Christian Wartena und Koraljka Golub

Evaluierung von Verschlagwortung im Kontext des Information Retrievals

1 Einleitung

Dieser Beitrag möchte einen Überblick über die in der Literatur diskutierten Möglichkeiten, Herausforderungen und Grenzen geben, Retrieval als eine extrinsische Evaluierungsmethode für die Ergebnisse verbaler Sacherschließung zu nutzen. Die inhaltliche Erschließung im Allgemeinen und die Verschlagwortung im Besonderen können intrinsisch oder extrinsisch evaluiert werden. Die intrinsische Evaluierung bezieht sich auf Eigenschaften der Erschließung, von denen vermutet wird, dass sie geeignete Indikatoren für die Qualität der Erschließung sind, wie formale Einheitlichkeit (im Hinblick auf die Anzahl zugewiesener Deskriptoren pro Dokument, auf die Granularität usw.), Konsistenz oder Übereinstimmung der Ergebnisse verschiedener Erschließer:innen. Bei einer extrinsischen Evaluierung geht es darum, die Qualität der gewählten Deskriptoren daran zu messen, wie gut sie sich tatsächlich bei der Suche bewähren. Obwohl die extrinsische Evaluierung direktere Auskunft darüber gibt, ob die Erschließung ihren Zweck erfüllt, und daher den Vorzug verdienen sollte, ist sie kompliziert und oft problematisch. In einem Retrievalsystem greifen verschiedene Algorithmen und Datenquellen in vielschichtiger Weise ineinander und interagieren bei der Evaluierung darüber hinaus noch mit Nutzer:innen und Rechercheaufgaben. Die Evaluierung einer Komponente im System kann nicht einfach dadurch vorgenommen werden, dass man sie austauscht und mit einer anderen Komponente vergleicht, da die gleiche Ressource oder der gleiche Algorithmus sich in unterschiedlichen Umgebungen unterschiedlich verhalten kann. Wir werden relevante Evaluierungsansätze vorstellen und diskutieren, und zum Abschluss einige Empfehlungen für die Evaluierung von Verschlagwortung im Kontext von Retrieval geben.

Der Beitrag ist folgendermaßen aufgebaut: Zunächst schauen wir uns an, was der Zweck von Verschlagwortung ist und wie die Qualität von vergebenen Stich- und Schlagwörtern definiert werden kann. In Abschnitt 3 behandeln wir die Frage, inwieweit es möglich ist, die Qualität der Verschlagwortung indirekt zu bestimmen, also aufgrund von Suchergebnissen, die sich auf die Sacherschließung stützen. Dabei konzentrieren wir uns insbesondere auf Faktoren, die mit den Deskriptoren interagieren und die Auswirkungen auf die Ergebnisse haben könnten. In Abschnitt 4 stellen wir eine Auswahl von Studien vor, die

Open Access. © 2021 Christian Wartena und Koraljka Golub, publiziert von De Gruyter. CC BY Dieses Werk ist lizenziert unter der Creative Commons Attribution 4.0 Lizenz.
https://doi.org/10.1515/9783110691597-016

tatsächlich Retrievalszenarien für die Evaluierung von Sacherschließung nutzen. Wir schließen das Kapitel mit einigen allgemeinen Empfehlungen zur extrinsischen Evaluierung von Sacherschließung ab.

2 Schlagwörter

Inhaltsbeschreibende Terme können entweder direkt aus dem zu beschreibenden Text oder aus einem Erschließungsvokabular, das für die thematische Beschreibung von Dokumenten entwickelt worden ist, entnommen werden. Im ersten Fall werden die gewählten Terme üblicherweise Stichwörter genannt; im zweiten Fall sprechen wir einerseits von (kontrollierter) Verschlagwortung, wenn Wörter oder Wortgruppen aus einer natürlichen Sprache verwendet werden, die in einem Thesaurus oder in einer Normdatei organisiert sind, und andererseits von Klassierung,[1] wenn Notationen oder Klassen aus einem Klassifikationssystem genutzt werden. Zwischen Stichwörtern und kontrollierten Schlagwörtern können wir noch die freien Schlagwörter einordnen, die die gleiche Funktion wie kontrollierte Schlagwörter haben, aber weder aus dem zu erschließenden Text noch aus einem Vokabular entnommen werden müssen. Da (kontrollierte) Schlagwörter auch normale (deutsche oder anderssprachige) Wörter sind, ist es häufig der Fall, dass viele Schlagwörter genau wie Stichwörter verbatim im Text vorkommen. Da das Vorkommen im Dokument jedenfalls für längere Texte ein wichtiger Indikator für die Relevanz eines Begriffes ist, wird bei der maschinellen Verschlagwortung die Extraktion aus dem Text fast immer als Ausgangspunkt genommen. Andererseits können die Notationen in einem Klassifikationssystem auch natürlichsprachige Begriffe sein oder mit solchen verknüpft sein. Insbesondere wenn ein (hierarchisches) Klassifikationssystem sehr feingliedrig unterteilt ist, verschwimmen die Grenzen zwischen Klassierung und Verschlagwortung. Wenn dagegen eine grobkörnige Klassifikation oder die oberen Ebenen einer feingliedrigen Systematik verwendet werden, unterscheidet sich die Klassierung erheblich von der Verschlagwortung. Schließlich müssen wir noch die Volltextindexierung unterscheiden, die sämtliche inhaltstragenden Wörter eines Textes verzeichnet und die für die Volltextsuche gebraucht wird. Die Relevanz eines Begriffes für einen Text wird dann später in Abhängigkeit von der jeweiligen Abfrage bestimmt. Hierfür werden zum

1 In Anlehnung am englischen Wort *Classification* spricht man, vor allem in der Informatik, hier auch häufig von Klassifikation oder Klassifizieren.

Beispiel probabilistische Modelle genutzt, die wiederum Relevanzwahrscheinlichkeiten nutzen, um ein Ranking der Suchergebnisse zu ermöglichen.

Der Hauptzweck der Verschlagwortung ist es, das Auffinden eines Dokuments aus verschiedenen Perspektiven zu ermöglichen. Typischerweise werden drei bis zwanzig einzelne oder teilweise präkombinierte Schlagwörter pro Dokument vergeben. Ziel des Klassierens (also der Erschließung mit Notationen, die Klassifikationssystemen entnommen werden) ist die Gruppierung ähnlicher Dokumente, um deren Sichtung zu unterstützen: In der analogen Welt bedeutet das eine Sichtung von in Bibliotheksregalen aufgestelltem Bestand; in einer Online-Umgebung werden Trefferlisten gesichtet. Die Erschließung beschränkt sich beim Klassieren auf wenige thematische Aspekte in starker Präkombination, typischerweise auf eine einzige Klasse (siehe bezüglich der Ähnlichkeiten zwischen verbaler Erschließung und Klassierung auch Lancaster 2003: 20–21).

Schlagwortbasiertes Retrieval wird vor allem in sogenannten *Online Public Access Catalogs* (OPACs) in Bibliotheken angewandt, um Bücher und andere Medien zu finden, für die oft kein Volltext zur Verfügung steht, so dass die Suche vollständig auf das Vorhandensein von Metadaten angewiesen ist. Im Folgenden werden wir uns hauptsächlich auf schlagwortbasiertes Retrieval und OPACs beziehen, aber immer wieder auf die Volltextsuche schauen, um die hierfür entwickelten Methoden zu verstehen.

Verschlagwortung kann intellektuell[2] von Menschen ausgeführt werden, in automatischen Verfahren erfolgen oder in einem hybriden Prozess, bei dem entweder das menschliche Erschließen softwaregestützt erfolgt oder ein automatisierter Erschließungsprozess intellektuell gesteuert und korrigiert wird. Wir sind der Meinung, dass an jede Verschlagwortung dieselben Maßstäbe angelegt werden sollten, unabhängig davon, wie sie generiert wurde. Daher werden wir im Folgenden prinzipiell keinen Unterschied zwischen intellektueller und automatisch generierter Erschließung machen.

2.1 Sacherschließungsqualität

Die Sacherschließungsnorm ISO 5963:1985 (International Organization for Standardization 1985) gibt eine dreistufige dokumentorientierte Definition für die intellektuelle Sacherschließung: 1) thematische Erfassung des Dokuments, 2) begriffliche Analyse, um zu entscheiden, welche Aspekte des Inhalts dargestellt werden sollen, 3) Übersetzung dieser Begriffe oder Aspekte in ein kontrolliertes

2 In der Informatik und im Information Retrieval ist der Term *intellektuell* unüblich. Stattdessen wird das Adjektiv *manuell* verwendet.

Vokabular. Folgerichtig wird die Qualität der Verschlagwortung und der Sacherschließung im Allgemeinen in erster Linie über die Qualität des Erschließungsprozesses definiert. Der erste und zweite Schritt der ISO-Definition haben mit *aboutness* oder Thematik zu tun, also mit der Frage, was der Gegenstand oder die Themen des zu erschließenden Dokuments sind. Über *aboutness* und die Schwierigkeiten geeigneter Sacherschließung wurde bereits ausführlich diskutiert (für einen Überblick siehe Lancaster 2003: 13–19). Die Bestimmung der relevanten Aspekte eines Textes ist keine eindeutig definierte Aufgabe, da Texte ein komplexes kognitives und soziales Phänomen darstellen, und das kognitive Verstehen von Texten viele Wissensquellen einbezieht, verschiedene Schlussfolgerungen zulässt und mit individueller Interpretation einhergeht (Moens 2000: 7–10). Verschiedene Studien haben gezeigt, wie unterschiedlich Texte interpretiert werden können. Z. B. führte Morris (2010) ein Experiment durch, bei dem 26 Teilnehmer:innen drei Texte anhand lexikalischer Ketten (Gruppen semantisch verwandter Wörter) interpretieren sollten: Die Interpretationsergebnisse unterschieden sich um etwa 40 %.

Die zweite und dritte Stufe der ISO-Norm basieren auf spezifischen Vorentscheidungen bezüglich der Art des Bestands und der anvisierten Nutzer:innen (Schüler:innen, Lai:innen, Spezialist:innen usw.) z. B. im Hinblick auf Abdeckung (ein Erschließungsvorgang mit einer niedrigen Relevanzschwelle bei der Auswahl der Begriffe führt zu einer hohen Abdeckung) und Spezifität (die hierarchische Ebene, auf der erschlossen wird).

Anders als bei der dokumentorientierten Erschließung stehen bei der anfrageorientierten (problemorientierten, nutzungsorientierten) Erschließung (Brenner und Moers 1958; Soergel 1985; Fidel 1994) die potentiellen Nutzer:innen und Anwendungen im Vordergrund; die Aufgabe der Erschließer:innen ist es, den Text zu verstehen und dann vorherzusehen, für welche Themen oder Anwendungen dieser Text relevant werden könnte. Die abfrageorientierte Erschließung berücksichtigt also nicht nur die *aboutness*, sondern auch die erwartete Relevanz eines Texts im Hinblick auf Themen, Zwecke, Aufgaben und zu lösende Probleme, wodurch weitere Kontextabhängigkeit und Vielfalt ins Spiel kommen. Es mag für einzelne Erschließer:innen unmöglich sein, alle Ideen und Bedeutungen, die mit einem Text assoziiert werden könnten, zu identifizieren, da es immer weitere Ideen und Bedeutungen geben wird, die verschiedene Personen zu verschiedenen Zeiten und an verschiedenen Orten in dem Text finden können (vgl. z. B. Mai 2001: 606). Trotzdem sollten Erschließer:innen versuchen, sich in die Bedürfnisse der Zielgruppe hineinzudenken.

2.2 Evaluierung von Verschlagwortung

Da der (Haupt-)Zweck von Schlagwörtern die Unterstützung des Retrievals ist, liegt es nahe, die ausgewählten Terme indirekt zu bewerten, also nicht die Schlagwörter selbst zum Gegenstand der Betrachtung zu machen, sondern den extrinsischen Ansatz zu wählen und den Nutzen der vorhandenen Erschließung für die Suche in einem Retrievalsystem zu untersuchen. Wie wir im nächsten Abschnitt erörtern werden, ist dies nicht unproblematisch und nicht einfach zu bewerkstelligen. Beliebter sind daher die intrinsischen Methoden der Erschließungsevaluierung.

Eine der intrinsischen Methoden, die gern für die Evaluierung maschineller Verschlagwortung verwendet wird, ist der Vergleich von maschinell und intellektuell vergebenen Termen. Bei einem solchen Vergleich sollten aber immer die für den jeweiligen Bestand bestehenden Erschließungsvorgaben berücksichtigt werden. Ein Schlagwort, das in einer Umgebung korrekt ist, in der ein hoher Vollständigkeitsgrad erwartet wird, kann in einem System mit einer hohen Relevanzschwelle und niedrigerer Abdeckungsvollständigkeit fehl am Platz sein (Soergel 1994). Bei diesem Vergleich muss bedacht werden, dass auch eine intellektuelle Verschlagwortung fehlerhaft sein kann. Lancaster (2003: 86–87) listet folgende Arten von Erschließungsfehlern auf: Fehler im Zusammenhang mit dem angestrebten Abdeckungsgrad (wenn zu viele oder wenige Deskriptoren ausgewählt wurden), Fehler im Zusammenhang mit der Spezifität (was in der Regel bedeutet, dass der ausgewählte Deskriptor nicht der spezifischste verfügbare ist), Auslassung wichtiger Deskriptoren und die Zuweisung offensichtlich falscher Deskriptoren.

Zudem berichtet u. a. Hjørland (2016), dass verschiedene Personen, seien es Nutzer:innen oder professionelle Sacherschließer:innen, ein und denselben Text thematisch unterschiedlich einordnen. Ein Grund dafür sind Unterschiede in der Herangehensweise, die entweder von der rationalistischen Vorstellung ausgehen kann, es gebe genau eine korrekte Erschließung für einen Text, oder eben von der pragmatischen Vorstellung, dass für verschiedene Zwecke und Nutzer:innen jeweils eine unterschiedliche Erschließung erforderlich sein könne (Hjørland 2018). Daher können in einem Bestand vorgefundene Metadatensätze nicht ohne Weiteres „den Goldstandard“ darstellen: Schlagwörter, die von einer Maschine vergeben wurden, aber nicht von einem Menschen, können falsch sein oder aber bei der intellektuellen Erschließung versehentlich oder aufgrund bestimmter Erschließungsziele ausgelassen worden sein. Als eine Möglichkeit, dieses Problem zumindest teilweise zu überwinden, schlagen Gazendam et al. (2009) eine weniger strikte Verwendung des Goldstandards bei

der Bewertung automatisch vergebener Schlagwörter vor, bei der keine exakte Übereinstimmung, sondern eine semantische Nähe zu intellektuell ausgewählten Termen (gemessen an der Anzahl und Art der Relationen zwischen zwei Deskriptoren in einem Thesaurus) gefordert wird.

Ein weiterer häufig verwendeter Indikator für die Erschließungsqualität ist die Konsistenz zwischen Arbeitsergebnissen verschiedener Erschließer:innen (*inter-indexer consistency*) oder des- bzw. derselben Erschließer:in zu verschiedenen Zeiten (*intra-indexer consistency*) (für einen Überblick siehe Lancaster 2003: 68–82). Die Konsistenz kann entweder auf der Ebene der Begriffe oder auf der Ebene der Terme zur Bezeichnung der Begriffe gemessen werden. Markey (1984) vergleicht 57 Untersuchungen zur Erschließungskonsistenz und berichtet, dass der Übereinstimmungsgrad zwischen 4 % und 84 % liegt, wobei nur 18 Untersuchungen eine Übereinstimmung von über 50 % aufweisen. Es scheint dabei zwei Haupteinflussfaktoren zu geben: 1) je höher die Abdeckung und Spezifität der Erschließung ist, desto geringer wird die Konsistenz; das heißt, die Erschließer:innen wählen denselben ersten Deskriptor für das Hauptthema des Texts, aber die Übereinstimmung nimmt ab, je mehr Deskriptoren sie wählen; 2) je umfangreicher das Vokabular (je mehr Auswahlmöglichkeiten also die Erschließer:innen haben), desto geringer ist die Wahrscheinlichkeit, dass sie dieselben Deskriptoren wählen (Olson und Boll 2001: 99–101).

Lancaster (2003: 71) fügt weitere mögliche Faktoren hinzu: Sacherschließung auf der Grundlage unkontrollierten versus kontrollierten Vokabulars, Spezifität des Vokabulars, Merkmale des Fachgebiets und seiner Terminologie, individuelle Faktoren (z. B. Erfahrungsniveau), den Erschließer:innen zur Verfügung stehende Werkzeuge und der Umfang des zu erschließenden Dokumentes. Auch kann eine Erschließung sowohl konsistent schlecht als auch konsistent gut sein (Cooper 1969). Das bedeutet, dass eine hohe Konsistenz nicht unbedingt ein Zeichen für eine hohe Erschließungsqualität ist (Lancaster 2003: 91). Hohe Konsistenz ist eine notwendige, aber keine hinreichende Bedingung für Korrektheit, und daher sollte die Konsistenz nicht als Hauptindikator für Erschließungskorrektheit angesehen werden (Soergel 1994). Rolling (1981) zeigt außerdem, dass sich die Werte für Erschließungseinheitlichkeit, -qualität und -effektivität nicht unbedingt proportional zueinander verhalten. Er definiert Erschließungsqualität über die Frage, ob der Informationsgehalt eines erschlossenen Dokuments korrekt dargestellt wird (dokumentorientierte Definition), und den Erschließungserfolg über die Frage, ob ein erschlossenes Dokument jedes Mal, wenn es für eine Suche relevant ist, korrekt abgerufen wird (abfrageorientierte Definition). Trotz allem wird die Erschließungskonsistenz häufig als Maßstab für die Qualität eines Goldstandards und als Maßstab für die Erschlie-

ßungsqualität in operativen Systemen ohne Bezug auf einen Goldstandard verwendet.

3 Extrinsische Evaluierung von Schlagwörtern

Wie bereits erwähnt, kann und sollte die Qualität verbaler Sacherschließung im Kontext eines Retrievalsystems untersucht werden, vorzugsweise unter Einbeziehung realer Suchanfragen und realer Nutzer:innen (siehe Lancaster 2003: 99). Auch die oben erwähnte ISO-Norm 5963 empfiehlt, die Qualität der Erschließung durch die Analyse von Suchergebnissen zu prüfen.

Das Testen eines Retrievalsystems weist viele Probleme auf, die Ergebnisse hängen von vielen Faktoren ab und die Qualität der Indexierung kann daher nicht isoliert betrachtet werden. Wenn ein System bei einer Suche nicht die relevanten Dokumente auswirft, kann das Problem bei den gewählten Indextermen, beim Retrievalsystem selbst oder bei der Interaktion zwischen Index und Retrievalalgorithmus liegen, worin verschiedene Annahmen über die Eigenschaften des Index einfließen, und der möglicherweise an bestimmte Erschließungsvorgaben angepasst werden muss.

Soergel (1994) legt eine logische Analyse der Auswirkungen von Erschließungsvorgaben auf den Retrievalerfolg vor. Er identifiziert Erschließungswerkzeuge, *viewpoint*-basierte und *importance*-basierte Erschließungsabdeckung, Spezifität, Korrektheit und Konsistenz als Faktoren, die sich auf das Retrieval auswirken. Er kommt zu dem Schluss, dass der Retrievalerfolg hauptsächlich davon abhängt, wie gut die Erschließung mit den Anforderungen der jeweiligen Suche zusammenpasst und wie gut bei der Abfrageformulierung die Merkmale des Retrievalsystems beachtet werden. Diese Komplexität muss beim Entwurf und bei der Evaluierung von Retrievalsystemen berücksichtigt werden, denn:

> indexing characteristics and their effects on retrieval are so complex that they largely defy study in artificial test situations. Most experiments fail to account for important interactions among factors as they occur in the real world, and thus give results that mislead more than they enlighten, results that have little meaning for the assessment or improvement of operational retrieval systems (Soergel 1994: 589).

In den folgenden Abschnitten werden wir einige Aspekte der Evaluierung von Retrievalsystemen diskutieren, insbesondere wenn sie für die extrinsische Evaluierung von Schlagwörtern relevant erscheinen.

3.1 Relevanz im Kontext des Retrievals

Es gibt viele mögliche Beziehungen zwischen einem Dokument und einer Suchanfrage – was Relevanz zu einer komplexen Sache macht. Die Relevanz eines Dokuments für ein Suchwort wird oft durch ein menschliches Urteil bestimmt, wodurch eine gewisse Subjektivität mitspielt. Borlund (2003) betont, dass Relevanz mehrdimensional und dynamisch ist: Nutzer:innen legen bei der Beurteilung von Relevanz viele verschiedene Kriterien an, und die Wahrnehmung von Relevanz kann sich bei derselben Person im Laufe der Zeit ändern. Es gibt verschiedene Klassen, Typen, Grade und Ebenen von Relevanz. Saracevic (2007a, 2007b) und Huang und Soergel (2013) geben einen kritischen Überblick über das Wesen, die Eigenschaften und Erscheinungsformen von sowie über Theorien und Modelle für Relevanz.

Trotz der dynamischen und mehrdimensionalen Natur von Relevanz geht die Evaluierung von Systemen für das Information Retrieval in der Praxis oft nicht über die Verwendung bereits vorhandener Relevanzurteile hinaus. Die Cranfield-Tests (Cleverdon, Mills und Keen 1968) haben die vorherrschende Evaluierungsmethodik für das Information Retrieval etabliert, bei der ein Goldstandard eingesetzt wird: ein Testkorpus, das aus einer Menge von Dokumenten, einer Menge von *Topics* oder Anfragen und einer Menge von Relevanzbewertungen besteht (Buckley und Voorhees 2000). Ein *Topic* ist eine Beschreibung der jeweils gesuchten Information. Relevanzbewertungen geben an, welche Dokumente als Treffer zu jedem *Topic* ausgefunden werden sollen; sie sind in der Regel binär und enthalten keine Angaben zu den verschiedenen Graden der Relevanz. Man kann zwar sagen, ob A für B relevant ist, aber es ist viel schwerer, genau anzugeben, auf welche Weise A für B relevant ist, und herauszufinden, wie sich eine Information in die Gesamtstruktur eines Themas einfügt und welchen Beitrag sie zum Denken und Argumentieren der Nutzerin bzw. des Nutzers über das Thema leisten kann. Daher plädieren Huang und Soergel (2013) dafür, bei der Konzeptualisierung von Relevanz den Fokus von Entitäten hin zu Relationen zu verschieben. Sie fordern, die Erforschung verschiedener Arten von Relevanzbeziehungen mit einer hohen Priorität zu belegen, um die Tiefe und den Reichtum des Relevanzbegriffs voll zu durchdringen.

3.2 Einfluss von Suchszenarien und -anfragen

Suchszenarien und daraus resultierende Suchaufgaben beeinflussen sowohl Suchstrategien als auch Relevanzbewertungen. Beispielsweise würden die

Treffer einer Suche nach *Depression* von einer Ärztin, die sich über die neuesten Forschungsergebnisse zu Depressionen informieren möchte, und von einer medizinischen Forscherin, die einen umfassenden Übersichtsartikel über alle Aspekte der Depression vorbereitet, unterschiedlich bewertet werden. Daher müssen Suchszenarien und Suchaufgaben bei Retrievaluntersuchungen berücksichtigt werden (Liu und Belkin 2015).

Suchaufgaben werden üblicherweise nach ihrer Komplexität eingeteilt (vgl. z. B. Kim und Soergel 2005, und Belkin et al. 2014). Ingwersen und Järvelin (2005: 327) ordnen Arbeits- und Suchszenarien in ein Kontinuum von natürlich bis künstlich ein und verwenden sechs Dimensionen für die Beschreibung von Abfragen. Die Auswahl der Suchaufgaben spielt offensichtlich eine Rolle bei der Bewertung des Systems. Iivonen (1995) stellte fest, dass sich die Konsistenz der Ergebnisse zwischen verschiedenen Nutzer:innen und zwischen vier verschiedenen Arten von Suchaufgaben bei einem bzw. einer Nutzer:in signifikant unterschied. Kim (2008) zeigt, dass es erhebliche Auswirkungen auf das Suchverhalten hat, ob es um eine spezifische oder um eine allgemeine Aufgabe geht. Liu und Belkin (2015) führen ebenfalls verschiedene Studien an, in denen gezeigt wurde, dass es einen erheblichen Einfluss auf das Suchverhalten und die Evaluierung von Retrievalsystemen hat, wie vertraut die Nutzer:innen mit einem Thema sind.

Da der Vergleich zwischen zwei Systemen je nach Suchaufgabe zu unterschiedlichen Ergebnissen führen könnte, raten Golub et al. (2016) dazu, verschiedene Klassen von Suchaufgaben und Abfragen sowie verschiedene Nutzungsgruppen zur Evaluierung zu verwenden.

3.3 Einfluss von Retrieval und Rankingalgorithmus

Ein modernes Retrievalsystem ist nicht einfach ein Programm, das eine Trefferliste aller Texte ausgibt, in denen das Suchwort vorkommt. Vielmehr wurde in den letzten Jahrzehnten eine Reihe komplexer Algorithmen entwickelt, die die Qualität des Retrievals über einen einfachen Stringabgleich hinaus verbessern. Da diese Methoden alle für das Retrieval auf Basis eines Volltextindexes entwickelt wurden, ist es nicht von vornherein klar, ob diese Techniken auch auf OPACs angewandt werden können. Für die OPAC-Suche werden dagegen traditionell eine hohe Erschließungsqualität und sorgfältig formulierte Anfragen vorausgesetzt. Dennoch wird schon seit den 1980er Jahren untersucht (siehe z. B. Fox et al. 1993) und weiterhin diskutiert (siehe z. B. Yu und Young 2004 oder Antelman, Lynema und Pace 2006), wie sich Fortschritte aus dem Bereich des Information Retrievals auch auf OPACs übertragen lassen. OPACs, die für die

Suche im Internet entwickelte Techniken verwenden, werden manchmal als *Next-Generation*-OPACs oder WebPACs bezeichnet. Auch die sogenannten Resource-Discovery-Systeme, die meistens Ergebnisse aus mehreren Quellen einschließen, nutzen typischerweise Verfahren, die für die Volltextsuche entwickelt wurden.

Einige der Techniken, die für die unstrukturierte Volltextsuche entwickelt wurden, könnten auch für das Retrieval mit einem OPAC sehr relevant sein, da die meisten erschlossenen Bestände aus verschiedenen Gründen nicht so ideal erschlossen sind, wie wir es uns wünschen würden: 1) Es gibt Erschließungsfehler; 2) viele Sammlungen enthalten Deskriptoren aus verschiedenen Quellen, die nach unterschiedlichen Erschließungsregeln aus unterschiedlichen Vokabularen zugewiesen wurden; 3) es könnten Diskrepanzen zwischen Nutzer:innen und Erschließer:innen im Hinblick auf den Blickwinkel oder den erwarteten Grad der Erschließungsgranularität bestehen. Daher könnte man versuchen, mit Retrieval-Algorithmen Erschließungsfehler oder -schwächen auszugleichen. Tudhope et al. (2006) schlagen z. B. ein anspruchsvolles Matching vor, das, wie sie meinen, Inkonsistenzen aufgrund von Diskrepanzen in der Erschließungsspezifität oder von Erschließungsfehlern ausgleichen kann. Dasselbe könnte für viele andere Algorithmen gelten.

3.3.1 Termgewichtung

Probabilistische Modelle versuchen ausgehend von einigen einfachen und plausiblen Annahmen die Wahrscheinlichkeit zu schätzen, dass ein Dokument für eine bestimmte Anfrage relevant ist. Meistens stellt sich die resultierende Formel als eine Variation der tf.idf-Gewichtung dar. Die tf.idf-Gewichtung verbindet zwei grundlegende Ideen: Erstens ist es umso wahrscheinlicher, dass ein Text für eine Abfrage mit einem bestimmten Wort relevant ist, je häufiger dieses Wort in dem Text vorkommt. Bei dieser Komponente geht es also um die Wort- oder Termfrequenz (tf). Zweitens ist es umso wahrscheinlicher, dass ein Text für eine Abfrage mit einem bestimmten Wort relevant ist, wenn dieses Wort ein spezifisches Wort ist, das nur in wenigen Texten vorkommt. Das Maß für die Spezifität eines Wortes ist die inverse Dokumentfrequenz (idf).

Die Termfrequenz versucht das Problem zu lösen, dass ein Wort in einem Text zwar vorhanden, aber nicht relevant ist: Wenn ein Wort in einem Text häufig vorkommt, ist es unwahrscheinlich, dass es nur zufällig vorhanden ist. Wenn Nutzer:innen ein Wort suchen, das in vielen Texten vorkommt, kann die Termgewichtung darüber hinaus auch hilfreich sein, um die „relevantesten“

Dokumente[3] am Anfang der Trefferliste zu platzieren und um den Nutzer:innen zu helfen, eine kleine Auswahl aller relevanten Texte zu treffen. Für Schlagwörter kann dieses Maß nicht angewandt werden, denn es gibt in der Sacherschließung keine irrelevanten Deskriptoren (mit Ausnahme von Erschließungsfehlern), und jeder Deskriptor wird pro Dokument nur einmal vergeben.

Die idf-Komponente ist nur relevant, wenn die Suchanfrage mehr als ein Wort enthält. Bei einer Anfrage wie beispielsweise *Saxophon Geschichte* können wir vermuten, dass die Nutzer:innen Literatur über die Geschichte des Saxophons suchen. Wenn nun ein Text nur das Wort *Saxophon* enthält, das Wort *Geschichte* dagegen nicht, ist es mit hoher Wahrscheinlichkeit relevanter als ein Text, in dem nur *Geschichte* vorkommt. Dieser Ansatz ließe sich eventuell auch auf Schlagwörter übertragen, wenn häufig und selten verwendete Schlagwörter in einer Anfrage kombiniert werden.

3.3.2 Lemmatisierung

Bei Sprachen mit ausgeprägter Morphologie ist die Lemmatisierung ein wichtiger Aspekt der Termnormierung. In einem Volltextsuchsystem muss die Lemmatisierung sowohl für die Wörter im Text als auch für die Suchwörter durchgeführt werden, um den Abgleich in jeder Kombination zu gewährleisten. In einem auf verbaler Erschließung basierenden System wird in der Regel die Ansetzungsform (das *Lemma*) als Deskriptor verwendet, und die Nutzer:innen sollten dann möglichst nur Ansetzungsformen für die Abfrage verwenden (es sei denn, dass das Retrievalsystem eine korrekt funktionierende Lemmatisierung anwenden kann). Bei Substantiven werden jedoch in vielen Fällen Pluralformen verwendet. Hier hängen die Brauchbarkeit und die Qualität eines auf verbaler Erschließung basierenden Systems von der Klarheit der Regeln ab, in welchen Fällen Singularformen und in welchen Pluralformen zu verwenden sind.

Beim Information Retrieval wird oft anstelle der Lemmatisierung heuristisches Stemming angewandt. Heuristisches Stemming ist linguistisch nicht klar definiert und unterscheidet in der Regel nicht zwischen Ableitungs- und Flexionsmorphologie, so dass nicht nur Flexionsformen eines Wortes, sondern auch semantisch weit entfernte Wörter mit demselben Stamm zusammengeführt werden. Brandts (2004) sieht daher eine Überlagerung von positiven und negativen Effekten, die sich ungefähr ausgleichen. Singh und Gupta (2016) dagegen stellten fest, dass für verschiedene Datensätze und Sprachen die meisten Stemming-

3 Streng genommen unterscheiden probabilistische Modelle nicht zwischen mehr und weniger relevanten Ergebnissen.

algorithmen einen positiven Effekt auf die Retrievalqualität haben (bis zu 50 % Steigerung der durchschnittlichen Genauigkeit für bulgarische CLEF-Daten[4]). Auch bei Sing und Gupta (2016) gab es jedoch Fälle, in denen das heuristische Stemming die Retrievalergebnisse negativ beeinflusste.

3.3.3 Anfrageerweiterung

Viele Retrievalsysteme nutzen die Möglichkeit, einen Term durch andere Terme zu ersetzen, in der Regel durch Synonyme oder Ober- und Unterbegriffe (Hyperonyme und Hyponyme). Die Algorithmen und Systeme mit solchen Funktionalitäten sind wiederum meistens im Kontext von Verfahren zur Volltextsuche entwickelt worden, bei denen verschiedene Ausdrücke in einem Text oder einer Anfrage auf das gleiche Konzept verweisen können. Die Ersetzung von Ausdrücken kann entweder im Index oder bei der Bearbeitung der Anfrage erfolgen. Im ersten Fall werden, wenn ein Wort in einem Text gefunden wird, als synonym definierte Ausdrücke in den Index mit aufgenommen. Im zweiten Fall werden Synonyme der Suchwörter zur Anfrage hinzugefügt: Die Anfrage wird also um zusätzliche Terme erweitert. Die Anfrageerweiterung (bei der Volltextsuche) wird in der Regel mit einer Gewichtung kombiniert (Manning, Raghavan und Schütze 2008: 174), wobei die hinzugefügten Terme ein geringeres Gewicht erhalten. Relevante verwandte Begriffe können entweder einem Thesaurus oder einem Wörterbuch entnommen oder mit statistischen Methoden aus Textkorpora, Anfrageprotokollen usw. gewonnen werden. Der Fall, dass der indizierte Bestand selbst als Quelle für die Anfrageerweiterung genutzt wird, wird weiter unten in den Abschnitten zu Relevanzfeedback erörtert.

Es wird häufig berichtet, dass zusätzliche Ausdrücke die Trefferquote (*recall*), in vielen Fällen aber auch die Genauigkeit (*precision*) der höchst gerankten Ergebnisse verbessern, weil Texten, die sowohl den ursprünglich gesuchten Ausdruck als auch Synonyme davon enthalten, eine höhere Relevanzwahrscheinlichkeit zugewiesen wird. Wie Komarjaya, Poo und Kan (2004) betonen, ist das insbesondere dann der Fall, wenn häufige und allgemeine Ausdrücke in sehr kurzen Anfragen verwendet werden. Sie zeigten, dass die Genauigkeit für die Suche in einem OPAC durch eine Anfrageerweiterung um über 30 % verbessert werden konnte. Greenberg (2001) zufolge eignen sich für die automatische

4 CLEF (Cross-Language Evaluation Forum bzw. Conference and Labs of the Evaluation Forum) ist eine Initiative zur Evaluierung von Retrieval für Europäische Sprachen und eine Reihe von jährlich stattfindenden Workshops (seit 2000) mit gemeinsamen Aufgaben, die von verschiedenen Teams und Systemen gelöst werden.

Anfrageerweiterung engere Begriffe oder Hyponyme, während in einer interaktiven Suchumgebung auch andere verwandte Begriffe verwendet werden könnten. Eines der Hauptprobleme der Anfrageerweiterung ist die sogenannte *Query Drift* oder Anfrageverschiebung: Die Anfrage wird durch die zusätzlichen Begriffe nicht präziser oder vollständiger, sondern ein Aspekt erhält – womöglich nicht der Absicht der Anfrage entsprechend – ein zusätzliches Gewicht. Bei der Volltextsuche scheint die Anfrageerweiterung mit verwandten Ausdrücken aus einem intellektuell erstellten Thesaurus oder Wörterbuch nur für kurze Abfragen nützlich zu sein (Navigli und Velardi 2003). Azad und Deepak (2019) formulieren die noch stärker zugespitzte Aussage, dass das Ergebnis für gut formulierte Abfragen gar nicht, aber das für schlecht formulierte erheblich verbessert werde. Diese Verallgemeinerung scheint allerdings zu stark zu sein, da der Nutzen der Anfrageerweiterung von vielen Faktoren abhängt. Järvelin et al. (2001) berichten, dass in einem Experiment die Erweiterung mit Synonymen und spezifischeren Ausdrücken bei der Volltextsuche einen positiven Effekt hatte und dass gerade die Ergebnisse stark strukturierter Abfragen von der Erweiterung profitieren konnten. Neben anderen zeigen Navigli und Velardi (2003), dass der Nutzen dieser Art von Anfrageerweiterung gesteigert werden kann, wenn sie mit einer Disambiguierung auf der Ebene der Wortsemantik kombiniert wird. Raza et al. (2018) zufolge könnte speziell eine ontologiebasierte Anfrageerweiterung nützlich sein, wenn eine fachspezifische Ontologie verwendet wird – und zwar sowohl für die Genauigkeit (*precision*) als auch für die Trefferquote (*recall*).

Bei der intellektuellen Sacherschließung wird in der Regel nur der spezifischste Deskriptor gewählt, obwohl ein Text auch für einen allgemeineren Gegenstand relevant sein könnte. Angesichts der Tatsache, dass viele Vokabulare wie die Gemeinsame Normdatei (GND) oder die Library of Congress Subject Headings (LCSH) extrem spezifische Begriffe enthalten, würden wir erwarten, dass die Anfrageerweiterung mit spezifischeren Ausdrücken bei der Suche in OPACs nützlich sein könnte. Es gibt nur wenige Studien, die den Einsatz der Anfragerweiterung in OPACs systematisch evaluieren. Neben der oben erwähnten Arbeit von Komarjaya, Poo und Kan (2004) behandeln auch Vallet et al. (2005) ein System mit einem auf verbaler Sacherschließung basierenden Index, bei dem eine Anfrageerweiterung die Ergebnisse verbessern konnte. Alani, Jones und Tudhope (2000) und Tudhope et al. (2006) schlagen eine Art implizite Anfrageerweiterung vor, bei der der Matching-Algorithmus Suchbegriffe mit semantisch verwandten Indexbegriffen abgleichen kann. Es wurde jedoch keine zahlenmäßige Analyse der Leistung des Algorithmus bei einer großen Menge von Anfragen vorgelegt.

Schließlich scheint die Anfrageerweiterung in einem heterogenen Bestand unentbehrlich: Wenn verschiedene normierte Deskriptoren für denselben Begriff verwendet werden, was bei der Zusammenführung von Erschließungsdaten aus verschiedenen Quellen unvermeidlich ist, sind vollständige Ergebnisse nur möglich, wenn ein Suchwort um alle äquivalenten Ausdrücke oder um eng verwandte Ausdrücke aus den im Bestand verwendeten Erschließungssystemen ergänzt wird.

3.3.4 Relevanzfeedback

Wenn ein Retrievalsystem sich unsicher über die Intention der Anfrage ist, könnte es mit den Nutzer:innen interagieren. Es könnte z. B. einige Treffer auswerfen und fragen, welcher davon dem Sinn der Abfrage am besten entspricht, und dann weitere Ergebnisse liefern, die dem gewählten ähnlich sind. Diese Art des Feedbacks wurde in einem Retrievalsystem von Hancock-Beaulieu und Walker (1992) für einen OPAC implementiert, das den Nutzer:innen die Möglichkeit bot, weitere ähnliche Ergebnisse (*more like this*) anzufordern. In einer Studie mit Nutzer:innen stellten sie fest, dass in fast der Hälfte der Fälle, in denen diese Funktion genutzt wurde, mindestens ein weiteres relevantes Dokument gefunden wurde.

In vielen Systemen werden wieder und wieder dieselben Abfragen eingegeben. Das ermöglicht es, ohne Interaktion mit den Nutzer:innen ein implizites Relevanzfeedback aus deren Nutzungsverhalten zu gewinnen. Wird z. B. der zweite Treffer einer Liste angeklickt, aber nicht der erste, können wir davon ausgehen, dass die jeweiligen Anzeigedetails den zweiten Treffer im Sinne der Abfrage relevanter erscheinen lassen (Jung, Herlocker und Webster 2007). Bei der Volltextsuche kann ein implizites Relevanzfeedback insbesondere für die Verarbeitung mehrdeutiger Suchbegriffe, sehr allgemeiner Begriffe und unklarer Relationen zwischen Suchtermen nützlich sein. Es könnte auch darauf hindeuten, dass ein bestimmtes Dokument, obwohl es einen bestimmten Term enthält, für die entsprechende Anfrage nicht relevant ist.

Ob implizites Feedback für einen OPAC nützlich sein könnte, ist schwer zu sagen. Uns sind keine Studien zu diesem Thema bekannt.

3.3.5 Pseudo-Relevanzfeedback

Anstatt echtes Feedback zu verwenden, könnten wir auch einfach annehmen, dass ein Retrievalsystem im Wesentlichen gut funktioniert und alle weit oben-

stehenden Treffer relevant sind. Das System kann diesen obersten Treffern dann typische Ausdrücke entnehmen und die Abfrage im Hintergrund um diese erweitern (Manning, Raghavan und Schütze 2008: 171 f.). Ein solches Verfahren löst zwar nicht das Problem der Mehrdeutigkeit, aber es hilft, Texte zu finden, die für einen Suchterm relevant sind, diesen aber nicht enthalten. Wie bei der Abfrageerweiterung besteht bei dem Pseudo-Relevanzfeedback die Gefahr der Abfrageverschiebung.

Bei einem intellektuell mit normiertem Vokabular erschlossenen Bestand sollte kein Pseudo-Relevanzfeedback notwendig sein, da im Idealfall keine Synonyme fehlen. Hier würden wir erwarten, dass die negativen Auswirkungen der Abfrageverschiebung mögliche positive Effekte zunichtemachen. Bei einem heterogenen Bestand hingegen könnte Pseudo-Relevanzfeedback sehr wohl nützlich sein. Nehmen wir etwa an, unser Bestand enthielte einige Texte mit dem Deskriptor *Vögel* aus Thesaurus A und einige mit dem Deskriptor *Ornithologie* aus Thesaurus B. Bei einer ausreichend großen Schnittmenge wird das Pseudo-Relevanzfeedback ergeben, dass viele Texte, die bei der Suche nach *Ornithologie* gefunden wurden, auch mit *Vögel* erschlossen sind, und das System wird dann bei einer Abfrage mit *Ornithologie* auch Texte auswerfen, die nur den Ausdruck *Vögel* enthalten. Dieser Mechanismus könnte das Fehlen von Beziehungen zwischen Deskriptoren oder das Fehlen eines Mappings zwischen verschiedenen Vokabularen ausgleichen. Der Thesaurus, der von Komarjaya, Poo und Kan (2004) für die oben erwähnte Anfrageerweiterung verwendet wurde, wurde von genau dieser Idee ausgehend konstruiert, um für häufig verwendete Suchbegriffe entsprechende Library of Congress Subject Headings zu finden. Aus demselben Impetus heraus zeigen Lüschow und Wartena (2017), dass fehlende Schlagwörter automatisch ergänzt werden können. Sie betrachten einen Bestand, der hauptsächlich mit Medical Subject Headings (MeSH) erschlossen ist. Aus anderen Bibliotheken übernommene Datensätze enthalten jedoch in der Regel keine MeSH-Deskriptoren. Es wurde gezeigt, dass viele MeSH-Deskriptoren auf der Grundlage anderer Erschließungsdaten extrapoliert werden können.

3.3.6 Deep Learning

Neuronale Netze können Darstellungen von Wörtern durch Vektoren in einem hochdimensionalen Raum (vereinfacht gesagt also Zahlenreihen), sogenannte *Word Embeddings* lernen, die nicht unmittelbar interpretierbar sind, aber die es ermöglichen, semantische Beziehungen zwischen Wörtern zu modellieren. Neuerdings berechnen Modelle wie ELMO und BERT kontextabhängige Darstellungen. Zu einem gewissen Grad spiegelt die Worteinbettung dann die kontex-

tuell korrekte Bedeutung wider, aber auch unterschiedliche syntaktische Kontexte werden erfasst. Das Wort *Kinder* wird z. B. in den Phrasen *für Kinder* und *über Kinder* nicht durch dieselben Vektoren dargestellt (Devlin et al. 2018). Dai und Callan (2019) zeigen, dass Systeme, die diese kontextualisierten Worteinbettungen nutzen, das Potenzial haben, in natürlicher Sprache geschriebene Anfragen zu verstehen und bessere Ergebnisse zu liefern als eine unstrukturierte schlagwortbasierte Suche. Google nutzt solche Techniken jetzt schon für englischsprachige Abfragen, und rät, statt Folgen von Suchbegriffen ganze Sätze zur Formulierung von Abfragen zu verwenden (Nayak 2019). Ob diese Art der Suche auch auf Titelsätze angewandt werden kann, ist, soweit wir wissen, noch nicht untersucht.

3.3.7 Implikationen für die Evaluierung verbaler Sacherschließung

Die bisher diskutierten Volltextmethoden versuchen vor allem, die Effekte fehlender oder überflüssiger Deskriptoren auszugleichen. In manchen Fällen können mit diesen Methoden auch Mehrdeutigkeiten von Wörtern in einem Text oder in einer Abfrage aufgelöst werden. Obwohl einige der diskutierten Methoden Schwächen der Sacherschließung wahrscheinlich teilweise ausgleichen können und einige Studien tatsächlich einen positiven Effekt im OPAC-Retrieval zeigen, bleibt es unklar, wie gut die Erschließungsfehlertoleranz eines *Next-Generation*-OPACs sein kann.

Es gibt einige Möglichkeiten, in einem Retrievalsystem das Fehlen relevanter Deskriptoren auszugleichen, obwohl sich bei der Volltextsuche herausgestellt hat, dass anscheinend alle Arten der Anfrageerweiterung problembehaftet sind. Eine effektive Anfrageerweiterung scheint am ehesten dann möglich zu sein, wenn ein Bestand in nicht zu geringem Umfang verbal erschlossen ist, so dass auf dieser Basis weitere Begriffe ergänzt werden können. Es besteht auch kein Grund zu der Sorge, dass zu viele Terme ergänzt werden könnten, sofern geeignete Rankingalgorithmen angewandt werden, die sich bei der Arbeit mit einer großen Anzahl von Termen pro Dokument als sehr effektiv erwiesen haben. Dies steht im Einklang mit der Schlussfolgerung von Soergel (1994), dass eine möglichst vollständige Erschließung eine der wichtigsten Voraussetzungen für ein effektives Retrieval ist.

3.4 Thesaurus- oder ontologiebedingte Effekte

Bei der oben beschriebenen Erschließung wird Vokabular einem Begriffssystem (*Knowledge Organization System*, KOS) entnommen. Die Qualität des KOS beeinflusst auch die Qualität der Suchergebnisse. Wenn das KOS für eine Form der Abfrageerweiterung, Facettierung oder Disambiguierung verwendet wird, hat auch die Struktur des KOS direkten Einfluss auf die Suchergebnisse. Bhogal, Macfarlane und Smith (2007) bemerken, dass ein korrektes, aktuelles KOS mit einer guten Abdeckung des Fachgebiets für eine nützliche Abfrageerweiterung entscheidend ist.

Strasunken und Tomassen (2008) schlagen ein Framework für die Bewertung von Ontologien im Kontext des Information Retrieval vor. In einem Experiment stellten sie fest, dass die Suchergebnisse durch das Hinzufügen von Instanzen, Objekteigenschaften und Äquivalenzrelationen erheblich verbessert werden können. Buscaldi und Suárez-Figueroa (2013) untersuchen die Folgen von 30 Problemtypen, die in Ontologien vorkommen können, für elf verschiedene ontologiebasierte Retrievalsysteme, und kommen zu dem Ergebnis, dass die meisten dieser Ontologieprobleme bei vielen Systemen das Retrieval negativ beeinflussen. Trotz der Unterschiede zwischen ontologiebasierten Retrievalsystemen und OPACs können wir ein ähnliches Resultat für OPACs erwarten, die in der einen oder anderen Weise mit einer KOS-Struktur operieren.

4 Ausgewählte Studien zum Einfluss von Erschließung auf das Funktionieren des Retrievals

Obwohl die genannten Faktoren die Suchergebnisse beeinflussen und gleichzeitig mit der Erschließung interagieren, verwenden einige Studien das Retrieval tatsächlich zur Evaluierung einer Erschließung durch Verschlagwortung. Die klassische MEDLARS-Studie von Lancaster (1968) ist eine der wenigen Studien, bei der reale Nutzer:innen, reale Abfragen und die daraus resultierenden Treffer betrachtet wurden. Die Ergebnisse von 300 MEDLARS-Abfragen wurden untersucht, Auswirkungen der Erschließung wurden analysiert und eine Fehleranalyse durchgeführt (siehe auch: Saracevic 1998). Hliaoutakis, Zervanou und Petrakis (2009) untersuchten die Auswirkungen von zwei automatischen Erschließungsmethoden auf das Retrieval bei der Suche nach Abstracts und Volltexten. Die Evaluierung basierte bei den Abstracts auf 64 TREC-Anfragen und

bei den Volltexten auf 15 TREC-Anfragen; in beiden Fällen wurden TREC-Relevanzbewertungen herangezogen.
Lancaster (2003: 87) empfiehlt die folgende Simulationsmethode:

1. Auswahl eines Korpus von Textdokumenten
2. Erstellung von beispielsweise drei Fragestellungen pro Dokument, für die es als wichtige Antwortquelle angesehen werden kann (wobei eine Fragestellung auf das Hauptthema des Texts, die anderen beiden nicht auf das Hauptthema, aber doch auf wichtige Aspekte abzielen)
3. Formulierung einer Anfrage zu jeder Fragestellung durch erfahrene Suchspezialist:innen
4. Davon unabhängige Erschließung der Dokumente auf herkömmliche Art und Weise
5. Vergleich der Erschließung mit den Formulierungen der Abfragen, um festzustellen, ob die relevanten Dokumente aufgrund der gewählten Deskriptoren gefunden werden. (es sollte auch geprüft werden, ob andere relevante Dokumente gefunden werden)

Lykke und Eslau (2010) verglichen den Einfluss maschineller Klassierung und Volltextindexierung auf die Ergebnisse eines Retrievalsystems bei einem Pharmaunternehmen. Sie wählten zehn echte Topics aus dem Suchprotokoll des Unternehmens aus und führten für jedes davon drei Suchen durch: 1) eine Suche mit Hilfe der intellektuell vergebenen Deskriptoren aus dem fachspezifischen Thesaurus des Unternehmens; 2) eine Volltextsuche; 3) eine Volltextsuche mit einer Abfrageerweiterung unter Verwendung des Firmenthesaurus. Die Anfragen basierten auf den Termen aus der ersten Anfrage, die die ursprünglichen Nutzer:innen im Rahmen einer interaktiven Suche zu einem Topic formuliert hatten. Im Test wurde für jedes Topic nur eine Anfrage ohne Interaktion durchgeführt. Die Relevanz der gefundenen Texte wurde von den tatsächlichen Nutzer:innen auf einer 4-Punkte-Skala entsprechend der Arbeitsaufgabe bewertet. Die folgende Tabelle zeigt die Ergebnisse gemittelt über die zehn Themen:

Tab. 1: Genauigkeit (*precision*) und relative Trefferquote (*recall*) für drei Suchtypen in einer Studie von Lykke und Eslau (2010)

	Suche mit kontrollierten Deskriptoren	**Freitextsuche**	**Freitextsuche mit Anfrageerweiterung**
Relative Trefferquote	24 %	41 %	89 %
Genauigkeit	17 %	33 %	24 %

Was die Genauigkeit betrifft, sind die Ergebnisse überraschend; denn es wäre zu erwarten, dass – vor allem im Kontext eines Unternehmens, wo das Retrieval in zielgerichtete Informationsaufgaben eingebettet und auf diese ausgerichtet ist – menschliche Erschließer:innen die Relevanz der Themen eines Texts besser beurteilen können. Es ist denkbar, dass die einzelnen Suchthemen in natürlicher Sprache genauer hätten formuliert werden können. Da bei dem Test nur einzelne Abfragen ohne Interaktion durchgeführt wurden, lassen sich die Ergebnisse nicht unbedingt auf ein interaktives Information Retrieval übertragen.

Svarre und Lykke (2014) verglichen das Retrieval auf der Grundlage automatischer Kategorisierung mit dem Volltextretrieval im Intranet der dänischen Regierung, das von den Steuerbehörden genutzt wird. 32 Teilnehmer:innen führten in jeweils vier Sitzungen drei simulierte und eine reale Suche durch. Insgesamt wurden 128 Sitzungen durchgeführt, 64 in jedem der beiden Testsysteme mit insgesamt 564 Anfragen. Suchverhalten und -ergebnisse wurden durch Protokolle, Relevanzbewertungen auf einer Skala von 1 bis 3 sowie Interviews nach der Suche dokumentiert. Die Interviews nach der Suche lieferten auch qualitative Daten – Einblicke in die Überlegungen und Entscheidungen der Nutzer:innen jeweils bei der Taxonomiesuche und bei der schlagwortbasierten Suche. Der Erfolg wurde anhand von zwei Metriken gemessen: (a) Anfrageerfolg: Prozentsatz der Anfragen, die zu mindestens einem Text führten, dessen Relevanz mit 2 oder 3 bewertet wurde, und (b) Sitzungserfolg: Prozentsatz der Sitzungen, in denen das Problem des Suchszenarios gelöst wurde. Am besten schnitt die Volltextsuche ab: 31 % beim Anfrageerfolg und 89 % beim Sitzungserfolg im Vergleich zu 22 % und 84 % für die automatische Kategorisierung. Das Ergebnis wurde auf verschiedene Ursachen zurückgeführt. Der restriktivere UND-Operator war in beiden Systemen mit der gleichen Häufigkeit verwendet worden, was in dem System, das die Klassierung nutzte, zu sehr kleinen Ergebnismengen führte. Einige Teilnehmer:innen gaben auch an, dass sie aufgrund zu geringer Kenntnis der Taxonomie Schwierigkeiten hatten, geeignete Kategorien für ihre Suche zu finden.

Die letzten beiden Studien sind gute Beispiele dafür, wie eine extrinsische Bewertung verbaler Erschließung durchgeführt werden kann. Sie zeigen auch die Leistungsfähigkeit der Volltextsuche und bekräftigen die Auffassung, dass mehr Terme bessere Ergebnisse bringen als weniger Terme. Entsprechend stellt sich die Frage, ob verbale Erschließung überhaupt noch sinnvoll ist, wenn Volltexte und Volltextsuche zur Verfügung stehen. Diese Frage zu behandeln würde jedoch den Rahmen dieses Beitrags sprengen.

5 Fazit

Ein Retrievalsystem besteht aus vielen Komponenten, die ineinandergreifen müssen. Nicht nur die einzelnen Komponenten, sondern auch die Abstimmung der Komponenten untereinander beeinflussen die Qualität der Ergebnisse. Daher ist ein Retrievalsystem ein äußerst schwierig zu nutzendes Messinstrument für die Qualität von Erschließung. Soergel (1994) schloss daraus, dass es unmöglich sei, Erschließung auf der Grundlage von Suchergebnissen zu bewerten, und dass qualitative Studien vorzuziehen seien. Da jedoch die Suche in einem Retrievalsystem der eigentliche Zweck der Inhaltserschließung ist, scheint es dennoch erstrebenswert, in Erfahrung zu bringen, wie gut eine Suche auf der Basis von Deskriptoren in der Realität tatsächlich funktioniert. Da das Argumentieren über das erwartete Verhalten eines komplexen Retrievalsystems noch schwieriger und problematischer ist, scheint realistisches Testen doch der beste Weg zu sein.

Bei der Evaluierung der Qualität von verbaler Erschließung im Kontext des Retrievals sollten wir möglichst realistische Umstände und Systeme verwenden. Wenn möglich, sollte mit einem breiten Spektrum von Aufgaben, Nutzer:innen und Systemeinstellungen gearbeitet werden, um die Ergebnisse belastbarer und weniger von bestimmten Einstellungen abhängig zu machen. Einen guten Einstieg für die Auswahl eines Verfahrens und das weitere Vorgehen könnte der Vorschlag in Golub et al. (2016) bieten. Dieser Ansatz sollte jedoch noch eingehend empirisch überprüft werden, und es ist zu erwarten, dass viel zusätzliche, weiterführende Forschung nötig sein wird, um geeignete Evaluierungsdesigns für solch komplexe Phänomene wie Erschließung und Retrieval zu entwickeln.

Da allerdings bisher wenig über die Wirkungen verschiedener Herangehensweisen und einzelner Bausteine bekannt ist, ist vielleicht die wichtigste Erkenntnis die, dass man sehr genau dokumentieren sollte, welche Algorithmen, welche Aufgaben usw. in die Evaluierung einbezogen wurden. Zweitens sollte man mit Verallgemeinerungen äußerst vorsichtig sein, da es keine Gewähr gibt, dass Verfahren und Richtlinien für die Verschlagwortung, die sich in dem einen Kontext bewährt haben, sich auch für einen anderen eignen.

6 Literaturverzeichnis

Alani, Harith, Christopher Jones und Douglas Tudhope: Associative and Spatial Relationships in Thesaurus-Based Retrieval. In: Research and Advanced Technology for Digital Libraries. Proceedings of the 4th European Conference, ECDL 2000 Lisbon, Portugal, September

18–20, 2000. Hrsg. von José Borbinha und Thomas Baker. Berlin, Heidelberg: Springer 2000. S. 45–58. https://doi.org/10.1007/3-540-45268-0_5.

Antelman, Kristin, Emily Lynema und Andrew K. Pace: Toward a 21st century library catalog. In: Information technology and libraries (2006) Bd. 25 H. 3. S. 128–139. https://doi.org/10.6017/ital.v25i3.3342.

Azad, Hiteshwar Kumar und Akshay Deepak: Query expansion techniques for information retrieval: a survey. In: Information Processing & Management (2019) Bd. 56 H. 5. S. 1698–1735. https://doi.org/10.1016/j.ipm.2019.05.009.

Beaulieu, Micheline: Approaches to user-based studies in information seeking and retrieval: A Sheffield perspective. In: Journal of Information Science (2003) Bd. 29 H. 4. S. 239–248. https://doi.org/10.1177%2F01655515030294002.

Belkin Nick, Kalervo Järvelin, Evangelos Kanoulas, Birger Larsen, Thomas Mandl, Elaine Toms und Pertti Vakkari: Task-Based Information Retrieval. In: Evaluation Methodologies in Information Retrieval. Dagstuhl Seminar 13441. S. 117–119. https://doi.org/10.4230/DagRep.3.10.92.

Bhogal, Jagdev, Andrew Macfarlane und Peter Smith: A review of ontology based query expansion. In: Information Processing & Management (2007) Bd. 43 H. 4. S. 866–886. https://doi.org/10.1016/j.ipm.2006.09.003.

Borlund, Pia: The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. In: Information research (2003) Bd. 8 H. 3 S. http://informationr.net/ir/8-3/paper152.html (30.12.2020).

Buckley, Chris, und Ellen M. Voorhees: Evaluating evaluation measure stability. In: ACM SIGIR Forum (2017) Bd. 51 H. 2. S. 235–242. https://doi.org/10.1145/3130348.3130373.

Brants, Thorsten: Natural Language Processing in Information Retrieval. In: Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands. Hrsg. von Bart Decadt, Guy De Pauw und Véronique Hoste. Antwerpen: University of Groningen 2004. S. 1–13.

Brenner, Claude W. und Calvin N. Mooers: A case history of a Zatocoding information retrieval system. In: Punched cards: Their application to science and industry. Hrsg. v. Robert Casey, et al., 2. Aufl. New York: Reinhold Publishing 1958. S. 340–356.

Buscaldi, Davide und Mari Carmen Suarez-Figueroa: Effects of Ontology Pitfalls on Ontology-based Information Retrieval Systems. In: Proceedings of the International Conference on Knowledge Engineering and Ontology Development – Volume 1: KEOD. Hrsg. v. Joaquim Filipe und Jan Dietz. SciTePress 2013. S. 301–307. https://doi.org/10.5220/0004550203010307.

Cleverdon, Cyril W., Jack Mills und E. Michael Keen: Factors determining the performance of indexing systems. Bd. 1: Design. Cranfield 1966.

Dai, Zhuyun und Jamie Callan: Deeper Text Understanding for IR with Contextual Neural Language Modeling. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19). New York, NY: Association for Computing Machinery 2019. S. 985–988. https://doi.org/10.1145/3331184.3331303.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova: Bert:Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805 (1.3.2021).

Fidel, Raya: Moves in Online Searching. In: Online Review (1985) Bd. 9 H. 1. S. 61–74. https://doi.org/10.1108/eb024176.

Fox, Edward A., Robert K. France, Eskinder Sahle, Amjad Daoud und Ben E. Cline: Development of a modern OPAC: from REVTOLC to MARIAN. In: Proceedings of the 16th annual interna-

tional ACM SIGIR conference on Research and development in information retrieval (SIGIR '93). Hrsg. v. Robert Korfhage, Edie M. Rasmussen und Peter K. Willett. New York, NY: Association for Computing Machinery 1993. S. 248–259. https://doi.org/10.1145/160688.160730.

Gazendam, Luit, Christian Wartena, Véronique Malaisé, Guus Schreiber, Annemieke de Jong und Hennie Brugman: Automatic Annotation Suggestions for Audiovisual Archives: Evaluation Aspects. In: Interdisciplinary Science Reviews (2009) Bd. 34 H. 2–3. S. 172–188. https://doi.org/10.1179/174327909X441090.

Golub, Koraljka und Marianne Lykke: Automated classification of web pages in hierarchical browsing. In: Journal of documentation (2009) Bd. 65 H. 6. S. 901–925. https://doi.org/10.1108/00220410910998915.

Golub, Koraljka, Dagobert Soergel, George Buchanan, Douglas Tudhope, Marianne Lykke und Debra Hiom: A framework for evaluating automatic indexing or classification in the context of retrieval. In: Journal of the Association for Information Science and Technology (2016) Bd. 67 H. 1. S. 3–16. https://doi.org/10.1002/asi.23600.

Greenberg, Jane: Automatic query expansion via lexical–semantic relationships. In: Journal of the American Society for Information Science and Technology (2001) Bd. 52 H. 5. S. 402–415. https://doi.org/10.1002/1532-2890(2001)9999:9999%3C::AID-ASI1089%3E3.0.CO;2-K.

Hancock-Beaulieu, Micheline und Stephen Walker: An evaluation of automatic query expansion in an online library catalogue. In: Journal of Documentation (1992) Bd. 48 Nr. 4. S. 406–421. https://doi.org/10.1108/eb026906.

Hliaoutakis, Angelos, Kaliope Zervanou und Euripides GM Petrakis: The AMTEx approach in the medical document indexing and retrieval application. In: Data & Knowledge Engineering (2009) Bd. 68 H. 3. S. 380–392. https://doi.org/10.1016/j.datak.2008.11.002.

Huang, Xiaoli und Dagobert Soergel: Relevance: An improved framework for explicating the notion. In: Journal of the American Society for Information Science and Technology (2013) Bd. 64 H. 1. S. 18–35. https://doi.org/10.1002/asi.22811.

Hjørland, Birger: Subject (of Documents). 2016. In: ISKO Encyclopedia of Knowledge Organization. Hrsg. v. Birger Hjørland und Claudio Gnoli. http://www.isko.org/cyclo/subject (30.12.2020).

Hjørland, Birger: Indexing: concepts and theory. 2018. In: ISKO Encyclopedia of Knowledge Organization. Hrsg. v. Birger Hjørland und Claudio Gnoli. http://www.isko.org/cyclo/indexing (30.12.2020).

Iivonen, Mirja: Consistency in the selection of search concepts and search terms. In: Information Processing & Management (1995) Bd. 31 H. 2. S. 173–190. https://doi.org/10.1016/0306-4573(95)80034-Q.

Ingwersen, Peter und Kalervo Järvelin: Information retrieval in context: IRiX. ACM Sigir Forum (2005) Bd. 39. Nr. 2. S. 31–39. https://doi.org/10.1145/1113343.1113351.

Järvelin, Kalervo, Jaana Kekäläinen und Timo Niemi: ExpansionTool: Concept-Based Query Expansion and Construction. In: Information Retrieval (2001) H. 4. S. 231–255. https://doi.org/10.1023/A:1011998222190.

Jung, Seikyung, Jonathan L. und Janet Webster: Click data as implicit relevance feedback in web search. In: Information Processing & Management (2007) Bd. 43 H. 3. S. 791–807. https://doi.org/10.1016/j.ipm.2006.07.021.

Kim, Kyung-Sun: Effects of emotion control and task on web searching behavior. In: Information Processing & Management (2008) Bd. 44 H. 1. S. 373–385. https://doi.org/10.1016/j.ipm.2006.11.008.

Kim, Soojung und Dagobert Soergel: Selecting and measuring task characteristics as independent variables. In: Proceedings of the American Society for Information Science and Technology (2005) Bd. 42 H. 1. https://doi.org/10.1002/meet.14504201111.

Komarjaya, Jeffry, Danny C. C. Poo und Min-Yen Kan: Corpus-Based Query Expansion in Online Public Access Catalogs. In: Research and Advanced Technology for Digital Libraries. ECDL 2004. Hrsg. v. Rachel Heery und Liz Lyon. Berlin, Heidelberg: Springer 2004. S. 221–231. https://doi.org/10.1007/978-3-540-30230-8_21.

Lancaster, Frederick W.: Evaluation of the MEDLARS demand search service. Bethesda: U. S. Dept. of Health, Education, and Welfare, Public Health Service 1968.

Lancaster, Frederick W.: Indexing and Abstracting in Theory and Practice. 3. Aufl. London: Facet Publishing 2003.

Liu, Jingjing und Nicholas J. Belkin: Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. In: Journal of the Association for Information Science and Technology (2015) Bd. 66 H. 1. S. 58–81. https://doi.org/10.1002/asi.23160.

Lüschow, Andreas und Christian Wartena: Classifying Medical Literature Using K-Nearest-Neighbours Algorithm. In: Proceedings of the 17th European Networked Knowledge Organization Systems Workshop co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017), Thessaloniki, Greece, September 21st, 2017. http://ceur-ws.org/Vol-1937/paper3.pdf (30.12.2020).

Lykke, Marianne und Anna Gjerluf Eslau: Using thesauri in enterprise settings: Indexing or query expansion? In: The Janus faced scholar: A Festschrift in honour of Peter Ingwersen. Hrsg. v. Birger Larsen, Jesper Wiborg Schneider und Fredrik. Åström. Copenhagen: Det Informationsvidenskabelige Akademi 2010. S. 87–97.

Mai, Jens-Erik: Semiotics and indexing: an analysis of the subject indexing process. In: Journal of documentation (2001) Bd. 57 H. 5. S. 591–622. https://doi.org/10.1108/EUM0000000007095.

Manning, Christopher D., Prabhakar Raghavan und Hinrich Schütze: Introduction to Information Retrieval. Cambridge: Cambridge University Press 2008.

Markey, Karen: Interindexer Consistency Tests: A Literature Review and Report of a Test of Consistency in Indexing Visual Materials. In: Library and Information Science Research (1984) Bd. 6 H. 2. S. 155–77.

Moens, Marie-Francine: Automatic Indexing and Abstracting of Document Texts. Boston: Kluwer 2000.

Morris, Jane: Individual Differences in the Interpretation of Text: Implications for Information Science. In: Journal of the American Society for Information Science and Technology (2010) Bd. 61 Nr. 1. S. 141–149. https://doi.org/10.1002/asi.21222.

Navigli, Roberto und Paola Velardi: An analysis of ontology-based query expansion strategies. In: Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia. 2003. S. 42–49. http://staffwww.dcs.shef.ac.uk/people/F.Ciravegna/ATEM03/ATEM03-Proceedings.pdf (7.1.2021).

Nayak, Pandu: Understanding searches better than ever before. 2019. https://blog.google/products/search/search-language-understanding-bert/ (7.1.2021).

Olson, Hope A. und John J. Boll: Subject analysis in online catalogs. 2. Aufl. Englewood, CO: Libraries Unlimited 2001.

Raza, Muhammad Ahsan, Rahmah Mokhtar, Noraziah Ahmad, Maruf Pasha und Urooj Pasha: A Taxonomy and Survey of Semantic Approaches for Query Expansion. In: IEEE Access (2019) Bd. 7. S. 17823–17833. https://doi.org/10.1109/ACCESS.2019.2894679.

Rolling, L.: Indexing consistency, quality and efficiency. In: Information Processing & Management (1981) Bd. 17 H. 2. (1981): S. 69–76. https://doi.org/10.1016/0306-4573(81)90028-5.

Saracevic, Tefko (2007a): Relevance: A review of the literature and a framework for thinking on the notion in information science: Part II: nature and manifestations of relevance. In: Journal of the American Society for Information Science and Technology (2007) Bd. 58 H. 13. S. 1915–1933. https://doi.org/10.1002/asi.20682.

Saracevic, Tefko (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science: Part III: Behavior and effects of relevance. In: Journal of the American Society for Information Science and Technology (2007) Bd. 58 H. 13. S. 2126–2144. https://doi.org/10.1002/asi.20681.

Singh, Jasmeet und Vishal Gupta: Text Stemming: Approaches, Applications, and Challenges. In: ACM Computing Surveys (2016) Bd. 49 Nr. 3. https://doi.org/10.1145/2975608.

Soergel, Dagobert: Organizing information: Principles of data base and retrieval systems. Orlando u. a.: Academic Press 1985.

Soergel, Dagobert: Indexing and retrieval performance: The logical evidence. In: Journal of the American Society for Information Science and Technology (1994) Bd. 45 H. 8. S. 589–599. https://doi.org/10.1002/(SICI)1097-4571(199409)45:8%3C589::AID-ASI14%3E3.0.CO;2-E.

Strasunskas, Darijus und Stein L. Tomassen: Empirical Insights on a Value of Ontology Quality in Ontology-Driven Web Search. In: On the Move to Meaningful Internet Systems: OTM 2008. Hrsg. v. Robert Meersman und Zahir Tari. Berlin, Heidelberg: Springer 2008.

Suomela, Sari und Jaana Kekäläinen: User evaluation of ontology as query construction tool. In: Information Retrieval (2006) Bd. 9 H. 4. S. 455–475. https://doi.org/10.1007/s10791-006-6387-3.

Svarre, Tanja J. und Marianne Lykke: Experiences with automated categorization in e-government information retrieval. In: Knowledge Organization (2014) Bd. 41 H. 1. S. 76–84. https://doi.org/10.5771/0943-7444-2014-1-76.

Tudhope, Douglas, Ceri Binding, Dorothee Blocks und Daniel Cunliffe: Query expansion via conceptual distance in thesaurus indexed collections, In: Journal of Documentation (2006) Bd. 62 Nr. 4. S. 509–533. https://doi.org/10.1108/00220410610673873.

Vallet, David, Miriam Fernández und Pablo Castells: An ontology-based information retrieval model. In: The Semantic Web: Research and Applications. Proceedings of the Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29–June 1, 2005. Hrsg. v. Asunción Gómez-Pérez und Jérôme Euzenat. Berlin, Heidelberg: Springer 2005. S. 455–470. https://doi.org/10.1007/11431053_31.

Yu, Holly und Margo Young: The impact of web search engines on subject searching in OPAC. In: Information technology and libraries (2004) Bd. 23 Nr. 4. S. 168–180. https://doi.org/10.6017/ital.v23i4.9658.