Skalarprodukt hinter der Retrieval-Funktion $\rho_{\mathcal{R}}$ an, drückt Ungleichung 25 aus, dass einer der N Summanden in der Retrieval-Wert-Berechnung $\rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j)$ größer ist als die entsprechenden Summanden bei der Berechnung der erwarteten Ähnlichkeiten $\overline{\rho_{\mathbf{d}}}(\mathbf{d}_i)$ und $\overline{\rho_{\mathbf{d}}}(\mathbf{d}_j)$. Die Wahrscheinlichkeit $p(X)$, mit dem Ereignis $X = \left(\rho(\mathbf{d}_i, \mathbf{d}_j) > \max_{k \in \{i,j\}} \overline{\rho_{\mathbf{d}}}(\mathbf{d}_k)\right)$, ist mit der Beobachtung bei Term t_1 also gestiegen.

Sei nun Term t_2 ein in D weit verbreiteter Term, $\mathbf{D}_2 \subseteq \mathbf{D}$ die Menge der Dokumentrepräsentationen, deren Dokumente t_2 enthalten. Im Gegensatz zu Term t_1 besitzt Term t_2 kaum Diskriminanzkraft. Die Termgewichte $w(t_2, \mathbf{d} : \mathbf{d} \in \mathbf{D}_2)$ streuen aufgrund der vielen Gewichte größer Null in Formel 24 um das erwartete Gewicht $w(t_2, \overline{\mathbf{d}})$ (Abbildung 22, grüne Elemente). Für zwei Dokumentrepräsentationen $\mathbf{d}_m, \mathbf{d}_n \in \mathbf{D}_2$ ist es daher möglich, dass sich die Wahrscheinlichkeit $p(X)$, einen Retrieval-Wert $\rho_{\mathcal{R}}(\mathbf{d}_m, \mathbf{d}_n)$ größer als $\overline{\rho_{\mathbf{d}}}(\mathbf{d}_m)$ und $\overline{\rho_{\mathbf{d}}}(\mathbf{d}_n)$ zu erreichen, mit dem Wert für den Summanden $[\mathbf{d}_m]_{t_2} \cdot [\mathbf{d}_n]_{t_2}$ verringert. Wie die folgende Überlegung zeigt, ist im Mittel, durch Beobachtung von Term t_2, jedoch keine Verkleinerung der Wahrscheinlichkeit $p(X)$ zu erwarten.

Sei $\overline{w}_{t_2}$ das mittlere Gewicht von Term t_2 in den Dokumentrepräsentationen aus $\mathbf{D}_2$,

$$\overline{w}_{t_2} = \frac{\sum_{i=1}^{\mathrm{N}} w(t_2, \mathbf{d}_i)}{|\mathbf{D}_2|}.$$

Das Produkt $\overline{w}_{t_2} \cdot \overline{w}_{t_2}$ stellt, bei Annahme des Skalarprodukts, den gemittelten

Summanden für Term t_2 in der Retrievalfunktion $\rho_{\mathcal{R}}(\mathbf{d}_m, \mathbf{d}_n)$ dar. Entsprechend entsteht im Mittel der Summand $\overline{w}_{t_2} \cdot [\overline{\mathbf{d}}]_{t_1}$ zur Bestimmung der erwarteten Ähnlichkeiten $\overline{\rho_{\mathbf{d}}}(\mathbf{d}_m)$ und $\overline{\rho_{\mathbf{d}}}(\mathbf{d}_n)$.

Da $|\mathbf{D}_2| \leq$ N und damit $\overline{w}_{t_2} \geq [\overline{\mathbf{d}}]_{t_2}$ gilt

$$\overline{w}_{t_2} \cdot \overline{w}_{t_2} \geq \overline{w}_{t_2} \cdot [\overline{\mathbf{d}}]_{t_1} \ . \tag{26}$$

Wie bei Ungleichung 25 folgt aus 26 ein Anstieg (bzw. bei $|\mathbf{D}_2| =$ N auch Gleichbleiben) der Wahrscheinlichkeit $p(X)$.

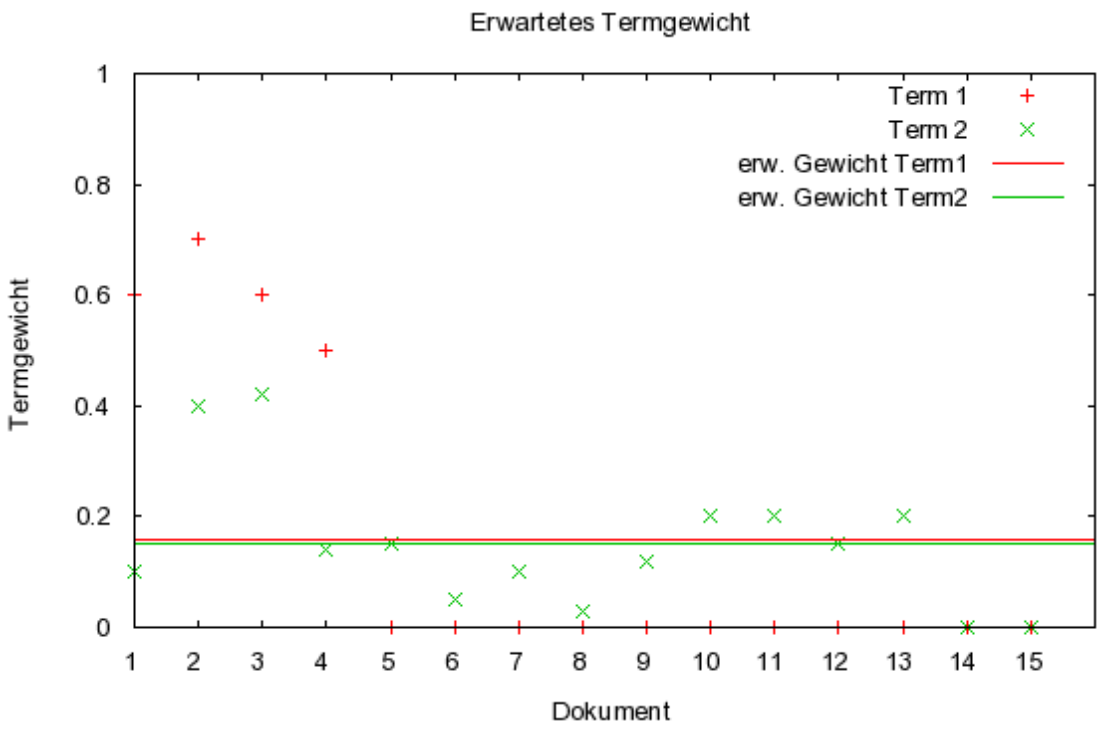


Abbildung 22: Erwartetes Termgewicht $w(t, \overline{\mathbf{d}})$. Während für Term t_1 das erwartete Termgewicht $w(t_1, \overline{\mathbf{d}})$ weit unter den Gewichten $w(t_1, \mathbf{d}_{1,\ldots,4})$ der, den Term t_1 enthaltenden, Dokumente 1 bis 4 liegt (rote Elemente), streuen die Termgewichte $w(t_2, \mathbf{d}_{1,\ldots,13})$ von Term t_2 (grün) um das erwartete Gewicht $w(t_2, \overline{\mathbf{d}})$.

Für die Modellvereinfachung ist ein Referenzdokument $\mathbf{d}^*$ gesucht, mit dem im Gegensatz zu $\overline{\mathbf{d}}$ bei Beobachtung von Term t_2 die Wahrscheinlichkeit $p(X)$ im Mittel nicht größer wird. Gleichzeitig soll bei Beobachtung von Term t_1 die Wahrscheinlichkeit $p(X)$ steigen.

Das gesuchte Gewicht $[\mathbf{d}^*]_t$ für einen Term t liegt zwischen dem zu erwarteten Gewicht $w(t, \overline{\mathbf{d}})$ und dem maximalen Gewicht $w_{max}(t, \mathbf{D}) = \max\limits_{i=1,\ldots,N} w(t, \mathbf{d}_i)$ des

Terms t in der Dokumentkollektion $\mathbf{D}$. Dabei stellt das erwartete Gewicht die bessere Annäherung bereit. Für eine Abschätzung des Referenzdokuments $\mathbf{d}^*$ bietet sich daher das harmonische Mittel

$$[\mathbf{d}_{harm}]_t = \frac{2 \cdot w(t, \overline{\mathbf{d}}) \cdot w_{max}(t, \mathbf{D})}{w(t, \overline{\mathbf{d}}) + w_{max}(t, \mathbf{D})}$$

an, dass den kleineren Wert stärker gewichtet. Abbildung 23 zeigt die Höhe der Gewichte $w(t_1, \mathbf{d}^*)$ und $w(t_2, \mathbf{d}^*)$ der Terme $t_1$und t_2. Die Termgewichte von Term t_1 in den Dokumenten 1 bis 4 liegen, wie in Abbildung 22, alle über dem Gewicht $w(t_1, \mathbf{d}^*)$ des Referenzdokuments. Bei der Retrieval-Wert-Berechnung $\rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j)$, mit $\mathbf{d}_i, \mathbf{d}_j \in \mathbf{D}_1$, ensteht durch den Summanden für Term t_1 eine Erhöhung der Wahrscheinlichkeit $p(X)$. Bei Term t_2 überragen nur noch zwei Termgewichte das harmonische Gewicht $w(t_2, \mathbf{d}^*)$. Im Mittel ergibt sich für $\rho_{\mathcal{R}}(\mathbf{d}_m, \mathbf{d}_n)$, mit $\mathbf{d}_m, \mathbf{d}_n \in \mathbf{D}_2$, durch den Summanden für Term t_2 eine Verkleinerung der Wahrscheinlichkeit $p(X)$.

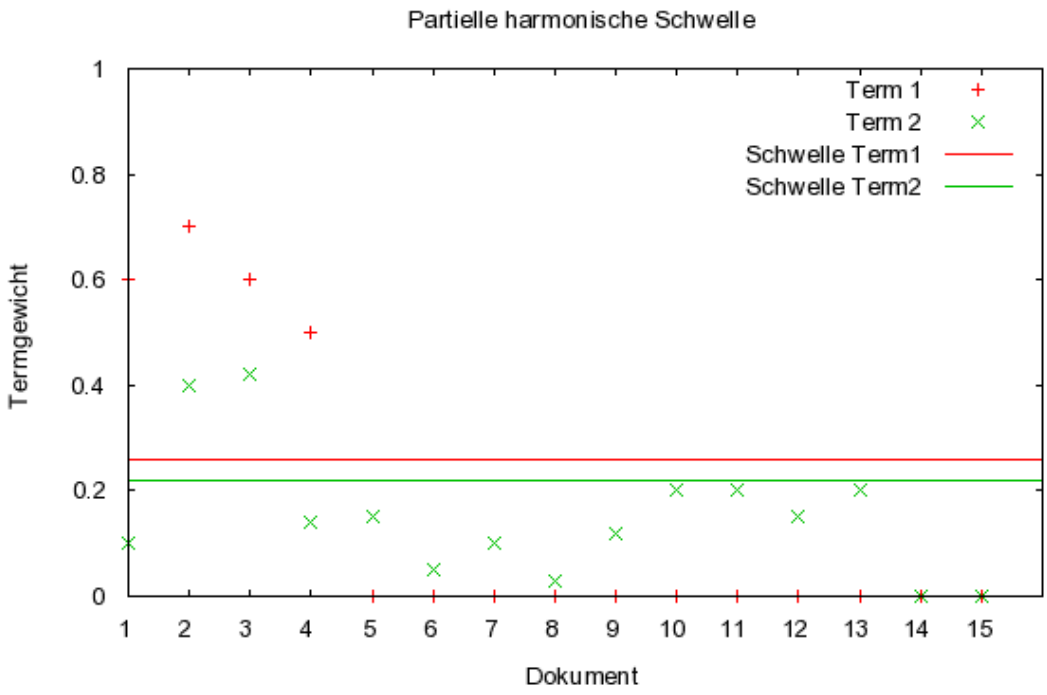


Abbildung 23: Termgewichte des Referenzdokuments $\mathbf{d}^*$. Das Referenzgewicht $w(t_1, \mathbf{d}^*)$ für Term 1 ist niedriger als die Gewichte in den Dokumenten 1 bis 4 (rot). Im Gegensatz dazu schaffen es nur zwei Dokumente, für Term 2 das Referenzgewicht $w(t_2, \mathbf{d}^*)$ zu überschreiten (grün).

Zusammengefasst ergibt sich für die Modellvereinfachung mit dem harmonischen Referenzdokument $\mathbf{d}^*$ die Vorschrift

$$\varphi_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j) = \begin{cases} \rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j) & \text{falls } \rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j) > \max_{k \in \{i,j\}} \overline{\rho_{\mathbf{d}^*}}(\mathbf{d}_k) \\ 0 & \text{sonst.} \end{cases}$$

3.5.3 Local Average

Eine vergleichsweise simple Vorschrift zur Modellvereinfachung ergibt sich aus der Hypothese, dass thematisch verwandte Dokumente durch einen überdurchschnittlich hohen Ähnlichkeitswert in der Ähnlichkeitsmatrix M vertreten sind. Der Mindestwert $\rho_{avg}(\mathbf{d})$, den ein Retrieval-Wert $\rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j)$ für das Dokument $\mathbf{d}_i \in \mathbf{D}$ annehmen muss, ergibt sich aus dem arithmetischen Mittelwert

$$\rho_{avg}(\mathbf{d}_i) = \frac{\sum_{j=1}^{N} \rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j)}{\mathrm{n}_{\mathbf{d}_i}} \ ,$$

mit $\mathrm{n}_{\mathbf{d}_i}$ als Anzahl der Dokumentpaare, die mit $\mathbf{d}_i$ einen Ähnlichkeitswert größer Null teilen.

Alle Retrieval-Werte $\rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j)$, die unter dem Mindestwert $\rho_{avg}(\mathbf{d})$ einer der beteiligten Dokumente $\mathbf{d}_i$ oder $\mathbf{d}_j$ liegen, werden eliminiert,

$$\rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j) = \begin{cases} \rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j) & \text{falls } \rho_{\mathcal{R}}(\mathbf{d}_i, \mathbf{d}_j) > \max_{k=\{i,j\}} \rho_{avg}(\mathbf{d}_k) \\ 0 & \text{sonst.} \end{cases}$$

3.5.4 Relevance Ratio

Der maximale Spannbaum verbindet die Knoten eines gewichteten Graphen so, dass der Graph kreisfrei und die Masse der verwendeten Kanten maximal ist. Zur Generierung des maximalen Spannbaums existiert der Algorithmus von Kruskal [Kruskal56]: Aus der Liste der noch nicht betrachteten Kanten wird jeweils die schwerste Kante zum Graphen hinzugefügt, falls sie den Graphen kreisfrei lässt. Ersetzt man die Forderung nach Kreisfreiheit mit dem Abbruchkriterium „bis der Graph zusammenhängend ist“, ergibt sich der Ausgangspunkt für die Modellvereinfachung „Relevance Ratio“.

In Abbildung 20, zur Verteilung der Masse im Ähnlichkeitsgraphen, war beispielhaft zu erkennen, dass unter den Kanten mit hohem Gewicht ausschließlich solche innerhalb von Clustern sind. Bildet man den Graphen $\mathrm{G}_{MST}(\mathrm{V}, \mathrm{E}_{MST}, w)$ nach eingangs beschriebener Vorschrift mit Abbruchbedingung, werden zunächst diese

cluster-internen Kanten dem Graphen zugefügt. Erreicht das höchste noch verfügbare Kantengewicht die Stufe, an der auch Ähnlichkeitswerte zwischen den Clustern auftauchen, verbinden diese die bis dato gebildeten Cluster nach und nach zu einem zusammenhängenden Graphen und der Algorithmus stoppt. Der so erzeugte Graph G_{MST} entspricht einer Modellvereinfachung mit einem globalen Grenzwert in Höhe des Gewichts der zuletzt eingefügten Kante.

Setzt man die Zahl der in G_{MST} enthaltenen Kanten $|\mathrm{E}_{MST}|$ ins Verhältnis zur Zahl der insgesamt im Graphen $\mathrm{G}(\mathrm{V}, \mathrm{E}, w)$ möglichen Kanten $|\mathrm{E}|$ ($|\mathrm{E}| = \frac{|V| \cdot |V-1|}{2}$), lässt sich eine Aussage über die im Schnitt pro Knoten verwendete Kantenzahl $\overline{k}$ machen,

$$\overline{k} = \frac{|\mathrm{E}_{MST}|}{|\mathrm{E}|} \cdot (|V| - 1) = \frac{2 \cdot |\mathrm{E}_{MST}|}{|V|}$$

Wendet man auf Graph G den Algorithmus des gegenseitigen k-Nearest- Neighbour-Graphen mit dem nächsten ganzzahlige Wert von $\overline{k}$ an, ergibt sich die Modellvereinfachnung nach dem „Relevance Ratio“-Verfahren,

$$\mathrm{G}_{Rel} = kNN(\mathrm{G}, \overline{k}) \ .$$

Interpretiert man einen Ähnlichkeitswert φ von Null in der Ähnlichkeitsmatrix M mit dem Fehlen der entsprechenden Kante im Ähnlichkeitsgraphen G, ist die Anzahl der zu einem Knoten $v \in V$ adjazenten Knoten variabel und die Zahl der in G enthaltenen Kanten $|\mathrm{E}'|$ kollektionsspezifisch (kleiner gleich $\frac{|V| \cdot |V-1|}{2}$). Es lässt sich somit kein allgemeiner Wert $\overline{k}$ angeben. Stattdessen ergibt sich aus dem Kantenverhältnis $|\mathrm{E}_{MST}|/|\mathrm{E}'|$ für jeden Knoten $v_i \in V$ ein eigenes k_{v_i} aus

$$k_{v_i} = \frac{|\mathrm{E}_{MST}|}{|\mathrm{E}'|} \cdot |\mathrm{E}_{v_i}| \ .$$

$|\mathrm{E}_{v_i}|$ gibt dabei an, wieviele Knoten aus G adjazent zu Knoten v_i sind. Zur Erstellung des Graphen G'_{Rel} muss der Algorithmus des gegenseitigen k-Nearest-Neighbour-Graphen so modifiziert sein, dass die variablen Knotenwerte k_v verwendet werden können.

3.5.5 Major Expected Density

Die bisher vorgestellten Ansätze zur Modellvereinfachung modifizieren die Ähnlichkeitsmatrix im Vorfeld der Fusionierung. Das hier, unter dem Namen „Ma-

jor Expected Density“, vorgestellte Verfahren hingegen integriert die Vereinfachung mit in den Fusionierungsprozess. Zentrale Idee dabei ist, bei jeder Cluster-Zuweisung des Fusionierungsalgorithmus zu prüfen, ob durch Eliminierung kleinster Ähnlichkeitswerte ein besseres temporäres Clustering erreicht werden kann. Für die Fusionierung wird der Major-Clust-Algorithmus (3.3.3) verwendet. Die Entscheidung, ob sich durch Modellvereinfachung ein besseres Clustering ergeben hat, wird von einem internen Qualitätsindex, der Expected Density, getroffen. Im Vorgriff auf die detailierte Beschreibung in Kapitel 4.2.3 sei hier nur erwähnt, dass der Index unter gegebenen Vorschlägen jenes Clustering bestimmt, welches die besten strukturellen Eigenschaften besitzt.

Bei der Fusionierung mit Major Clust wird, wie bereits beschrieben, in einem Fusionierungsdurchgang für jeden Knoten $\mathbf{d}_i \in \mathbf{D}$ der Cluster C_x ermittelt, der die stärkste Anziehungskraft f auf ihn auswirkt,

$$x = \arg\max_{j \in \mathcal{C}} \; f(C_j, \mathbf{d}_i)\,.$$

Für die Modellvereinfachung wird nun auch der Cluster C_y gesucht, welcher nach C_x die stärkste Kraft ausübt,

$$y = \arg \max_{j : j \in \mathcal{C}, j \neq x} \; f(C_j, \mathbf{d}_i)\,.$$

Nun wird so lange die jeweils kleinste Kante zwischen den Knoten aus C_x und $\mathbf{d}_i$ entfernt, bis sich für Cluster C_y die stärkere Anziehungskraft f ergibt. Es ergeben sich damit zwei temporäre Cluster-Konfigurationen: Zum einen die gewöhnliche Vereinigung von Dokument $\mathbf{d}_i$ mit Cluster C_x unter Beibehaltung aller Kanten, zum anderen die Vereinigung von $\mathbf{d}_i$ und C_y unter Egalisierung des Kräfteunterschieds. Die beiden Konfigurationen werden dem Qualitätsindex vorgestellt. Jene Konfiguration, die den höheren Qualitätswert erzielt, wird für die weiteren Betrachtungen übernommen, die andere wird verworfen.

Da die Modellvereinfachung erst während der Fusionierung stattfindet, lässt sich der Ansatz mit anderen Verfahren kombinieren. So kann im Vorfeld eine erste, vorsichtige Eliminierung von Ähnlichkeitswerten durchgeführt werden, die dann während der Fusionierung vom Major-Expected-Density-Verfahren verfeinert wird.

4 Qualitätsindizes

Von der Erstellung der Dokumentrepräsentationen bis zur Fusionierung zu homogenen Gruppen[12] steht beim Dokumenten-Clustering in jedem Schritt eine Vielzahl an Methoden zur Auswahl. Ohne einen Hinweis auf die Beschaffenheit der Dokumentkollektion (wie z.B. Anzahl oder Struktur[13] der zu erwartenden Cluster) ist es schwer zu entscheiden, welche Kombination an Verfahren die besten Ergebnisse produziert. Hinzu kommt, dass Retrieval-Modelle wie das Best-Match-Modell (2.5) oder die Gewichtungsvorschrift „Lnu“ (2.2.1 auf Seite 22) parametrisiert sind, Fusionierungsalgorithmen (3.3) benötigen eine Vorgabe der Clusterzahl oder liefern Ergebnisse in Abhängigkeit von der Startkonfiguration. Im Vorfeld der Analyse sind kaum einschränkende Aussagen machbar, die geringe Laufzeit zur Clustergenerierung ermöglicht jedoch einen Vergleich mehrerer erzeugter Ergebnisse.

Ein Qualitätsindex (engl. *validity measure*) versucht unter einer Reihe an Lösungsvorschlägen den Kandidaten zu nominieren, der das beste Clustering verkörpert. Fällt die Entscheidung dabei unter Einbeziehung einer Referenzlösung, spricht man von einem externen Qualitätsmaß, interne Maße beziehen sich ausschließlich auf strukturelle Eigenschaften in den Clusterings. Im folgenden werden, in Anlehnung an die Ausführungen in [MeyerZuEissen07], einige Qualitätsindizes vorgestellt.

4.1 Externe Qualitätsindizes

Externe Qualitätsindizes vergleichen die vorgeschlagene Lösung eines Clusteralgorithmus $\mathcal{C}$ mit einer manuell erstellten Referenzlösung $\mathcal{C}^*$. Sie liefern sichere Erkenntnise über die Qualität eines Clusterings. Stellvertretend für die existierenden externen Qualitätsindizes sei hier der bekannteste und für die im Rahmen der Arbeit durchgeführten Experimente verwendete Qualitätsindex, das F-Measure, vorgestellt.

4.1.1 F-Measure

Für jedes Clusterpaar eines Referenz- und ermittelten Clusterings $(C_i^*, C_j) \in \mathcal{C}^* \times \mathcal{C}$ können zwei Kenngrößen aus der Schnittmenge der in C_i^* und C_j enthaltenen Dokumente berechnet werden, die Aufschluss über die Qualität geben.

[12] Über Termgewichtung, Ähnlichkeitsbestimmung und Modellvereinfachung.
[13] Ausreißerbehaftet? Cluster gleicher Größe? ...

Precision bezeichnet das Verhältnis der in der Schnittmenge $C_i^* \cap C_j$ enthaltenen Elemente zur Anzahl der Dokumente im Cluster $C_j \in \mathcal{C}$,

$$prec\,(C_i^*, C_j) = \frac{|C_i^* \cap C_j|}{|C_j|} \ .$$

Die Größe macht eine Aussage über die „Reinheit“ des Clusters C_j. Sind alle Dokumente des Clusters relevant (also auch in C_i^* enthalten), ergibt sich ein Precision-Wert von 1,0.

Die Größe *Recall* stellt die Zahl der gemeinsamen Dokumente dagegen ins Verhältnis zur Anzahl der Referenzdokumente im Cluster $C_i^* \in \mathcal{C}_i^*$,

$$rec\,(C_i^*, C_j) = \frac{|C_i^* \cap C_j|}{|C_i^*|} \ .$$

Der Wert impliziert, wieviele der relevanten Dokumente Teil des Clusters C_j sind. Auch hier ist der maximale Wert 1,0.

Während es trivial ist, ein Clustering zu entwerfen, das *einen* der beiden Werte maximiert[14], zeichnet sich ein qualitativ hochwertiges Clustering durch einen großen Precision- *und* Recall-Wert aus. Das F-Measure ist formaler Ausdruck dieses Gedankens. Der partielle Wert des F-Measures für ein Clusterpaar $(C_i^*, C_j) \in \mathcal{C}^* \times \mathcal{C}$ errechnet sich aus dem harmonischen Mittel von Precision und Recall,

$$F_{i,j} = \frac{1}{\frac{1}{2} \cdot \left(\frac{1}{prec(C_i^*, C_j)} + \frac{1}{rec(C_i^*, C_j)} \right)} = \frac{2 \cdot prec(C_i^*, C_j) \cdot rec(C_i^*, C_j)}{prec(C_i^*, C_j) + rec(C_i^*, C_j} \ .$$

Nimmt man für jeden Referenzcluster C_i^* den „stärksten“ Partner $F_i^{'} = \max_{j \in \mathcal{C}} F_{i,j}$ aus $\mathcal{C}$, lässt sich der gemittelte Gesamtwert schreiben als

$$F_{\mathcal{C}^*,micro} = \sum_{i \in \mathcal{C}^*} \frac{|C_i^*|}{|D|} \cdot F_i^{'} \ .$$

Die Berechnung des F- Measures erlaubt dabei zwei orthogonale Variationen.

Zum einen können alle Cluster ungeachtet ihrer Größe gleichen Anteil am Endergebnis haben,

$$F_{\mathcal{C}^*,macro} = \sum_{i \in \mathcal{C}^*} \frac{1}{|\mathcal{C}^*|} \cdot F_i^{'} \ ,$$

[14] Mit $prec(C_i^*, C) = \max_{j \in C} prec(C_i^*, C_j)$ ergibt sich eine Precision von 1.0 für ein Clustering, das jedem Dokument einen eigenen Cluster zuweist. Der Recall wird maximal bei einem Cluster mit allen Dokumenten.

zum anderen liefert, bei ungleicher Clusterstärke ($|\mathcal{C}^*| \neq |\mathcal{C}|$) , die Iteration über $\mathcal{C}$ (statt $\mathcal{C}^*$) zur Bestimmung des maximalen $F_{i,j}$ mit $F_j^{''} = \max_{i \in \mathcal{C}^*} F_{i,j}$ andere Endergebnisse. Während $F^{'}$ eine Unterschätzung der richtigen Clusterzahl bevorzugt (Cluster aus $\mathcal{C}$ werden dann einfach mehrfach referenziert), priviligiert $F^{''}$ eine Übertreibung der Cluster-Anzahl (Abbildung 24). Ein probater Kompromiss ist es, über das Clustering mit weniger Clustern zu iterieren. Dadurch ist dafür gesorgt, dass jeder Cluster nur einmal in die Bewertung einfließt (zu sehen oben rechts bzw. unten links in Abbildung 24).

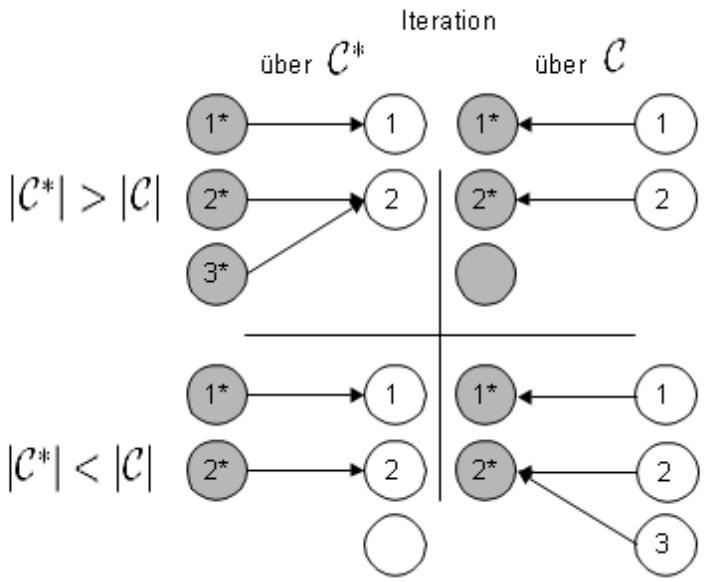


Abbildung 24: Bestimmung des „stärksten“ Partners (*best match*). Die Entscheidung, über welches Clustering iteriert wird, beeinflusst bei ungleicher Zahl der Cluster den Wert des F-Measures.

4.2 Interne Qualitätsindizes

Interne Qualitätsindizes nutzen zur Einschätzung der Qualität ausschließlich Informationen, die sich unmittelbar aus den gegebenen Clusterings gewinnen lassen. Im Gegensatz zu den externen Maßen sind sie damit auch für den Gebrauch im Anwendungsfall geeignet. Interne Indizes prüfen, inwieweit das vorgelegte Clustering eine Homogenisierung der Ausgangsmenge darstellt. Meist wird angenommen, dass Aussagen über die *Kompaktheit* eines Clusters oder die *Distanz* zwischen den Clustern[15] positiv mit der Homogenität korrelliert sind. Die Beurteilung

[15] der Maxime der Clusteranalyse entsprechend.

der Clusterqualität kann sich aus Komplexitätsgründen naturgemäß[16] nicht auf den Maximalwert der zugrunde liegenden Gütekriterien beziehen, weshalb Qualitätsangaben stets nur relativ zu den konkurrierenden Clusterings möglich sind.

4.2.1 Dunn-Index

Der Index von Dunn stellt die Eigenschaften Kompaktheit und Separiertheit ins Verhältnis. Ein Clustering $\mathcal{C}$ erzielt dann einen hohen, und damit guten, Indexwert, wenn die Distanzen (Unähnlichkeiten) $\delta(C_i, C_j)$ zwischen den Clustern aus $\mathcal{C}$ groß sind, während die maximale Streuung in den Clustern $\triangle(C_l)$ klein ist,

$$I(C) = \frac{\min\limits_{i \neq j} \delta(C_i, C_j)}{\max\limits_{l \in C} \triangle(C_l)} \ .$$

Ursprünglich verwendete Dunn als Distanzfunktion δ den minimalen Abstand zwischen zwei Objekten aus unterschiedlichen Clustern und als Funktion für die Streuung $\triangle$ den maximalen Abstand innerhalb eines Clusters [Dunn74]. Ein robusteres Maß wurde 1995 von J. Bezdek mit

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} \varphi(x, y)$$

und

$$\triangle(C_l) = 2 \left(\frac{\sum_{x \in C_i} \varphi(x, c_i)}{|C_i|} \right)$$

vorgestellt [Bezdek95]. Die Funktion $\varphi(x, y)$ bestimmt die Distanz zwischen zwei Objekten, c_i und c_j bezeichnen die Centroiden des Clusters C_i bzw. C_j.

4.2.2 Davis-Boldin-Index

Für Davis und Boldin [Davis79] ergibt sich die Qualität eines Clusterings $\mathcal{C}$ aus der Güte $R_i(\mathcal{C})$ der einzelnen Cluster $C_i \in \mathcal{C}$,

$$DB(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \cdot \sum_{i \in \mathcal{C}} R_i(\mathcal{C}) \ .$$

[16] Ansonsten wäre der Qualitätsindex ein Fusionierungsalgorithmus.

Die Güte $R_i(\mathcal{C})$ des Clusters C_i ist wiederum Ausdruck von Streuung $\triangle$ und Separiertheit δ,

$$R_i(\mathcal{C}) = \max_{i,j \in \mathcal{C}, i \neq j} R_{ij}(\mathcal{C}) \text{ mit}$$

$$R_{ij}(\mathcal{C}) = \frac{(\triangle(C_i) + \triangle(C_j))}{\delta(C_i, C_j)} \ .$$

Ein typisches Maß für die Streuung ist $\triangle(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} |x_i - c_i|$, Separiertheit wird als $\delta(C_i, C_j) = |c_i - c_j|$ mit c_i als Centroid des Clusters C_i quantifiziert [MeyerZuEissen07]. Ein wohlgeformtes Clustering mit geringer Streuung und großer Interclusterdistanz minimiert den Davis-Bouldin-Index.

4.2.3 Expected Density

Das Maß der Expexcted Density kommt auf die Interpretation der Ähnlichkeitsmatrix als gewichteten Graphen G(V,E,w) zurück. Dort lässt sich über die Summe aller Kantengewichte[17] $w(G) = |V| + \sum_{e \in \mathrm{E}} w(e)$ eine Aussage über die Dichte θ des Graphen G formulieren,

$$w(\mathrm{G}) = |\mathrm{V}|^{\theta} \ \Leftrightarrow \ \theta = \frac{\ln(w(\mathrm{G}))}{\ln|\mathrm{V}|} \ .$$

Die Dichte θ stellt die Bezugsgröße für die Bewertung eines Clusterings $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ dar. Es seien $\mathrm{G}_{\mathrm{i}=1,\ldots,\mathrm{k}}(\mathrm{V_i}, \mathrm{E_i}, w_i) \subset \mathrm{G}$ die k Teilgraphen, die durch $C_1, C_2, \ldots, C_k \subset V$ induziert werden. Ein Teilgraph G_i wird dann für dichter als der Gesamtgraph erachtet, wenn der Quotient $\frac{w(\mathrm{G}_i)}{|\mathrm{V}_i|^{\theta}}$ größer als 1,0 ist. Das dazu erforderliche mittlere Kantengewicht im Teilgraphen G_i fällt in Abhängigkeit von θ mit der Größe $|\mathrm{V}_i|$ des Teilgraphen.

Ein Clustering $\mathcal{C}$ sollte den Graphen G so in Teilgraphen $\mathrm{G}_{i=1,\ldots,k}$ zerlegen, dass diese zusammengenommen eine möglichst hohe innere Dichte besitzen. Die Expected Density eines Clusterings $\mathcal{C}$ summiert die Dichtequotienten der Teilgraphen daher im Verhältnis ihrer Größe zu

$$\overline{\rho}(\mathcal{C}) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^{\theta}} \ .$$

[17] $|V|$ dient dabei als Anpassungwert für Graphen mit kleinen Kantengewichten [MeyerZuEissen07].

5 Experimente

Externe Qualitätsindizes beurteilen die Güte eines Clusterings objektiv. Mit ihrer Verwendung lässt sich die Leistungsfähigkeit der am Dokumenten-Clustering beteiligten Komponenten experimentell untersuchen. Dafür notwendig sind Dokumentsammlungen, für die sich ein Referenz-Clustering zur Repräsentation der korrekten Lösung erzeugen lässt. Für die im Rahmen dieser Arbeit durchgeführten Experimente wurden zwei englischsprachige Nachrichtensammlungen verwendet, deren Artikel mit Metainformationen zur Einordnung in Sachgebiete ausgestattet sind:

Der „Reuters Corpus Volume 1“ (RCV1) ist eine Sammlung von über 800000 englischsprachigen Meldungen, die von der Nachrichtenagentur Reuters[18] für wissenschaftliche Untersuchungen freigegeben sind [Rose02]. Jeder Artikel ist in einer Hierarchie von Themenfeldern (*sog. topics*) eingeordnet. Ausgehend von den vier Top-Level-Kategorien „Corporate/Industrial“, „Economics“, „Government/Social“ und „Markets“ verzweigt sich die Hierarchie in zunehmend speziellere Unterkategorien (Abbildung 25). Ein Artikel gehört jeder Kategorie an, die auf dem Pfad von der speziellsten zugeordneten Kategorie bis zur Top-Level-Kategorie liegt.

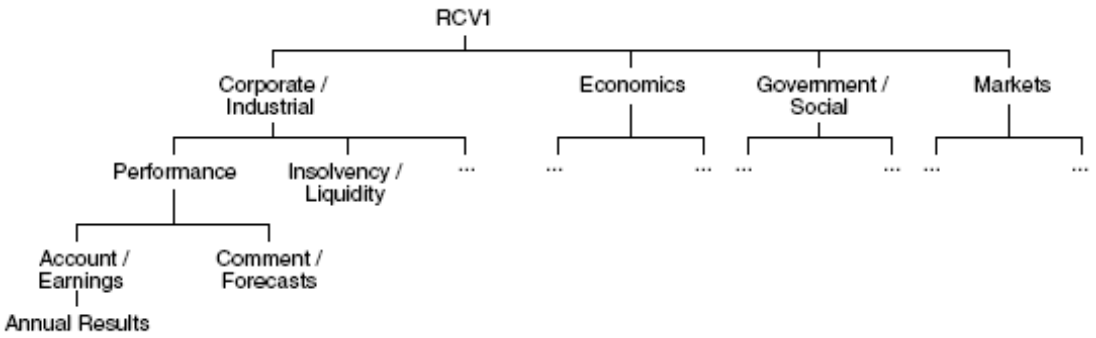


Abbildung 25: Teil der Themenhierarchie in der Reuters-Sammlung RCV1 (Quelle: [MeyerZuEissen07])

Die zweite Dokumentsammlung ist eine von der Text Retrieval Conference (TREC)[19] verwendete Zusammenstellung an Nachrichtenartikeln, die in der LA Times[20] zwischen 1989 und 1990 veröffentlicht wurden. In jedem Artikel enthalten ist die Metainformation, in welcher Sparte (engl. section) der Zeitung der Artikel abgedruckt wurde. Analog zum Vorgehen bei [Steinbach00] wird dieser Spartenname

[18] `www.reuters.com`.
[19] `http://trec.nist.gov/overview.html`.
[20] `www.latimes.com`.

(z.B. „Sports“, „Business“ oder „Metro“ (dt. Regionales)) zur Gruppierung der Artikel in verschiedene Themengebiete verwendet.

Für die Experimente wurden aus den beiden Dokumentsammlungen jeweils sieben Testkollektionen generiert. Eine *Testkollektion* bezeichnet eine Teilmenge der in einer Sammlung enthaltenen Dokumente. Sie ist definiert über die Anzahl der Themengebiete, aus denen die einzelnen Dokumente stammen, der Zahl insgesamt enthaltener Dokumente sowie deren Verteilung über die Themenbereiche. Die Dokumente eines Themenbereichs bilden jeweils einen Cluster im Referenz-Clustering der Testkollektion.

Tabelle 1 zeigt eine Übersicht der mit der Reuters-Sammlung RCV1 erstellten Testkollektionen.

Reuters-Kollektionen	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
# Kategorien	3	3	3	5	5	5	6
# Dokumente	300	1500	1500	500	1500	2500	600
gleichverteilt	ja	ja	nein	ja	ja	nein	ja

Tabelle 1: Reuters Testkollektionen. Der Name der Kollektion leitet sich aus ihrer Zusammensetzung ab. Das 'R' steht für die Reuters-Sammlung, 'Cx' bezeichnet die Anzahl der Kategorien, 'E' bzw. 'U' definiert die Art der Verteilung. Die darauffolgende Zahl verkörpert die Mächtigkeit der Dokumentmenge.

Bei der Auswahl der Kategorien wurde darauf geachtet, dass keine Kategorie auf dem Pfad zur Top-Level-Kategorie einer anderen Kategorie liegt. Die Dokumente einer Testkollektion wurden zufällig unter allen Dokumenten einer Kategorie gezogen. Um repräsentative Ergebnisse wahrscheinlicher zu machen gab es pro Testkollektion fünf Ziehungen, die in den Experimenten verwendet werden konnten.

Tabelle 2 zeigt die Zusammensetzung der Testkollektionen aus der LA Times-Sammlung. Auch hier wurden die Dokumente der Kollektion in fünf Ziehungen zufällig ausgewählt.

LATimes-Kollektionen	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
# Cluster	3	3	3	5	5	5	6
# Dokumente	300	1500	1500	500	1500	2500	600
gleichverteilt	ja	ja	nein	ja	ja	nein	ja

Tabelle 2: LA Times Testkollektionen. Ein vorangestelltes 'L' im Namen der Kollektion zeigt an, dass die Dokumente der LA Times-Sammlung entstammen.

Jedes der folgenden Experimente wurde mit den 14 beschriebenen Testkollektionen durchgeführt. Bei der Indexierung wurden stets Layout- und Navigations-Elemente vom Text getrennt, Stoppworte mit Standardlisten entfernt und der Porter-Stemming-Algorithmus [Porter80] angewendet.

5.1 Allgemeines Korrelationsexperiment

Unter einer Reihe erzeugter Clusterings ohne Kenntnis der Ideallösung das beste Clustering nominieren zu können, ist eine wichtige Fähigkeit für das Dokumenten-Clustering. So können Parameter bestimmt werden, für die im Vorfeld der Analyse keine fundierte Abschätzung gemacht werden konnte. Die in Kapitel 4.2 vorgestellten internen Qualitätsindizes ermöglichen eine solche Nominierung. Im allgemeinen Korrelationsexperiment wurde untersucht, wie gut die Ergebnisse interner Indizes den tatsächlichen Sachverhalt aufdecken können.

Zur Vorbereitung des Experiments wurden die Dokumente der Testkollektionen mit der lnc-Gewichtung (Formel 3) des Vektorraummodells gewichtet und die Ähnlichkeitswerte zwischen den Dokumenten mit dem Skalarprodukt bestimmt. Die in Abschnitt 3.3 beschriebenen Fusionierungsprinzipien Group-Average Link, k-Means und Major Clust erzeugten anschließend durch Variation der Parameter (Anzahl der Cluster, Zahl der Ähnlichkeitswerte pro Dokument) für jede Testkollektion 29 Clusterings. Die Qualität der Clusterings wurde mit dem F-Measure objektiv quantifiziert. Mit dem Referenz-Clustering (F-Measure von 1,0) ergaben sich insgesamt 30 Clusterings pro Testkollektion.

Die Aufgabe der internen Qualitätsindizes bestand nun darin, für die 30 Clusterings ebenfalls einen Qualitätswert zu berechnen. Sollten die Einschätzungen der internen Indizes stimmen, korrespondiert ein verhältnismäßig großer (bei Davis-Boldin kleiner) Qualitätswert für ein Clustering mit einem hohem Wert für das F-Measure, ein kleiner (bei D.B. großer) Wert mit einem entsprechend niedrigen F-Measure-Wert. Insgesamt stellt sich ein monotoner Zusammenhang zwischen den internen Qualitätsindizes und dem F-Measure ein.

Abbildung 26 zeigt das Ergebnis der Messungen für die Testkollektion „LC3E1500“ in einem Streudiagramm. Auf der y-Achse sind die Werte des F-Measures abgetragen, auf der x-Achse die Werte der Qualitätsindizes. Um die Vergleichbarkeit der Diagramme zu verbessern, wurde das Vorzeichen des Davis-Boldin-Index, den es als einzigen zu minimieren gilt, geändert. Ein kleinerer Wert deutet somit bei allen Indizes auf ein schlechteres Clustering hin. Ein monotoner Zusammehang

wäre am Verlauf der Wertepaare von unten links nach oben rechts im Diagram zu erkennen.

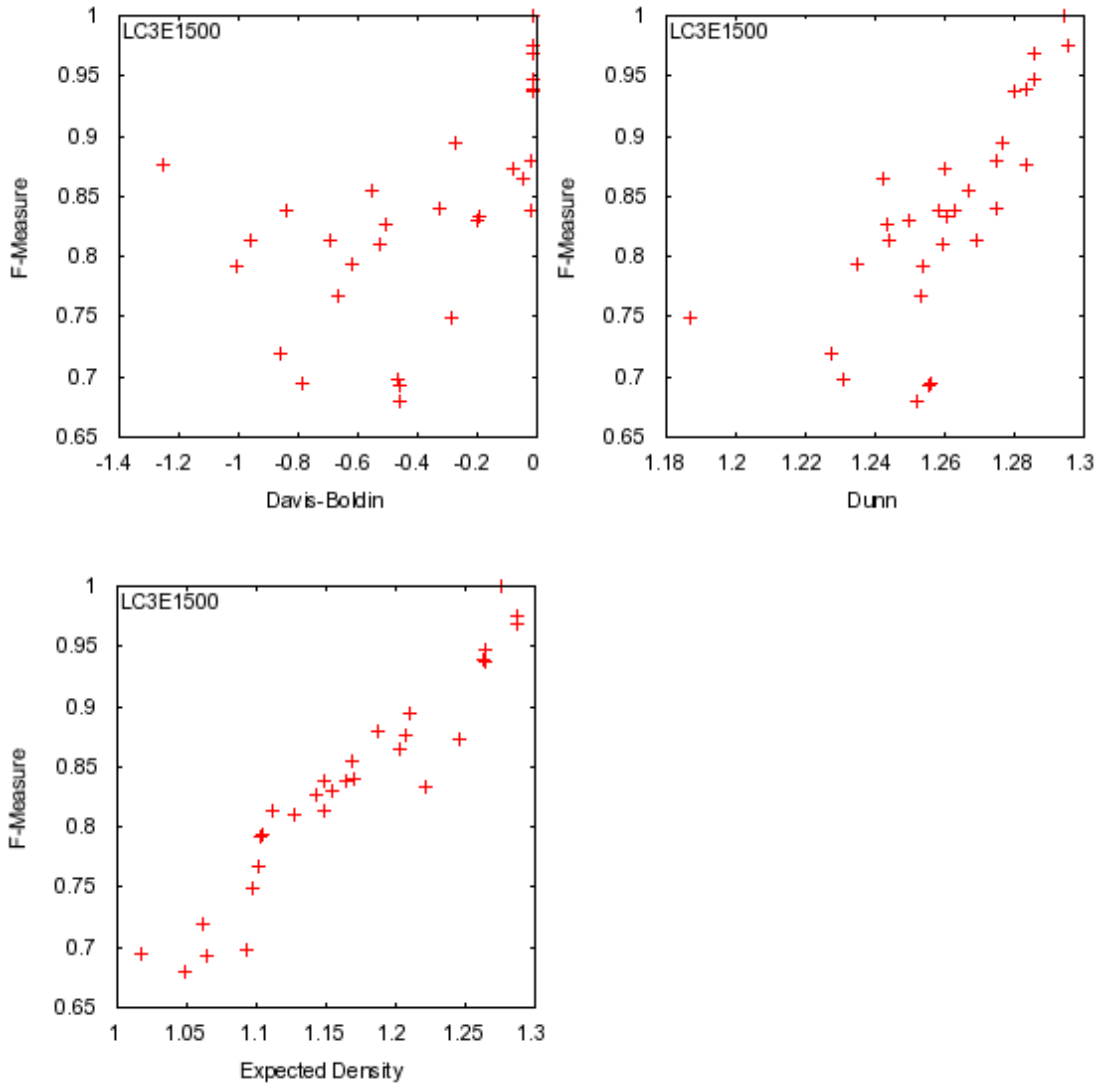


Abbildung 26: Korrelation zwischen F-Measure und Davis-Boldin-Index (oben links), Dunn-Index (oben rechts) und Expected Density (unten links).Da ein kleinerer Wert für den Davis-Boldin-Index ein besseres Clustering beschreibt, wurde zur besseren Vergleichbarkeit das Vorzeichen der Indexwerte geändert.

In den Streudiagrammen aus Abbildung 26 zeigt sich vor allem für die Expected Density der erwartete monotone Verlauf. Bei den anderen beiden Indizes ist die Tendenz schwächer, aber noch offensichtlich.

Um eine statistische Aussage über die Signifikanz der Korrelation zu erhalten, wurde der Spearman-Korrelationskoeffizient [Spearman04] für die internen Indizes berechnet. Für 30 Wertepaare ergibt sich ein Signifikanzniveau von $\alpha = 0,01$ ab

Spearmans Rho	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
Davis-Boldin	0.57*	0.68*	0.41	0.34	0.3	0.03	0.49*
Dunn	0.44	0.81*	0.61*	0.41	0.45	0.45	0.41
Exp. Density	0.94*	0.97*	0.85*	0.47*	0.59*	0.8*	0.66*
Spearmans Rho	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
Davis-Boldin	0.79*	0.71*	0.88*	0.07	-0.18	-0.07	0.18
Dunn	-0.03	-0.33	-0.07	-0.1	-0.09	-0.36	0.19
Exp. Density	0.6*	0.91*	0.88*	0.86*	0.82*	0.94*	0.95*

Tabelle 3: Spearmans Korrelationskoeffizent der drei untersuchten Indizes. Die mit Asterisk gekennzeichneten Werte erreichen das Signifikantniveau von 0,01.

einem Korrelationskoeffizienten von 0,47 [Hole05]. Tabelle 3 zeigt die Koeffizienten für die einzelnen Testkollektionen.

Aus Tabelle 3 ist abzulesen, dass die Expected Density in allen Testkollektionen das Signifikanzniveau von 0.01 erreicht. Der Dunn-Index erreicht für die Kollektionen der LA Times-Sammlung eine signifikante Korrelation, für die Reuters-Sammlung ergibt sich kein monotoner Zusammenhang. Der Index von Davis und Boldin liefert für die Testkollektionen mit drei Kategorien gute Resultate. In den übrigen Kollektionen wird, mit einer Ausnahme, keine signifikante Korrelation mehr erreicht.

5.2 Spezielles Korrelationsexperiment

Im zweiten Korrelationsexperiment wurde die Expected Density im Zusammenhang mit der Modellvereinfachung durch den kNN-Graphen (Abschnitt 3.5.1) untersucht. Während das zu bewertende Clustering immer gleich blieb, variierte der mit dem gegenseitgen kNN-Graphen vereinfachte Ähnlichkeitsgraph G.

Die Erstellung des Graphen G erfolgte für jede Testkollektion nach dem Vektorraummodell mit lnc-Gewichtung. Durch Variation des Parameters k (maximale Anzahl der Kanten pro Knoten) wurden 15 Modellvereinfachungen von G mit dem gegenseitigen kNN-Graphen erzeugt. Um zu prüfen, wie gut die Vereinfachungen das Fusionierungsergebnis verbessern, wurde mit Major Clust jeweils ein Clustering erzeugt und das F-Measure errechnet.

Zu den 15 vereinfachten Graphen wurde nun unter Verwendung des Referenz-Clusterings die Expected Density errechnet und den Ergebnissen des F-Measures gegenübergestellt. Analog zum allgemeinen Korrelationsexperiment sollte sich ein monotoner Zusammenhang zwischen den Wertepaaren ergeben, wenn die Expected Density das Potential der vereinfachten Graphen richtig einschätzt. Abbil-

dung 27 zeigt die Streudiagramme für die Testkollektionen RC6E600, LC6E600, RC5E500 und LC5E500.

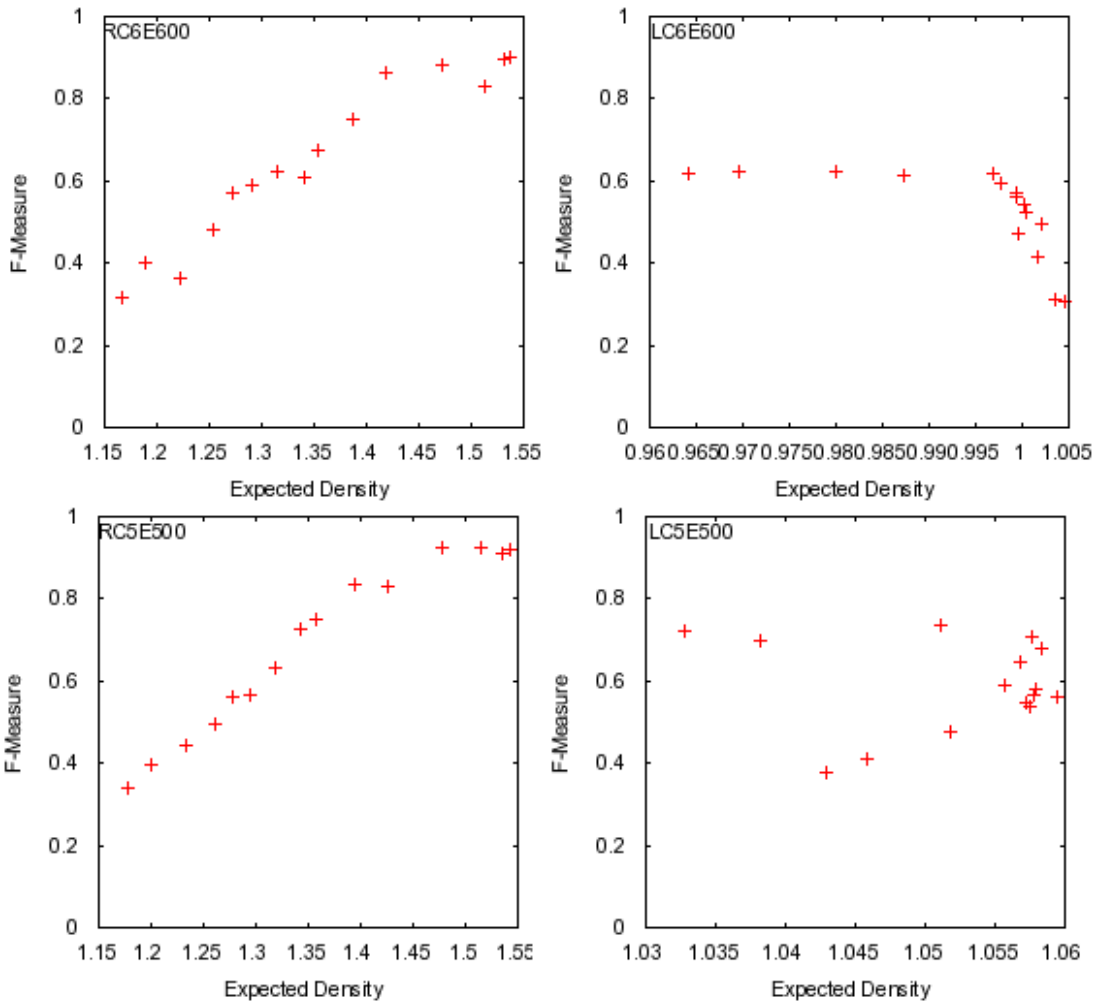


Abbildung 27: Streudiagramm für das spezielle Korrelationsexperiment. Auf der linken Seite der Abbildung befinden sich die Kollektionen RC6E600 (oben) und RC5E500 (unten) aus der Reuters-Sammlung, auf der rechten Seite sind die entsprechenden Kollektionen aus der LA Times-Sammlung zu sehen.

Die Streudiagramme zeigen zwei sehr unterschiedliche Tendenzen. Die Korrelation für die Reuters-Kollektionen auf der linken Seite von Abbildung 27 ist nahzu perfekt. Ein größerer Wert für die Expected Density führt zu einem höherem F-Measure-Wert. Für die nicht gezeigten Verläufe der restlichen Testkollektionen zeichnet sich das gleiche Bild ab. Die Diagramme der Testkollektionen LC5E500 und LC6E600 stellen somit Ausnahmen der allgemein gemachten Beobachtung dar. Bei genauerem Blick auf die beiden Streudiagramme fällt auf, dass die Werte

der Expected Density auf einen kleinen Wertebereich nahe 1 abbilden. Nach Definition der Expected Density geht dies einher mit der Aussage, dass sich die Dichte in den Clustern des Referenz-Clusterings bei den gewählten Modellvereinfachungen nicht von der Dichte im Gesamtgraphen abhebt. Für die mit Major Clust erzeugten Clusterings kann bemerkt werden, dass der in den Testkollektionen LC5E500 und LC6E600 maximal erreichte F-Measure-Wert hinter den Maxima der übrigen Testkollektionen zurückbleibt.

Zur Quantifizierung des Korrelationsgrads wurde wieder der Spearman-Korrelationskoeffizient berechnet. Ein Signifikanzniveau von $\alpha = 0,01$ wird für 15 Wertepaare bei 0.54 erreicht [Hole05]. Tabelle 4 zeigt die Werte des Korrelationskoeffizienten für die einzelnen Testkollektionen.

	LC3E300	LC3E1500	LCU1500	LC5E500	LC5E1500	LC5U2500	LC6E600
Exp. Density	0.93*	0.83*	0.98*	0.2	0.74*	0.69*	-0.85
	RC3E300	RC3E1500	RCU1500	RC5E500	RC5E1500	RC5U2500	RC6E600
Exp. Density	0.89*	0.95*	0.85*	0.94*	0.99*	0.96*	0.94*

Tabelle 4: Spearmans Korrelationskoeffizient für das spezielle Korrelationsexperiment. Zwölf der 14 Kollektionen übertreffen das Signifikanzniveau von 0,01 deutlich (angedeutet mit einem Asterisk).

Für zwölf der 14 Testkollektionen zeigt sich ein hochsignifikanter Zusammenhang zwischen den Werten der Expected Density und den F-Measure-Werten der mit Major Clust erzeugten Clusterings.

5.3 Vergleich der Retrieval-Modelle

Die in Kapitel 2 vorgestellten Retrieval-Modelle wurden ursprünglich entwickelt, um die Dokumente $\mathbf{d} \in \mathbf{D}$ bezüglich der Relevanz zu einer Anfrage $\mathbf{q}$ in eine Rangordnung zu bringen. Welche Effizienz die einzelnen Modelle dabei erreichen, ist Fragestellung des *ad hoc-Runs* im Rahmen der jährlich stattfindenden „Text Retrieval Conference“[21] und daher gut dokumentiert. Für den Einsatz der Modelle zur Ähnlichkeitsbestimmung im Dokumenten-Clustering hingegen gibt es bislang keine dem Autor bekannten vergleichbaren Studien. Die dort zu Grunde liegende Aufgabe unterscheidet sich zunächst insofern vom Ranking-Problem, als die Dokumente aus D in der Regel um ein Vielfaches länger sind als eine Anfrage. Hinzu kommt, dass im Fusionierungsschritt der Cluster-Analyse das Verhältnis in dem die Ähnlichkeitswerte zueinander stehen von Bedeutung ist. So geht es

[21] `http://trc.nist.gov/pubs.html` (Stichwort: ad hoc-Run).

nicht primär darum, die Ähnlichkeitswerte in einen monotonen Zusammenhang mit dem wahren thematischen Verwandtschaftsgrad zu bringen. Vielmehr wird das Retrieval-Modell erfolgreich sein, das zusammengehörige Dokumentmengen $C \subset D$ durch eine homogene Ähnlichkeitsverteilung innerhalb der Teilmenge auszeichnet. Im Experiment wurde untersucht, wie erfolgreich sich Retrieval-Modelle zur Ähnlichkeitsbestimmung im Dokumenten-Clustering einsetzen lassen.

Folgende Retrieval-Modelle wurden dabei betrachtet:

- das Vektrorraummodell (2.2)
- das Modell „Divergence From Randomness“ (2.4)
- das Best-Match-Modell (2.5)
- das statistische Sprachmodell von Ponte und Croft (2.6)
- das explizit semantische Indexierungsmodell (2.7)

Die Retrieval-Modelle wurden zum Teil mit verschiedenen Gewichtungsvorschriften verwendet. Um die Beschreibung zu vereinfachen, wird im folgenden die Gewichtungsfunktion zur Benennung der Verfahren verwendet. So bezeichnet beispielsweise der Audruck „lnc-Modell“ das Vektorraummodell mit lnc-Gewichtung. Insgesamt standen so neun Verfahren zur Verfügung:

Das Vektorraummodell wurde mit den Gewichtungsfunktionen lnc (Formel 3), TfIdf (Formel 6), Lnu (Formel 4) und Ltu (Formel 7) verwendet.

Für das „Divergence from Randomness“-Modell wurden zwei Gewichtungsvorschriften gewählt, die bei den Experimenten in [Amati02] überzeugten: zum einen die Kombination aus Poisson-Modell (Formel 10) zur Berechnung des Informationsgehalts, Laplace-Normalisierung (Formel 13) zur Abschätzung des Risikos und der zweiten Hypothese (Formel 15) zur Normalisierung der Dokumentlänge (kurz: PLH2); zum anderen die Kombination aus dem TfIdf-Modell nach Formel 12, der Abschätzung des Risikos über das Verhältnis zweier Bernoulli-Experimente (Formel 14) und der zweiten Hypothese zur Normalisierung der Dokumentlänge (kurz: InBH2).

Für das Best-Match-Modell wurde die Formel BM25 analog zu [Jin01] mit den Parametern $k_1 = k_3 = 0.5$, b=0.75 und $k_2 = 0$ eingesetzt.

Bei dem statistischen Sprachmodell von Ponte und Croft (kurz: LM) wurde auf die Fallunterscheidung in Formel 21 verzichtet. Stattdessen wurden die Ähnlichkeitswerte über das Skalarprodukt nach Formel 22 bestimmt.

Für das explizit semantische Indexierungsmodell wurde auf Grund der hohen Laufzeit die Anzahl der Wikipedia-Artikel auf 10000 zufällig gezogene Artikel beschränkt.

Abbildung 28 zeigt schematisch den Ablauf des Experiments.

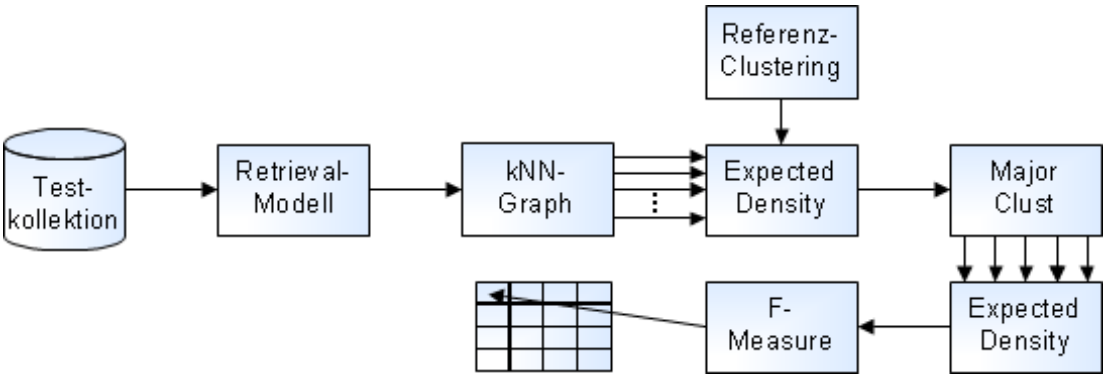


Abbildung 28: Schematischer Ablauf des Experiments. Dargestellt für eine Testkollektion und ein Retrieval-Modell.

Die Retrieval-Modelle bestimmten die Ähnlichkeitswerte zwischen den Dokumenten der 14 Testkollektionen. Die daraus resultierenden Ähnlichkeitsgraphen wurden mit dem gegenseitgen kNN-Graphen vereinfacht. Um die Neutralität der Vereinfachung zu gewährleisten, wurde der Parameter k, die maximale Anzahl an Ähnlichkeitswerten pro Dokument, für jedes Modell individuell bestimmt. Aus 100 Werten für k aus dem Intervall $[\frac{\overline{n}}{10}, \overline{n} + \frac{\overline{n}}{10}]$ (mit $\overline{n}$ als mittlere Anzahl an Dokumenten in den Kategorien einer Testkollektion) wurde die Modellvereinfachung gewählt, die den höchsten Wert bei der Messung der Expected Density mit dem Refernz-Clustering erhielt. Die Ergebnisse des speziellen Korrelationsexperiments zeigen, dass dieses Vorgehen für zwölf der 14 Testkollektionen signifikant mit der Generierung des besten Clusterings für die Menge der vereinfachten Graphen zusammenhängt. Für die beiden Testkollektionen LC5E500 und LC6E600 ist die Objektivität der Ergebnisse nicht gewährleistet. In den Tabellen 5 und 6 sind die ermittelten Werte für den Parameter k eingetragen. Der Wert spiegelt jeweils den Mittelwert der aus fünf unabhängigen Dokumentziehungen gewonnenen Ergebnisse wider.

k	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600	Mittel
lnc	60.4	276.8	304	55.8	155.2	358.8	62.4	181.9
LM	59.2	270.4	304	53.4	152	353.6	60.8	179.1
Lnu	47.6	211.2	238.4	43.2	129.6	275.6	54.4	142.9
TfIdf	61.6	276.8	267.2	57.6	161.6	332.8	73.6	175.9
Ltu	45.6	187.2	188.8	45	128	239.2	52.8	126.7
InBH2	60.8	284.8	300.8	62.4	171.2	377	75.2	190.3
PLH2	61.2	251.2	203.2	60.6	144	257.4	64.8	148.9
BM25	50	214.4	200	43.8	128	244.4	54.4	133.6
ESA	75.2	355.2	414.4	76.8	225.6	522.6	88.8	251.2
Mittel	58.0	258.7	269.0	55.4	155.0	329.0	65,2	170

Tabelle 5: Über die fünf Ziehungen gemittelter Parameter k in den Testkollektionen der LA Times-Sammlung. Der Parameter gibt an, wieviele Ähnlichkeitswerte pro Dokument nach der Vereinfachung mit dem kNN-Graphen maximal noch vorhanden sind.

k	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600	Mittel
lnc	77.6	384	256	79.8	227.2	470.6	80.8	225.1
LM	80	387.2	254.4	82.2	230.4	488.8	82.4	229.2
Lnu	74	348.8	238.4	77.4	222.4	291.2	80.8	190.4
TfIdf	75.6	379.2	241.6	79.8	228.8	304.2	81.6	198.7
Ltu	69.2	297.6	219.2	70.8	206.4	215.8	73.6	164.7
INBH2	79.2	379.2	252.8	86.4	254.4	257.4	88	200.0
PLH2	78.8	358.4	252.8	82.8	228.8	413.4	84	214.1
BM25	63.2	283.2	224	69	196.8	215.8	72.8	160.7
ESA	79.2	355.2	313.6	76.8	233.6	231.4	79.2	195.6
Mittel	75.2	352.5	250.3	78.3	225.4	321.0	80.4	197.6

Tabelle 6: Über die fünf Ziehungen gemittelter Parameter k in den Testkollektionen der Reuters-Sammlung.

Für die Tabellen lassen sich einige Beobachtungen machen.

Zunächst fällt auf, dass für die Testkollektionen der Reuters-Sammlung im Mittel mehr Ähnlichkeitswerte (197,6) verwendet werden als für die Testkollektionen der LA Times-Sammlung (170). Lediglich bei dem ESA-Modell bleiben mehr Ähnlichkeitswerte für die LA Times-Sammlung erhalten. Das ESA-Modell verwendet insgesamt die meisten Ähnlichkeitswerte. Danach folgen das statistische Sprachmodell von Ponte und Croft (LM) und das lnc-Modell. Die Modelle BM25, Ltu und Lnu verwenden vergleichsweise kleine Werte für k.

Der Wert von k hängt weniger von der Mächtigkeit einer Testkollektion ab als von der Zahl $\overline{n}$ der im Schnitt pro Kategorie enthaltenden Dokumente. So ergeben

sich die gemittelten Werte für k jeder Kollektion durch Multiplikation von $\overline{n}$ mit einem Faktor s zwischen 0,5 und 0,8. In Tabelle 7 ist für jedes Modell die mittlere Zahl der verwendeten Ähnlichkeitswerte pro Dokument $\overline{k}$ und das Intervall, in dem der Faktor s liegt, angegeben.

Modell	$\overline{k}$	s	Modell	$\overline{k}$	s	Modell	$\overline{k}$	s
ESA	223.4	[0.46, 1.05]	INBH2	195.2	[0.51, 0.88]	Lnu	166.7	[0.42, 0.8]
LM	204.2	[0.5, 0.98]	TfIdf	187.3	[0.48, 0.82]	BM25	147.2	[0.4, 0.73]
lnc	203.5	[0.51, 0.94]	PLH2	181.5	[0.4, 0.84]	Ltu	145.7	[0.37, 0.74]

Tabelle 7: Übersicht über die im Mittel verwendete Zahl an Ähnlichkeitswerten pro Dokument $\overline{k}$ und dem Intervall, in dem der Faktor s zur Erfüllung der Gleichung $s \cdot \overline{n} = k$ für jede Testkollektion liegt.

Im weiteren Verlauf des Experiments wurde für den besten kNN-Graphen jedes Modells mit Major Clust ein Clustering erzeugt. Um zu vermeiden, dass durch eine ungünstige Reihenfolge der Knotenbetrachtung kein repräsentatives Clustering entsteht, wurde die Fusionierung sieben Mal wiederholt und mittels Expected Density entschieden, welches Clustering zur Bewertung herangezogen wird. Im letzen Schritt des Experiments wurde die Qualität des erzeugten Clusterings mit dem F-Measure quantifiziert. In den Tabellen 8 und 9 sind die F-Measure-Werte aller Modelle eingetragen.

k-Max	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600	Mittel
lnc	0.95	0.96	0.97	0.61	0.56	0.69	0.59	0.76
LM	0.95	0.97	0.97	0.67	0.66	0.69	0.59	0.79
Lnu	0.84	0.8	0.91	0.58	0.6	0.7	0.56	0.71
TfIdf	0.97	0.98	0.98	0.72	0.5	0.68	0.48	0.76
Ltu	0.87	0.94	0.95	0.58	0.66	0.74	0.59	0.76
INBH2	0.94	0.96	0.97	0.57	0.71	0.74	0.61	0.79
PLH2	0.93	0.95	0.9	0.2	0.15	0.35	0.25	0.53
BM25	0.95	0.97	0.97	0.68	0.66	0.69	0.62	0.79
ESA	0.93	0.92	0.93	0.69	0.69	0.59	0.54	0.76
Mittel	0.93	0.94	0.95	0.59	0.58	0.65	0.54	0.74

Tabelle 8: Ergebnisse für die Testkollektionen der LA Times-Sammlung. Die Tabelle enthält den über die fünf Ziehungen der Dokumente gemittelten F-Measure-Wert für den mit Major Clust fusionierten vereinfachten Ähnlichkeitsgraphen der einzelnen Modelle.

k-Max	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600	Mittel
lnc	0.98	0.98	0.99	0.95	0.9	0.8	0.89	0.93
LM	0.95	0.98	0.99	0.96	0.92	0.74	0.92	0.92
Lnu	0.98	0.99	0.99	0.98	0.85	0.93	0.86	0.94
TfIdf	0.97	0.98	0.97	0.93	0.6	0.74	0.75	0.85
Ltu	0.98	0.99	0.99	0.95	0.80	0.72	0.85	0.9
INBH2	0.95	0.98	0.96	0.95	0.74	0.61	0.78	0.85
PLH2	0.95	0.98	0.92	0.88	0.7	0.61	0.74	0.83
BM25	0.96	0.96	0.97	0.91	0.77	0.83	0.91	0.9
ESA	0.96	0.95	0.94	0.84	0.58	0.58	0.74	0.8
Mittel	0.96	0.98	0.97	0.93	0.76	0.73	0.83	0.88

Tabelle 9: F-Measure-Werte für die Testkollektionen der Reuters-Sammlung, gemittelt über die fünf Dokumentziehungen.

Für die Testkollektionen mit drei Kategorien ergibt sich für alle Modellen ein sehr gutes Ergebnis. Oftmals wird mit einem F-Measure-Wert größer 0.95 ein nahezu perfektes Clustering erzeugt. In den Testkollektionen mit fünf und sechs Kategorien sinkt die Qualität der Clusterings. Vor allem in den Kollektionen der LA Times-Sammlung ist ein Abfall zu verzeichnen. Das Modell PLH2 ist dabei am stärksten betroffen. Die Modelle BM25, InBH2 und das statistische Sprachmodell (LM) erreichen mit 0,79 den größten mittleren F-Measure-Wert für die LA Times-Sammlung. Bei den Testkollektionen der Reuters-Sammlung erzielen die Modelle Lnu, BM25, Ltu, lnc und das statistische Sprachmodell von Ponte und Croft mit einem mittleren F-Measure-Wert größer 0,9 die besten Ergebnisse. Das ESA-Modell liegt mit einem Wert von 0,8 an letzter Stelle.

Insgesamt ergeben sich nur sehr kleine qualitative Unterschiede zwischen den einzelnen Modellen. Selbst das ESA-Modell, das durch die Verwendung einer externen Wissensquelle eine zusätzliche Komponente einführt, folgt den beobachteten Qualitätsschwankungen unter den Testkollektionen. In Tabelle 10 sind zusammenfassend die über alle Testkollektionen gemittelten F-Measure-Werte $\overline{\mathrm{F}}$ der Modelle eingetragen.[22]

[22] Zur Berechnung der Mittelwerte wurden die Testkollektionen LC5E500 und LC6E600 einbezogen. Für die Rangfolge ergibt sich aber auch ohne die beiden Kollektionen die präsentierte Konstellation.

Modell	$\overline{F}$	Modell	$\overline{F}$	Modell	$\overline{F}$
LM	0.86	Ltu	0.83	TfIdf	0.80
lnc	0.85	Lnu	0.83	ESA	0.78
BM25	0.85	INBH2	0.82	PLH2	0.67

Tabelle 10: Ranking der Retrieval-Modelle. Das statistische Sprachmodell von Ponte und Croft erreicht für die 14 Testkollektion den größten mittleren F-Measure-Wert.

5.4 Vergleich der Ansätze zur Modellvereinfachung

In Kapitel 3.5 wurde gezeigt, dass durch zufällige Termüberschneidungen in den Dokumenten viele kleine Ähnlichkeitswerte entstehen, denen keine thematische Verwandtschaft zu Grunde liegt. Durch ihre Vielzahl machen diese Ähnlichkeitswerte einen beachtlichen Teil der Masse im Ähnlichkeitsgraphen aus. Um die damit verbundene Problemtik für die Fusionierung zu verdeutlichen, sind in den Tabellen 11 und 12 die Werte des F-Measures eingetragen, die sich ohne Modellvereinfachung für den Vergleich der Retrieval-Modelle im vorherigen Experiment ergeben hätten.

Roh	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
lnc	0.17	0.17	0.3	0.07	0.07	0.16	0.05
LM	0.17	0.33	0.3	0.07	0.07	0.16	0.05
Lnu	0.17	0.17	0.38	0.07	0.07	0.16	0.05
TfIdf	0.33	0.49	0.54	0.07	0.07	0.16	0.05
Ltu	0.33	0.17	0.3	0.07	0.07	0.16	0.05
INBH2	0.49	0.33	0.47	0.07	0.07	0.16	0.05
PLH2	0.17	0.17	0.38	0.07	0.07	0.16	0.05
BM25	0.17	0.17	0.3	0.07	0.07	0.16	0.05
ESA	0.44	0.51	0.62	0.07	0.07	0.38	0.05

Tabelle 11: F-Measure für die Testkollektionen der LA-Times - Sammlung ohne Modellvereinfachung.

Roh	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
lnc	0.55	0.55	0.3	0.27	0.07	0.34	0.07
LM	0.63	0.55	0.3	0.07	0.07	0.28	0.05
Lnu	0.81	0.64	0.3	0.07	0.07	0.16	0.05
TfIdf	0.52	0.32	0.51	0.46	0.19	0.28	0.17
Ltu	0.64	0.6	0.3	0.28	0.07	0.16	0.05
INBH2	0.68	0.68	0.42	0.51	0.16	0.16	0.18
PLH2	0.750	0.76	0.67	0.33	0.14	0.22	0.05
BM25	0.52	0.56	0.3	0.1	0.07	0.16	0.05
ESA	0.54	0.54	0.7	0.17	0.21	0.16	0.05

Tabelle 12: F-Measure für die Testkollektionen der Reuters-Sammlung ohne Modellvereinfachung.

Im Vergleich zu den Resultaten mit Modellvereinfachung (Kapitel 5.3) zeigen die Tabellen für alle Testkollektionen einen erheblichen Qualitätsverlust. Am deutlichsten ist er in den Testkollektionen der LA Times-Sammlung mit fünf und sechs Kategorien zu erkennen. Mit einer Ausnahme werden die Dokumente dort zu einem einzigen Cluster zusammengefasst.

Im vorangegangenen Experiment wurde zur Vereinfachung der Ähnlichkeitsmatrix ein kNN-Graph verwendet. Hier nun werden die Ergebnisse präsentiert, die mit den im Rahmen dieser Arbeit entwickelten und in Kapitel 3.5 vorgestellten parameterlosen Verfahren zur Modellvereinfachung erzielt werden. Um die Leistungsfähigkeit der Ansätze zu beurteilen, werden die Resultate des kNN-Graphen als Referenzwerte herangezogen. Die dort erreichten Qualitätswerte stehen, wie im speziellen Korrelationsexperiment gezeigt, mit dem besten Ergebnis für die Modellvereinfachung mit einem kNN-Graphen signifikant in Zusammenhang.[23]

5.4.1 Vereinfachung mit Local Average

Die Modellvereinfachung Local Average (3.5.3) eliminiert unter allen Ähnlichkeitswerten eines Dokuments diejenigen, die kleiner als das arithmetische Mittel der Werte sind. In den Tabellen 13 und 14 sind die F-Measure-Werte eingetragen, die mit der Local-Average-Vereinfachung für den Vergleich der Retrieval-Modelle erzielt werden.

[23] Ausahmen sind die beiden Kollektionen LC5E500 und LC6E600.

Local Av.	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
lnc	0.97*	0.97*	0.70	0.07	0.07	0.31	0.05
LM	0.97*	0.97*	0.75	0.07	0.07	0.25	0.05
Lnu	0.73*	0.24	0.64	0.07	0.12	0.16	0.053
TfIdf	0.97*	0.98*	0.87	0.07	0.14	0.43	0.07
Ltu	0.66	0.49	0.69	0.07	0.07	0.16	0.05
INBH2	0.96*	0.97*	0.97*	0.10	0.08	0.45	0.07
PLH2	0.93*	0.96*	0.83	0.07	0.1	0.3*	0.07
BM25	0.87*	0.81	0.69	0.07	0.07	0.16	0.05
ESA	0.94*	0.83*	0.95*	0.38	0.29	0.44	0.22

Tabelle 13: F-Measure-Werte der Testkollektionen aus der LA Times-Sammlung unter Verwendung der Local-Average-Modellvereinfachung. Die mit Asterisk gekennzeichneten Werte liegen nicht weiter als 0,1 unter den Referenzwerten aus Tabelle 8.

Local Av.	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
lnc	0.97*	0.98*	0.99*	0.83	0.64	0.58	0.66
LM	0.98*	0.98*	0.99*	0.83	0.64	0.58	0.58
Lnu	0.98*	0.99*	0.99*	0.86	0.47	0.54	0.59
TfIdf	0.96*	0.98*	0.99*	0.92*	0.61	0.4	0.52
Ltu	0.99*	0.9*	0.90*	0.98*	0.43	0.26	0.52
INBH2	0.96*	0.94*	0.99*	0.93*	0.51	0.54*	0.64
PLH2	0.92*	0.94*	0.99*	0.91*	0.5	0.52*	0.54
BestMatch	0.98*	0.85	0.90*	0.85	0.29	0.16	0.44
ESA	0.89*	0.64	0.96*	0.61	0.47	0.35	0.37

Tabelle 14: F-Measure-Werte der Testkollektionen aus der Reuters-Sammlung unter Verwendung der Local-Average-Modellvereinfachung. Die mit Asterisk gekennzeichneten Werte liegen nicht weiter als 0,1 unter den Referenzwerten aus Tabelle 9.

In den Testkollektionen mit drei Kategorien kann die Vereinfachung mit Local Average überzeugen. In der Testkollektion RC3U1500 erreichen die erzeugten Clusterings sechs von neun Mal einen F-Measure-Wert von 0,99. Mit Zunahme der Anzahl an Kategorien werden die Referenzwerte nur noch vereinzelt erreicht und tendenziell zu wenige Cluster gebildet. Tabelle 15 zeigt die gemittelte Abweichung $\overline{v}$ von der richtigen Anzahl an Clustern. Es wird vermutet, dass der Wert negativ mit dem Rauschanteil in der Ähnlichkeitsmatrix korreliert ist.

	LC3E300	LC3E1500	LCU1500	LC5E500	LC5E1500	LC5U2500	LC6E600
$\overline{v}$	+0.2	0	-0.6	-3.5	-3.6	-3.4	-4.6
	RC3E300	RC3E1500	RCU1500	RC5E500	RC5E1500	RC5U2500	RC6E600
$\overline{v}$	+0.8	+0.6	+0.3	+0.2	-1.7	-2.4	-1.6

Tabelle 15: Über die Retrieval-Modelle gemittelte Abweichung $\overline{v}$ von der richtigen Anzahl an Clustern bei Verwendung der Local-Average-Modellvereinfachung. Für die Testkollektion LC6E600 gibt der Wert -4.6 beispielweise an, dass die Clusterings der Kollektion im Schnitt $6 - 4.6 = 1.4$ Cluster enthalten.

5.4.2 Vereinfachung mit dem Relevance-Ratio-Verfahren

Die Modellvereinfachung nach dem Relevance-Ratio-Verfahren fügt aus der Menge der Ähnlichkeitswerte so lange den jeweils höchsten Ähnlichkeitswert in eine anfangs leere Ähnlichkeitsmatrix ein, bis der von der Matrix implizierte Ähnlichkeitsgraph zusammenhängend ist. Die Anzahl der eingefügten Ähnlichkeitswerte wird daraufhin in das Verhältnis zur Anzahl der insgesamt vorhandenen Ähnlichkeitswerte gesetzt. Dieses Verhältnis bestimmt für jedes Dokument den Anteil der zu erhaltenden Ähnlichkeitswerte.

In den Tabellen 16 und 17 sind die F-Measure-Werte abgebildet, die sich beim Vergleich der Rerieval-Modelle unter Verwendung der Relevance Ratio einstellen.

Rel. Ratio	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
lnc	0.93*	0.93*	0.97*	0.56*	0.64*	0.62*	0.52*
LM	0.95*	0.95*	0.97*	0.55	0.54	0.63*	0.41
Lnu	0.28	0.17	0.46	0.07	0.07	0.16	0.05
TfIdf	0.8	0.73	0.78	0.57*	0.61*	0.6*	0.51*
Ltu	0.7	0.33	0.62	0.07	0.07	0.16	0.05
INBH2	0.95*	0.8	0.81	0.34	0.14	0.62	0.11
PLH2	0.72	0.79	0.72	0.41	0.28	0.47*	0.32*
BM25	0.96*	0.98*	0.86	0.25	0.33	0.53	0.18
ESA	0.92*	0.9*	0.89*	0.69*	0.69*	0.65*	0.32

Tabelle 16: F-Measure-Werte der Testkollektionen aus der LA Times-Sammlung unter Verwendung der Relevance-Ratio-Modellvereinfachung.

Rel.Ratio	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
lnc	0.90*	0.93*	0.86	0.94*	0.91*	0.8*	0.84*
LM	0.93*	0.95*	0.87	0.94*	0.9*	0.76*	0.81
Lnu	0.98*	0.99*	0.99*	0.96*	0.74	0.57	0.54
TfIdf	0.8	0.75	0.57	0.83*	0.62*	0.67*	0.75*
Ltu	0.85	0.84	0.88	0.91*	0.73*	0.73*	0.84*
INBH2	0.59	0.64	0.59	0.81	0.74*	0.61*	0.68*
PLH2	0.48	0.55	0.39	0.67	0.47	0.66*	0.63*
BM25	0.62	0.52	0.67	0.68	0.63	0.69	0.76
ESA	0.95*	0.85*	0.90*	0.72	0.47	0.36	0.67*

Tabelle 17: F-Measure-Werte der Testkollektionen aus der Reuters-Sammlung unter Verwendung der Relevance-Ratio-Modellvereinfachung. Die mit Asterisk gekennzeichneten Werte liegen nicht weiter als 0,1 unter den Referenzwerten aus Tabelle 9.

Im Gegensatz zu den anderen Modellvereinfachungen zeigt sich für die Relevance Ratio ein deutlicher Qualitätsunterschied unter den Retrieval-Modellen. Die Modelle lnc, ESA, TfIdf und LM erreichen in allen Testkollektionen eine gute Anäherung an die Referenzergebnisse. Für die übrigen Modelle zeigt sich ein differenziertes Bild. Das Referenzniveau wird jeweils nur in einigen unterschiedlichen Testkollektionen erreicht.

5.4.3 Vereinfachung mit Major Expected Density

Das Verfahren Major Expected Density vereinfacht den Ähnlichkeitsgraphen während der Fusionierung mit Major Clust. Bei jeder Entscheidung über die Zuweisung eines Clusters zu einem Dokument wird mit dem Qualitätsindex Expected Density geprüft, ob sich durch Eliminierung kleinster Ähnlichkeitswerte ein besseres temporäres Clustering ergibt.

Die F-Measure-Werte, die die so erzeugten Clusterings erreichen, sind für die 14 Testkollektionen in den Tabellen 18 und 19 dargestellt.

Maj.Ex.	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
lnc	0.97*	0.96*	0.79	0.21	0.13	0.43	0.09
LM	0.97*	0.97*	0.85	0.14	0.07	0.48	0.07
Lnu	0.81*	0.78*	0.61	0.07	0.07	0.16	0.05
TfIdf	0.95*	0.97*	0.86	0.35	0.20	0.52	0.34
Ltu	0.9*	0.7	0.69	0.07	0.07	0.2	0.07
INBH2	0.95*	0.97*	0.97*	0.4	0.25	0.55	0.19
PLH2	0.93*	0.97*	0.97*	0.27*	0.13	0.54	0.13
BM25	0.95*	0.88*	0.4	0.21	0.14	0.3	0.09
ESA	0.94*	0.93*	0.92*	0.58	0.51	0.65*	0.27

Tabelle 18: F-Measure-Werte der Testkollektionen aus der LA Times-Sammlung unter Verwendung der Major-Expected-Density-Modellvereinfachung.

Maj.Ex.	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
lnc	0.96*	0.98*	0.99*	0.88*	0.57	0.59	0.65
LM	0.95*	0.98*	0.99*	0.87*	0.6	0.58	0.62
Lnu	0.96*	0.99*	0.99*	0.96*	0.52	0.59	0.69
TfIdf	0.94*	0.99*	0.96*	0.95*	0.57*	0.45	0.58
Ltu	0.97*	0.99*	0.94*	0.96*	0.56	0.59	0.61
INBH2	0.93*	0.98*	0.93*	0.94*	0.56	0.54*	0.64
PLH2	0.88*	0.98*	0.99*	0.89*	0.55	0.53*	0.69*
BM25	0.91*	0.94*	0.9*	0.97*	0.44	0.39	0.61
ESA	0.96*	0.95*	0.96*	0.69	0.47	0.36	0.47

Tabelle 19: F-Measure-Werte der Testkollektionen aus der Reuters-Sammlung unter Verwendung der Major-Expected-Density-Modellvereinfachung. Die mit Asterisk gekennzeichneten Werte liegen nicht weiter als 0,1 unter den Referenzwerten aus Tabelle 9.

Die Tabellen offenbaren einen Qualitätsverlauf ähnlich dem der Modellvereinfachung mit Local Average. Während für die Testkollektionen mit drei Kategorien sehr gute Clusterings erzeugt werden, liegen die F-Measure-Werte in den Testkollektionen mit fünf und sechs Kategorien unter den Referenz-Werten. Das Quallitätsniveau liegt insgesamt über dem des Local-Average-Verfahrens.

5.4.4 Vereinfachung mit dem Expected-Similarity-Verfahren

Das Modellvereinfachungsverfahren Expected Similarity schätzt auf Grundlage der Termgewichtsverteilung in der Dokumentkollektion den Wert, der bei der Ähnlichkeitsberechnung zwischen zwei Dokumenten zu erwarten ist. Nur Ähnlichkeitswerte, die signifikant über dem Erwartungswert liegen, bleiben der Ähnlichkeitsmatrix erhalten.

Die durch Anwendung der Expected-Similarity-Vereinfachung erhaltenen F-Measure-Werte für die Testkollektionen sind in den Tabellen 20 und 21 eingetragen.

Exp. Sim.	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
lnc	0.9*	0.96*	0.97*	0.62*	0.56*	0.68*	0.43
LM	0.92*	0.97*	0.97*	0.66*	0.64*	0.69*	0.6*
Lnu	0.81	0.92*	0.91*	0.55*	0.56*	0.61*	0.43
TfIdf	0.88*	0.97*	0.96*	0.51	0.62*	0.69*	0.42*
Ltu	0.79*	0.96*	0.96*	0.6*	0.65*	0.66*	0.55*
INBH2	0.75	0.96*	0.92*	0.63*	0.69*	0.73*	0.59*
PLH2	0.7	0.94*	0.83*	0.27*	0.13*	0.4*	0.13
BM25	0.88*	0.97*	0.97*	0.68*	0.55	0.69*	0.65*
ESA	0.9*	0.90*	0.9*	0.69*	0.68*	0.75*	0.61*

Tabelle 20: F-Measure-Werte der Testkollektionen aus der LA Times-Sammlung unter Verwendung der Expected-Similarity-Modellvereinfachung.

Exp. Sim.	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
lnc	0.90*	0.98*	0.99*	0.95*	0.89*	0.89*	0.86*
LM	0.94*	0.98*	0.98*	0.94*	0.9*	0.9*	0.87*
Lnu	0.91*	0.98*	0.99*	0.97*	0.8*	0.96*	0.86*
TfIdf	0.76	0.95*	0.89*	0.89*	0.54*	0.39	0.78*
Ltu	0.87	0.98*	0.99*	0.93*	0.79*	0.70*	0.88*
INBH2	0.68	0.92*	0.83	0.83	0.72*	0.67*	0.7*
PLH2	0.75	0.88*	0.86*	0.84*	0.75*	0.66*	0.75*
BM25	0.79	0.91*	0.97*	0.93*	0.84*	0.83*	0.90*
ESA	0.93*	0.92*	0.89*	0.85*	0.58*	0.38	0.75*

Tabelle 21: F-Measure-Werte der Testkollektionen aus der Reuters-Sammlung unter Verwendung der Expected-Similarity-Modellvereinfachung. Die mit Asterisk gekennzeichneten Werte liegen nicht weiter als 0,1 unter den Referenzwerten aus Tabelle 9.

Die Tabellen zeigen für alle Testkollektionen und Retrieval-Modelle eine gute Annäherung an die Referenzwerte. In 109 von 126 Fällen liegt der erzielte F-Measure-Wert nicht weiter als 0,1 unter dem Referenzwert. Die Resultate, die das Referenzniveau nicht erreichen, konzentrieren sich in den Testkollektionen LC3E300 und RC3E300.

5.4.5 Fazit

Jedes der entwickelten Verfahren zur Modellvereinfachung ist in der Lage, die Fusionierung mit Major Clust positiv zu beeinflussen. Mit dem Ansatz der Expected Similarity ist es indes gelungen, für alle Testkollektionen überzeugende

Qualitätswerte zu produzieren. Das parameterlose Verfahren erreicht in 86,5% der untersuchten Fälle das Qualitätsniveau der durch externes Wissen unterstützen Modellvereinfachung mit einem kNN-Graphen.

6 Zusammenfassung und Ausblick

Die vorliegende Arbeit beschäftigt sich zentral mit der Ähnlichkeitsbestimmung zwischen Dokumenten im Dokumenten-Clustering und der Evaluierung dazu eingesetzter Verfahren.

Eingehend auf die Problemstellung im Information Retrieval wird gezeigt, dass das Dokumenten-Clustering geeignet ist, um die Retrieval-Qualität zu verbessern. Es werden Retrieval-Modelle vorgestellt, die die Ähnlichkeit zwischen den Dokumenten einer Dokumentkollektion quantifizieren. Im Wesentlichen ergeben sich quantitative Größen aus der absoluten Häufigkeit der in einem Dokument verwendeten Terme sowie deren Verbreitung in der Dokumentkollektion. Die Berechnungsvorschriften unterscheiden sich in der Art und Weise wie die Größen kombiniert und normalisiert werden. Eine vielversprechende Ausnahme stellt das explizit semantische Indexierungsmodell dar, das mit einer externen Wissensquelle zusätzliche Informationen in die Berechnung einbringt. Im experimentellen Vergleich erweist sich das Modell jedoch nicht als überlegen. Eine Schwachstelle der verwendeten Implementierung könnte die zufällige Auswahl der zur Wissensrepräsentation genutzten Artikel sein. Es wäre daher zu evaluieren, ob eine gezieltere Auswahl der Artikel die Leistungsfähigkeit des Verfahrens steigert. Um mit den gewählten Artikeln eine gute thematische Abdeckung zu garantieren, könnten in einem zukünftigen Verfahren ein Clustering für die Wissensquelle generiert und die Artikel aus den entstandenen Clustern gewählt werden. Ebenso denkbar wäre die Verwendung des latent semantischen Indexierungsverfahrens, um die Artikel der Wissenquelle in einen Konzeptraum geeigneter Dimensionalität zu projizieren.

Die Arbeit stellt heraus, dass der Fusionierungsprozess im Dokumenten-Clustering von Ähnlichkeiten beeinflusst wird, die sich rein zufällig aus der gemeinsamen Verwendung unbedeutender Terme in den Dokumenten ergeben. Obwohl der Betrag dieser Ähnlichkeiten sehr klein ist, erwächst aus ihrer großen Anzahl die Gefahr falscher Mehrheitsentscheidungen bei der Fusionierung. Neue, parameterlose Verfahren werden präsentiert, die unbedeutende Ähnlichkeitswerte eliminieren. Insbesondere das Verfahren Expected Similarity zeigt vielversprechende Ergebnisse für die betrachteten Testkollektionen. Eine Aufgabe für zukünftige Untersuchungen wird es sein, das Verfahren für größere Dokumentkollektionen zu testen.

Literatur

[Amati02] Amati G., van Rijsbergen C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. In: ACM Trancactions on Information Systems, 20(4), oct 2002, S. 357–389.

[Backhaus06] Backhaus K, Erichson B., Plinke W., Weiber R.: Multivariate Analysemethoden. Springer, Berlin 2006. S. 489-555. ISBN: 3-540-27870-2

[Bezdek95] Bezdek J.C., Pal N. R.: Cluster Validation with Generalized Dunn's Indices. In: N. Kasabov and G. Coghill, Proceedings of the 2nd international two-stream conference on ANNES, S. 190–193, Piscataway, NJ, 1995. IEEE Press.

[Buckley98] Buckley C., Singhal A., Mitra M.: New Retrieval Approaches Using SMART: TREC 4. In: Text Retrieval Conference 4th, DIANE Publishing, 1998. ISBN: 0788171690

[Davis79] Davies D.L., Bouldin D.W.: A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Learning, 1(2), 1979.

[Dumais04] Dumais S.T.: Latent semantic analysis. In: Annual Review of Information Science and Technology 38, S. 189-229.

[Dunn74] Dunn J.C.: Well seperated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4(1974):95-104, 1974.

[MeyerZuEissen07] Meyer zu Eissen, S.: On Information Need and Categorizing Search. 2007, http://ubdata.uni-paderborn.de/ediss/17/2007/meyer_zu/ (Abruf: 13.3.2008)

[Feldmann03] Feldman, R.: Mining Text Data. In Ye, N. (Ed.), The Handbook Of Data Mining (pp. 481- 518). London 2003, Erlbaum Associates, Publishers.

[Fuhr96] Fuhr, N.: Fachgruppe Information Retrieval in der Gesellschaft für Informatik (GI). `http://www.uni-hildesheim.de/fgir/index.php?option=com_content&task=view&id=14&Itemid=41` (Abruf:13.3.2008)

[Fuhr06] Fuhr, N.: Information Retrieval Skriptum zur Vorlesung. `http://www.is.informatik.uni-duisburg.de/courses/ir_ss04/folien/irskall.pdf` (Abruf: 13.3.2008)

[Gabrilovich07] Gabrilovich E., Markovitch S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, Indien. 2007, S. 1606–1612.

[Harman86] Harman D.: An experimaental study of factors important in document ranking. In:Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval. ACM, New York 1986.

[Harter75] Harter S.P.: A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. In: J. ASIS 26, S. 197-216.

[Hearst96] Hearst M. A., Pedersen J. O.: Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In: SIGIR'96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, S. 76–84, New York, NY, USA, 1996. ACM Press. ISBN 0-89791-792-8. doi: `http://portal.acm.org/citation.cfm?doid=243199.243216`.

[Hole05] Hole G.: Table of critical values for Spearman's rho. In: Graham Hole's Resource Page, `http://www.sussex.ac.uk/Users/grahamh/RM1web/Spearmanstable2005.pdf`

[Jain90] Jain A.K., Dubes R.C.: Algorithms for Clustering in Data. Prentice Hall, Englewood Cliffs, NJ, 1990. ISBN 0-13-022278-X.

[Jin01] Jin R., Falusos C., Hauptmann A.G.: Meta-scoring: automatically evaluating term weighting schemes in IR without precision-recall. In:SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research

and development in information retrieval, S.83-89, New Orleans 2001. ISBN: 1-58113-331-6 doi:http://doi.acm.org/10.1145/383952.383964

[Kruskal56] Kruskal, J.: On the shortest spanning subtree and the traveling salesman problem. In: Proceedings of the American Mathematical Society 7, S.48-50, 1956.

[Ponte98] Ponte J. M., Croft W. B.: A language modeling approach to information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, S. 275-281, Melbournem Australia,1998. ACM Press. ISBN:1-58113-015-5. doi: http://doi.acm.org/10.1145/290941.291008

[Porter80] Porter M.F.: An Algorithm for Suffix Stripping. Program, 14(3):130-137, 1980.

[Rauch88] Rauch W.: Was ist Informationswissenschaft? In: Grazer Universitätsreden,32, Kienreich Graz, 1988.

[vanRijsbergen79] van Rijsbergen C. J. : Information Retrieval. Buttersworth, London, 1979.

[Robertson76] Robertson S. E., Sparck-Jones K.: Relevance weighting of search terms. In: Journal of the American Society for Information Science, 27, 3, 129-46, May-Jun 76

[Robertson94] Robertson S., Walker S.: Some simple approximations to the 2-Poisson model for probabilistic weighted retrieval. In: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (Dublin), Springer Verlag, New York 1994, S. 232-241.

[Robertson95] Robertson S.E., Walker S., Jones S., Hancock-Beaulieu M.M., Gatford M.: Okapi at TREC-3. In: Overview of the Third Text REtrieval Conference (TREC-3) NIST, Gaithersburg 1995, MD, pp. 109–126 NIST Special Publication 500-225.

[Robertson96] Robertson S.E.,Walker S., Hancock-Beaulieu M.M., Gatford M, Payne A.: Okapi at TREC-4.

[Robertson04] Robertson S.: Understanding inverse document frequency: on theoretical arguments for IDF. In: Journal of Documentation 60, S. 503-520.

[Rose02] Rose T.G., Stevenson M., Whitehead M.: The Reuters Corpus Volume 1 - From Yesterday's News to Tomorrows's Language Resources. In: Proceedings of the Third International Conference on Language Resources and Evaluation, 2002.

[Salton68] Salton G.: Automatic Information Organization and Retrieval. McGraw Hill Text, NewYork 1968.

[Singhal96] Singhal A., Buckley C., Mitra M., Slaton G.: Pivoted Document Length Normalization, Cornell University, nov. 1995. `http://techreports.library.cornell.edu:8081/Dienst/UI/1.0/Display/cul.cs/TR95-1560`

[Spearman04] Spearman C.: The proof and measurment of association between two rings", Amer. J. Psychol., 15, S. 72-101, 1904.

[Stein04a] Stein B. M.: Web-Technologie II (WS2004/05): Clusteranalyse , Einführung. `http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/machine-learning/part-cluster-analysis/unit-cluster-analysis-intro.ps.pdf` (Aufruf: 13.3.2008)

[Stein04b] Stein B. M.: Web-Technologie II (WS2004/05): Clusteranalyse , Hierarchische Verfahren. `http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/machine-learning/part-cluster-analysis/unit-cluster-analysis-hierarchical.ps.nup.pdf` (Aufruf: 13.3.2008)

[Stein04c] Stein B. M.: Web-Technologie II (WS2004/05): Clusteranalyse , Iterative Verfahren. `http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/machine-learning/part-cluster-analysis/unit-cluster-analysis-iterative.ps.nup.pdf` (Aufruf: 13.3.2008)

[Stein04d] Stein B. M.: Web-Technologie II (WS2004/05): Clusteranalyse (Fortsetzung), Dichtebasierte Verfahren. `http://www.`

uni-weimar.de/medien/webis/teaching/lecturenotes/machine-learning/part-cluster-analysis/unit-cluster-analysis-density.ps.pdf (Aufruf: 13.3.2008)

[Stein04e] Stein B.M.: Web-Technologie II (WS2004/05): Modelle und Prozesse im IR. http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/information-retrieval/part-retrieval-models/unit-retrieval-models.ps.pdf (Aufruf:13.3.2008)

[Stein04f] Stein B.M.: Web Technologie II (WS2004/05): Indexterme. http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/information-retrieval/part-ir-basics/unit-index-terms.ps.pdf (Aufruf:13.3.2008)

[Stein08] Stein B.M., Meyer zu Eissen S.: Computational Aspects of Genre Classification.

[Steinbach00] Steinbach M., Karypis G., Kumar V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining, 2000. steinbach citeseer.ist.psu.edu/steinbach00comparison.html (Aufruf: 30.3.2008)

[Stock07] Stock W. G.: Information Retrieval, Informationen suchen und finden. Oldenbourg, München 2007, S. 335-353. ISBN: 3-486-58172-4

[Swanson74] Swanson D., Bookstein A.: Probabilistic models for automatic indexing. Journal of the American Society for Information Science 25, S. 312-318.

[Tibshirani00] Tibshirani R., Walther G., Hastie t.: Estimating the number of clusters via the gap statistic. Journal of Royal Statistics and Social Behaviour, 63 (2):411–423, 2001.

[Vorhees85] Voorhees, E. M.: The cluster hypothesis revisited. In: SIGIR'85: Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, S. 188–196, New York, NY, USA, 1985.

ACM Press. ISBN 0-89791-159-8. doi:http://doi.acm.org/10.1145/253495.253524.

[Yamron97] Yamron J.: Topic Detection and Tracking Segmentation Task. In: Proceedings of the Topic Detection and Tracking Workshop, Oct. 1997.