Optimum Clustering Framework

Nicolas Schönfeld

Seminar: Information Retrieval
Universität Koblenz-Landau, Campus Koblenz
nschoenfeld@uni-koblenz.de

Zusammenfassung Um den vorhandenen Nutzen von Dokumenten-Clustering im interaktiven Retrieval zu festigen, wird in dieser Arbeit ein Rahmenwerk präsentiert, das es ermöglicht durch Vermeidung von Heuristiken eine theoretische Grundlage für ein optimales Clustering von Dokumenten zu schaffen und so den Weg zu ebnen für eine gezieltere Entwicklung effektiver Clustering-Methoden für Dokumente.
Durch Festlegen der dafür notwendigen Bestandteile einer Clustering-Methode und der Einführung einer auf Relevanzwahrscheinlichkeit beruhenden Metrik zur Bestimmung der Qualität eines Clusterings, wird schließlich mithilfe von Schätzungen der Relevanzwahrscheinlichkeit von Anfrage-Dokumenten-Paaren die Eigenschaft des optimalen Clusterings definiert.

1 Einleitung

Das Clustering von Dokumenten bzw. Informationen nimmt heutzutage nicht ohne Grund einen wichtigen Platz im Themengebiet des Information Retrieval ein, da es großes Potenzial besitzt, den Benutzer bei interaktivem Information Retrieval zu unterstützen. So kann es besonders dann von Nutzen sein, wenn Benutzer ihre Anfragen nicht genau formulieren können bzw. einen nur sehr unscharf zu charakterisierenden Informationsbedarf haben, da sie im Vorfeld noch nicht genau wissen, wonach sie eigentlich suchen und beispielsweise nur ein grobes Thema haben, über das sie Näheres erfahren möchten. Durch Verwendung des Scatter/Gather-Verfahrens [1] in Verbindung mit effektiven Cluster-Methoden können dem Benutzer die möglicherweise relevanten Informationen strukturiert in Form verschiedener Cluster mit kurzen Zusammenfassungen über deren Inhalt präsentiert werden, aus denen dieser die für ihn relevanten Cluster auswählt. Diese ausgewählten Cluster werden daraufhin immer weiter zerteilt, sodass es dem Benutzer möglich ist entweder direkt für ihn relevante Dokumente zu finden oder durch die gewonnenen Informationen seine Suchanfrage besser formulieren zu können.

Im Folgenden geht es nun weniger darum ein weiteres Clustering-Verfahren zu entwerfen, es soll vielmehr ein Rahmenwerk präsentiert werden, das die theoretischen Grundlagen legt, um bereits bestehende Clustering-Verfahren gezielt

verbessern zu können und ein optimales Clustering von Dokumenten zu ermöglichen. Neben der Vermeidung von Heuristiken bei der Designentscheidung von Clustering-Verfahren, spielt dabei das Miteinbeziehen der Relevanz von Dokumenten eine wichtige Rolle. Das in den folgenden Kapitel nun näher vorgestellte Rahmenwerk entstammt [2].

Bevor näher auf dieses Rahmenwerk eingegangen wird, sollten zunächst jedoch ein paar grundlegende Begriffe aus dem Information Retrieval erläutert werden, die für das Verständnis dieser Arbeit notwendig sind.

2 Grundlagen

2.1 Relevanz

Die Relevanz stellt im Information Retrieval die Beziehung zwischen einer Anfrage und einem Dokument dar. Ein Dokument gilt als relevant, wenn es den Wissensbedarf des Benutzer deckt. Dabei kommt es jedoch zu dem Problem, dass für eine Beurteilung von Retrievalergebnissen die richtige Antwort bekannt sein muss, um die Antwort des Systems bewerten zu können. In der Praxis gibt es daher keine einheitliche Definition von Relevanz, stattdessen unterscheidet man zwischen drei verschiedenen Arten von Relevanz: subjektive, objektive und geschätzte Relevanz. [3]

Wie in Abb. 1 verdeutlicht, beschreibt die **subjektive Relevanz** die Beziehung zwischen einem Dokument und dem Informationsbedürfnis des Benutzers. Der Benutzer selbst entscheidet, inwiefern das jeweilige Dokument zur Deckung seines Informationsbedürfnisses von Nutzen ist. Daher ist die Relevanzbeurteilung in diesem Fall stets subjektiv.

Die zweite Form der Relevanz ist die **objektive Relevanz**. Die Tatsache, dass ein Benutzer bei der Verwendung eines Retrieval-Systems seinen Informationswunsch in Form einer Anfrage formulieren muss, führt üblicherweise dazu, dass der Informationswunsch des Benutzers durch die gewählte Repräsentation nur unvollständig widergespiegelt wird. Als objektive Relevanz bezeichnet man die Beziehung zwischen einem Dokument und eben dieser, das Informationsbedürfnis nur unvollständig widerspiegelnden, Anfrage. Bestimmt wird die objektive Relevanz von außenstehenden Personen, beispielsweise einer Gruppe von unabhängigen Experten.

Wie schon die objektive Relevanz, so beschreibt auch die **geschätzte Relevanz** die Beziehung zwischen einem Dokument und einer Anfrage, wobei die Relevanz in diesem Fall von dem Retrieval-System anhand bestimmter Regeln berechnet wird.

Da oft beispielsweise ein von Experten als relevant eingestuftes Dokument in Wirklichkeit für das Informationsbedürfnis des Benutzers gar nicht relevant ist,

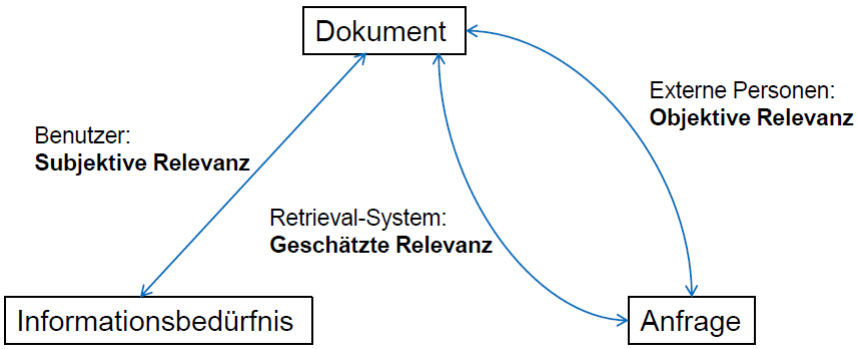

Abbildung 1. Die drei verschiedenen Arten von Relevanz

ist es nicht selten der Fall, dass die drei vorgestellten Arten der Relevanz beim Information Retrieval widersprüchliche Ergebnisse hinsichtlich der Relevanz liefern.

2.2 Effektivität

Effektivität im Information Retrieval lässt sich am besten beschreiben, als ein Maß für die Fähigkeit eines Systems, relevante Dokumente anzuzeigen, während nicht relevante Dokumente zurückgehalten werden. [4]

Neben einer subjektiven Einschätzung der Effektivität eines Retrieval-Systems gibt es auch allgemeine Maßzahlen, mit denen sich verschiedene Systeme vergleichen lassen. Die beiden Wichtigsten sind *Precision* und *Recall*.

Precision beschreibt dabei den Anteil der, vom Retrieval-System gefundenen, relevanten Dokumente im Verhältnis zu allen gefundenen Dokumenten. Je geringer der Wert, desto mehr irrelevante Dokumenten muss der Benutzer also betrachten.

Aufgabe des Recall ist es, den Anteil der relevanten Dokumente im Rechercheergebnis ins Verhältnis zu allen relevanten Dokumenten der Datenbasis zu setzen. Der Recall schätzt somit ab, wie gut das System zu einer bestimmten Anfrage alle relevanten Dokumente findet, was sich oft jedoch als schwierig berechenbar herausstellt, da dazu alle relevanten Dokumente der Datenbasis bekannt sein müssen.

Es sollte stets angestrebt werden, eine möglichst hohe Precision bei gleichzeitig hohem Recall zu erreichen. Welches der beiden Maße wichtiger ist, kann sich von System zu System unterscheiden. Für gewöhnlich besteht jedoch eine inverse Beziehung zwischen den beiden Maßen. Eine Steigerung des einen Maßes führt zu

einer Abnahme des anderen.

3 Überblick

Ähnlich wie seinerzeit das Probability Ranking Principle (PRP) [5] die Abhängigkeit von Heuristiken bei der Erstellung von Retrieval-Systemen beendete und eine theoretische Rechtfertigung für bestimmte probabilistische Retrieval-Modelle lieferte, die besagte, dass diese Modelle eine optimale Performance liefern, wenn die Resultate nach absteigender Wahrscheinlichkeit der Relevanz sortiert werden, so soll nun auch für das Clustering eine theoretische Grundlage geschaffen werden, die das Design eines Clustering-Verfahrens mit der Qualität seines Ergebnisses in Beziehung setzt. Statt weiterhin Clustering-Verfahren nur anhand von Experimenten zu bewerten, wird im Folgenden ein Rahmenwerk präsentiert, das bestmögliches Clustering ermöglicht, indem es die Dokumentendarstellung und das Ähnlichkeitsmaß mit der erwarteten Qualität des Clusterings in Verbindung bringt.

Ausgangspunkt des Optimum Clustering Frameworks (OCF) bildet die sogenannte Cluster-Hypothese. Diese besagt, dass *ähnliche Dokumente für gewöhnlich dazu neigen für dieselben Anfragen relevant zu sein.* [4]

Da das Ziel des OCF aber nicht das Auffinden möglicher relevanter Dokumente für eine bestimmte Anfrage ist, sondern es darum geht, ein bestmögliches Clustering von Dokumenten zu erreichen, wird die Cluster-Hypothese kurzerhand umgedreht: *Dokumente, die für dieselben Anfragen relevant sind, sollten zum selben Cluster gehören.*

Zu diesem Zweck wird eine Anfragensammlung zusammen mit Relevanzbeurteilungen eingeführt, sowie die Ähnlichkeit von Dokumenten neu definiert, sodass zwei Dokumente fortan als ähnlich bezeichnet werden, wenn sie für dieselben Anfragen relevant sind. Durch das Miteinbeziehen der Relevanz von Dokumenten bietet das Clustering, statt sich vorher auf einer rein semantischen Ebene zu bewegen, nun sogar einen pragmatischen Ansatz. Da die Relevanz üblicherweise jedoch nicht so einfach zu erkennen ist, wird stattdessen die Relevanzwahrscheinlichkeit betrachtet.

Um festzustellen, ob ein Clustering die Eigenschaft eines bestmöglichen Clusterings (*optimum clustering*) erfüllt, wird basierend auf Relevanzbeurteilungen eine Metrik für die Qualität eines Clusterings eingeführt. Das bestmögliche Clustering ist genau das Clustering, das die invertierte Cluster-Hypothese am besten erfüllt.

4 Cluster-Qualität

Um die Qualität eines Clusterings zu beurteilen, bedarf es einer geeigneten Metrik. Die üblicherweise verwendeten Metriken für eine externe Bewertung vergleichen ein Clustering mit einer gegebenen Klassifizierung eines Dokuments. Dies soll hier jedoch nicht geschehen, stattdessen soll die Bewertung hinsichtlich einer Anfragensammlung mit Relevanzbeurteilungen erfolgen. Die dafür erforderliche Metrik muss dazu folgende Anforderungen erfüllen:

- Die Metrik muss auf einer gegebenen Sammlung von Anfragen mit vollständigen Relevanzinformationen basieren.
- Es sollte möglich sein, Erwartungen dieser Metrik durch probabilistische Retrieval-Modelle zu berechnen.

Nach einigen formalen Definitionen wie der Clustering-Funktion oder der Relevanzequivalenz von Clusterings, die bei Bedarf in [2] nachgeschlagen werden können, kann nun damit begonnen werden die Metrik herzuleiten. Zum besseren Verständnis sei jedoch noch erwähnt, dass im Folgenden D für die Dokumentensammlung, Q für die Anfragensammlung, $\mathcal{R} \subset Q \times D$ für die Sammlung an relevanten Anfragen-Dokumenten-Paaren und $\mathcal{C}$ für ein Clustering, also eine Menge von Clustern, steht.

Aufbauend auf der Methode von Jardine und van Rijsbergen [6] zum Testen der Cluster-Hypothese, in der die Ähnlichkeiten von relevant-relevanten Dokumentenpaaren mit denen von relevant-irrelevanten Dokumentenpaaren verglichen wurden, wird die Metrik folgendermaßen definiert:

Für jede gegebene Anfrage werden die Paare von relevanten Dokumenten im selben Cluster gezählt und durch die Anzahl aller Paare innerhalb des Clusters geteilt. Diese Prozedur wird bei allen Clustern des Clusterings wiederholt und daraus der gewichtete Mittelwert gebildet. Da ähnlich wie der Precision im Information Retrieval die relevanten Dokumente bzw. Dokumentenpaare allen gefundenen Dokumenten bzw. Dokumentenpaaren gegenübergestellt werden, erhält das sich daraus ergebende neue Maß den Namen *pairwise precision* P_p:

$$P_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{1}{|D|} \sum_{\substack{C_i \in \mathcal{C} \\ c_i > 1}} c_i \sum_{q_k \in Q} \frac{r_{ik}(r_{ik} - 1)}{c_i(c_i - 1)} \tag{1}$$

Dabei steht c_i für die Größe des Clusters C_i und r_{ik} für die Anzahl der relevanten Dokumente im Cluster C_i bezüglich der Anfrage q_k.

Ähnlich wie bei der Bewertung der Effektivität von Retrieval-Systemen, so wird auch hier noch ein zweites Maß eingeführt, um einen weiteren Qualitätsaspekt abzudecken, der *pairwise recall* R_p. Dieser wird wieder ähnlich wie der klassische Recall definiert, mit dem Unterschied, dass Dokumentenpaare statt einzelne Dokumente für die Berechnung verwendet werden. Mit der Definition, dass g_k

die gesamte Anzahl der relevanten Dokumente für eine Anfrage q_k bezeichnet, ergibt sich folgende Definition für das Maß:

$$R_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{\sum_{q_k \in Q} \sum_{C_i \in \mathcal{C}} r_{ik}(r_{ik} - 1)}{\sum_{\substack{q_k \in Q \\ g_k > 1}} g_k(g_k - 1)} \tag{2}$$

Bei diesem Maß werden im Zähler wieder für jede Anfrage die Paare von relevanten Dokumenten in jedem Cluster gezählt und addiert und dann durch die Summe aller Paare von relevanten Dokumenten für diese Anfragen geteilt. Das getrennte Aufsummieren in Zähler und Nenner hat hierbei den Vorteil, dass nun unvoreingenommene Schätzwerte berechnet werden können, da sich der Nenner bei einer konstanten Anfragensammlung nicht verändert und somit ignoriert werden kann.

Die beiden vorgestellten Maße P_p und R_p werden abschließend noch durch die Bildung ihres harmonischen Mittels zu einem einzigen Maß zusammengefasst, dem *pairwise F-measure* F_p:

$$F_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{2}{\frac{1}{P_p(D,Q,\mathcal{R},\mathcal{C})} + \frac{1}{R_p(D,Q,\mathcal{R},\mathcal{C})}} \tag{3}$$

Eine ausführliche Begründung, warum das F-measure als Cluster-Metrik gut geeignet ist, findet sich in [2]. Dort wird gezeigt, dass es unter anderem die 4 Axiome von Ackerman und Ben-David [7] erfüllt, die für eine gute Metrik zur Bestimmung der Cluster-Qualität nötig sind.

5 Perfektes und optimales Clustering

Die im vorherigen Kapitel definierten Qualitätsmaße wurden eingeführt, um nun perfektes und optimales Clustering gegenüber zu stellen und näher definieren zu können. Die Begriffe *perfekt* und *optimal* wurden dabei analog zum klassischen Retrieval gewählt, bei dem perfektes Retrieval dadurch gekennzeichnet ist, dass die relevanten Dokumente allesamt vor dem ersten nicht-relevanten Dokument angeordnet werden. Dies ist jedoch nur mit externen Bewertungsmaßen möglich. Da ein Retrieval-System in der Praxis jedoch nicht den genauen Informationsbedarf des Benutzers kennt und auch die Semantik der einzelnen Dokumente nicht komplett erfassen kann, wird auf interne Bewertungsmaße zurückgegriffen und mit Repräsentationen dieser Dokumente gearbeitet, was jedoch lediglich ein optimales Retrieval bezüglich dieser Repräsentationen ermöglicht. Aufbauend auf den bereits eingeführten Qualitätsmaßen ist perfektes Clustering folgendermaßen definiert:

Ein Clustering $\mathcal{C}$ ist ein perfektes Clustering, wenn für eine Dokumenten- und Anfragensammlung, sowie dazugehöriger Relevanzrelation, kein anderes Clustering $\mathcal{C}'$ existiert, das entweder eine höhere pairwise precision und einen mindestens so hohen pairwise recall wie $\mathcal{C}$ aufweist, oder aber eine mindestens so hohe

pairwise precision und einen höheren pairwise recall.

Diese Definition stellt ein sogenanntes Pareto Optimum dar, das heißt es ist nicht möglich eines der beiden Maße zu verbessern ohne gleichzeitig das andere Maß zu verschlechtern. Wie bei Pareto Optima üblich, so kann es jedoch auch hier mehrere perfekte Lösungen geben.

Um auch optimales Clustering zu definieren, wird von den externen Bewertungsmaßen nun auf interne Maße gewechselt. Dazu kommen Erwartungswerte der pairwise precision und des pairwise recall zum Einsatz, die im Folgenden näher definiert werden. Unter der Voraussetzung, dass eine Retrieval-Methode vorhanden ist, die die Wahrscheinlichkeit der Relevanz $P(rel|q,d)$ eines gegebenen Anfrage-Dokumenten-Paares (q,d) schätzen kann, ist es möglich für jedes Dokumentenpaar die Wahrscheinlichkeit zu schätzen, dass beide Dokumente relevant sind, wobei zusätzlich angenommen wird, dass die Relevanzwahrscheinlichkeiten der beiden Dokumente voneinander unabhängig sind. Mithilfe der Wahrscheinlichkeiten, dass zwei Dokumente beide relevant sind, kann dann durch Betrachtung aller Dokumentenpaare die Anzahl der relevanten Dokumentenpaare in einem Cluster abgeschätzt werden.

Um nachfolgend die *expected precision* und den *expected recall* definieren zu können, wird zunächst eine Definition der *expected cluster precision* (ecp) benötigt:

$$\sigma(C) = \frac{1}{c(c-1)} \sum_{\substack{(d_l,d_m)\in C\times C \\ d_l \neq d_m}} \sum_{q_k \in Q} P(rel|q_k,d_l)P(rel|q_k,d_m) \tag{4}$$

Sollte ein Cluster nur ein einziges Element enthalten, so habe die expected cluster precision $\sigma(C)$ für diesen Cluster per Definition den Wert 0.

Mithilfe der Definition der expected cluster precision kann mit den nun folgenden Definitionen der *expected precision* und des *expected recall* die Qualität eines Clusterings geschätzt werden. Die expected precision stellt dabei den gewichteten Durchschnitt der ecp-Werte der Cluster dar. Gewichtet deshalb, weil die Größe der jeweiligen Cluster miteinbezogen wird. Für die nötigen Herleitungsschritte sei an dieser Stelle auf [2] verwiesen, die fertige Definition lautet wie folgt:

$$\pi(D,Q,\mathcal{C}) = \frac{1}{|D|} \sum_{C_i \in \mathcal{C}} c_i \sigma(C_i) \tag{5}$$

Um eine voreingenommene Schätzung des Recall zu vermeiden und auch, weil der Nenner für eine gegebene Anfragensammlung konstant bleibt, reicht zur Bestimmung des expected recall im Folgenden lediglich der Zähler aus. Dies ist deshalb möglich, da die Qualität der verschiedenen Clusterings für dieselbe Anfragen- und Dokumentensammlung verglichen wird. Der Zähler und damit die Definition des expected recall sieht dabei folgendermaßen aus:

$$\rho(D, Q, \mathcal{C}) = \sum_{C_i \in \mathcal{C}} c_i(c_i - 1)\sigma(C_i) \tag{6}$$

Wie bereits im vorherigen Kapitel, so lassen sich auch hier die beiden vorgestellten Qualitätsmaße zu einem einzigen Maß, dem *expected F-measure* kombinieren. Dies erfolgt wie zuvor durch den harmonischen Mittelwert:

$$eF(D, Q, \mathcal{C}) = \frac{2}{\frac{1}{\pi(D,Q,\mathcal{C})} + \frac{1}{\rho(D,Q,\mathcal{C})}} \tag{7}$$

Nachdem das perfekte Clustering bereits definiert wurde, ist es nun auch möglich, anhand der soeben vorgestellten Qualitätsmaße, eine analoge Definition für das optimale Clustering zu finden. Das Ersetzen der externen Relevanzbeurteilung durch Schätzung der Relevanzwahrscheinlichkeit stellt dabei den Hauptunterschied zwischen den beiden Definitionen dar:

Ein Clustering $\mathcal{C}$ ist ein optimales Clustering, wenn für eine Dokumenten- und Anfragensammlung, sowie dazugehöriger Retrievalfunktion zur Schätzung der Wahrscheinlichkeiten der Relevanz von Anfrage-Dokumenten-Paaren, kein anderes Clustering $\mathcal{C}'$ existiert, das entweder eine höhere expected precision und einen mindestens so hohen expected recall wie $\mathcal{C}$ aufweist, oder aber eine mindestens so hohe expected precision und einen höheren expected recall.

Eine einfache, wenn auch nicht effiziente, Möglichkeit ein solches optimales Clustering zu finden, wäre beispielsweise ein Brute-Force Clustering-Algorithmus. Dabei würden für eine gegebene Anfragen- und Dokumentensammlung, sowie einer probabilistischen Retrievalfunktion, einfach alle möglichen Clusterings erzeugt und für diese jeweils die expected precision und der expected recall berechnet und dann diejenigen Clusterings ausgewählt, die die Definition des optimalen Clusterings erfüllen.

6 Bestandteile des Optimum Clustering Frameworks

Um zu zeigen, wie die, für eine Bewertung von Clusterings mittels OCF, notwendigen Bestandteile auszusehen haben, ist es zunächst nötig festzulegen, um welche Bestandteile es sich dabei überhaupt handelt. Nach Meinung von Fuhr et al. [2] beruhen sämtliche Methoden zum Dokumenten-Clustering auf den folgenden drei Komponenten:

1. einer Anfragensammlung,
2. einer Retrievalfunktion, und
3. einem Ähnlichkeitsmaß für Dokumente.

Die meisten der bereits bestehenden Clustering-Methoden basieren auf Heuristiken und besitzen keine theoretische Grundlage, mit der die Wahl der drei aufgezählten Bestandteile mit der resultierenden Qualität des Clusterings in

Verbindung gebracht werden könnte. Mithilfe des Optimum Clustering Frameworks, das nahezu alle der bestehenden Methoden des Dokumenten-Clusterings abdeckt, soll dank der theoretischen Grundlage eine gezieltere Entwicklung besserer Clustering-Methoden angeregt werden.

Bei der Wahl der drei Bestandteile einer Clustering-Methode sollte stets darauf geachtet werden, diese so zu wählen, dass sie auch zum zugrundeliegenden theoretischen Modell passen. Während es beim OCF dabei für die Anfragensammlung kaum Einschränkungen und damit den größten Entscheidungsfreiraum gibt, ist die Wahl bei der Definition der Retrievalfunktion schon deutlich eingeschränkt, wohingegen bei der Ähnlichkeitsfunktion sogar gar keine Wahlfreiheit vorhanden ist. Jeder der drei Bestandteile wird im Folgenden nun näher erläutert.

6.1 Anfragensammlung

Ziel bei der Erstellung einer Anfragensammlung sollte es immer sein, Anfragen zu finden, die das aktuelle Informationsbedürfnis des Benutzers am besten widerspiegeln. Für das Clustering von Dokumenten kann man 3 Arten der Erstellung von Anfragensammlungen unterscheiden: lokal, global und extern.

Die lokalen und globalen Methoden erstellen Anfragensammlungen, indem sie die zugehörigen Dokumentensammlungen analysieren, wohingegen die externen Methoden meist auf die Miteinbeziehung von Domänenwissen setzen.

Ein einfaches Beispiel für ein lokale Methode wäre etwa, jedes Wort innerhalb der Dokumentensammlung als eigene Anfrage zu verwenden. Für eine etwas gezieltere Anfrage wäre es auch möglich Schlüsselwörter innerhalb der Dokumentensammlung ausfindig zu machen und diese als Anfragen zu benutzen. Aufgrund ihrer Einfachheit sind die lokalen Methoden auch die Art, die bei Clustering-Methoden am öftesten eingesetzt wird.

Beim globalen Ansatz hingegen werden eher globale Eigenschaften der Dokumentensammlung, wie beispielsweise strukturelle oder thematische Informationen, verwendet um eine entsprechende Anfragensammlung zu erzeugen.

Am effektivsten ist jedoch die externe Methode der Erstellung von Anfragensammlungen. Die Erstellung von Anfragen nach der externe Methode basiert ausschließlich auf externem Wissen. Eine Möglichkeit stellt dabei zum Beispiel eine manuelle Beurteilung der Relevanz oder etwa das Feedback von Benutzern dar. Weiterhin wäre es auch möglich eine externe Dokumentensammlung wie Wikipedia zu verwenden, wie es von dem ESA-Modell (explicit semantic analysis) von Gabrilovich und Markovitch [8] praktiziert wird.

Unabhängig von den Methoden zur Erstellung von Anfragesammlungen, bestände auch die Möglichkeit mehrere Anfragensammlungen zu verwenden, um so mehrere Clusterings zu erzeugen, aus denen der Benutzer dann wählen kann.

Durch mehrere Clusterings könnten beispielsweise verschiedene Dimensionen beziehungsweise Facetten dargestellt werden. Diese ließen sich unter anderem durch die Struktur der Dokumente definieren, so könnten zum Beispiel Emails abhängig vom Betreff, dem Sender oder dem Datum in verschiedene Cluster eingeteilt werden.

6.2 Probabilistische Retrievalfunktion

Prinzipiell besteht eine relativ große Auswahl an Retrievalfunktionen, die eingesetzt werden können. Es kommt jedoch auch öfters vor, dass Entscheidungen, die bei der Wahl der Erstellung der Anfragensammlung getroffen wurden, vorschreiben, welche spezifische Retrievalfunktion verwendet werden muss.

Ein Problem stellt jedoch das Schätzen der tatsächlichen Relevanzwahrscheinlichkeit dar, da die meisten Retrieval-Methoden eine Bewertungszahl berechnen, die lediglich rang-equivalent zum probabilistischen Retrieval ist und meist bestimmte anfragenspezifische Eigenschaften bei der Berechnung ignoriert werden. Da das OCF aber auf dem Vergleich der Relevanzwahrscheinlichkeiten verschiedener Anfragen beruht, müssen die Bewertungszahlen der entsprechenden Retrieval-Methoden, sofern sie keine direkte Schätzung der Relevanzwahrscheinlichkeit darstellen, erst noch in eine entsprechende Wahrscheinlichkeit umgewandelt werden.

6.3 Ähnlichkeitsmaß für Dokumente

Wie bereits angedeutet, gibt es für beim Ähnlichkeitsmaß für Dokumente keinerlei Wahlfreiheit. Das in der Herleitung zu Gleichung (5) (nachzulesen in [2]) bereits eingeführte Skalarprodukt der Vektoren $\tau(d)$, das die erwartete Anzahl an Anfragen liefert, für die beide Dokumente relevant sind, stellt das einzig zulässige Maß für die Ähnlichkeit von Dokumenten dar.

7 Hierarchische Clustering-Methoden im OCF

Generell lässt sich sagen, dass sowohl bei den agglomerativen als auch bei den divisiven Methoden des hierarchischen Clusterings, diejenigen Methoden am besten für das Optimum Clustering Framework geeignet sind, die nach der Fusion bzw. Division der Cluster die Clusterqualität erneut untersuchen. Im OCF können diese Methoden einfach umgesetzt werden, indem als Qualitätsmaß die expected cluster precision und/oder der expected recall verwendet wird.

Im Falle einer agglomerativen Clustering-Methode würde jeder Schritt dieser Methode zu einem Cluster mit einem mindestens so großen expected recall führen, wie die Cluster aus denen er zusammengesetzt wurde, während die expected precision bei jedem Schritt zwangsläufig abnimmt.

Bei divisiven Methoden verhält es sich entgegengesetzt. Hier wird mit einem einzigen Cluster mit maximalem expected recall aber eher geringer expected precision gestartet und dann in jedem Schritt die Cluster weiter aufgeteilt, um so die expected precision zu erhöhen, während der expected recall dabei nur möglichst minimal sinken soll.

Obwohl hierarchische Clustering-Methoden prinzipiell nicht in der Lage sind alle Pareto Optima zu finden, so gibt es doch Verfahren wie den min-cut-Algorithmus, mit denen zumindest ein einzelnes Optimum gefunden werden kann. Ein Begründung, warum dies mit dem min-cut-Algorithmus möglich ist, kann in [2] nachgelesen werden.

Für agglomerative Verfahren kann die Fähigkeit immer ein Pareto Optimum zu finden leider nicht nachgewiesen werden, da per Definition der pairwise precision, Cluster mit nur einem Element stets eine pairwise precision von 0 besitzen und somit mehrere Schritte notwendig sind bevor ein Optimum gefunden werden kann, was die Suche sehr schwierig macht.

Festhalten lässt sich jedoch, dass die Qualitätsmaße pairwise precision und pairwise recall gut dazu eingesetzt werden können, bestehenden Ähnlichkeitsmaßen oder Fusionsregeln von Clustern im Nachhinein eine theoretische Grundlage zu geben.

8 Zusammenfassung

Zusammenfassend lässt sich sagen, dass es mit Entwicklung des Optimum Clustering Frameworks erfolgreich gelungen ist, ein Rahmenwerk zu schaffen, das durch seine theoretische Grundlage aufzeigt, wie, abhängig von einer Anfragensammlung und einer probabilistischen Retrievalfunktion, eine optimale Clusteringqualität erzielt werden kann. Der Fokus lag dabei ganz klar auf der theoretischen Grundlage, da bisherige Clustering-Methoden zwar auch auf den drei Bestandteilen (Anfragensammlung, Retrievalfunktion und Ähnlichkeitsmaß für Dokumente) aufbauen, jedoch oftmals Heuristiken im Zusammenhang mit diesen verwenden. Dies konnte dadurch erreicht werden, dass geeignete Metriken für die Qualität eines Clusterings eingeführt wurden, die es ermöglichten optimales Clustering unter Beachtung der Abschätzungen der Relevanzwahrscheinlichkeit der jeweiligen Anfrage-Dokumenten-Paare zu definieren.
Ziel der nun vorhandenen theoretischen Grundlage soll es sein, eine Umgebung zu schaffen, in der eine gezieltere Entwicklung hinsichtlich besserer Methoden des Clusterings von Dokumenten möglich ist.

Literatur

[1] Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/Gather: A cluster-based approach to browsing large document collections. In: Proceedings

of the 15th Annual International ACM SIGIR Conference on Research an Development in Information Retrieval, ACM (1992) 318–329
[2] Fuhr, N., Lechtenfeld, M., Stein, B., Gollub, T.: The optimum clustering framework: implementing the cluster hypothesis. Information Retrieval (2011) 1–23
[3] SwissEduc: Informationsbeschaffung im Internet. `http://www.swisseduc.ch/informatik/internet/internet_recherche/informationsbeschaffung_im_internet/Relevanz_Dokuments.html` Letzter Zugriff: 13.06.2012.
[4] Van Rijsbergen, C.J.: Information Retrieval. 2nd edn. Butterworths (1979)
[5] Robertson, S.E.: The probability ranking principle in IR. Journal of Documentation **33** (1977) 294–304
[6] Jardine, N., van Rijsbergen, C.J.: The use of hierarchic clustering in information retrieval. Information Storage and Retrieval **7**(5) (1971) 217 – 240
[7] Ackerman, M., Ben-David, S.: Measures of clustering quality: A working set of axioms. In: Proceedings NIPS 200, MIT Press (2008) 121–128
[8] Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of IJCAI'07, Morgan Kaufmann Publishers (2007) 1606–1611