
Textindexierung durch beispielbasierte Termextraktion

Leonhard Voltmer, Dr. Oliver Streiter

Textindexierung durch beispielbasierte Termextraktion

Leonhard Voltmer, Dr. Oliver Streiter

EURAC
research

EUROPÄISCHE
AKADEMIE

ACCADEMIA
EUROPEA

EUROPEAN
ACADEMY

BOZEN - BOLZANO

2003

Bestellungen bei:
Europäische Akademie Bozen
Viale Druso, 1
39100 Bozen - Italien
Tel. +39 0471 055055
Fax +39 0471 055099
E-mail: press@eurac.edu
Verantwortlicher Direktor: Stephan
Ortner

Per ordinazioni:
Accademia Europea Bolzano
Drususallee, 1
39100 Bolzano - Italia
Tel. +39 0471 055055
Fax +39 0471 055099
E-mail: press@eurac.edu
Direttore responsabile: Stephan Ortner

Nachdruck und fotomechanische Wieder-
gabe - auch auszugsweise - nur unter An-
gabe der Quelle (Herausgeber und Titel)
gestattet.

Riproduzione parziale o totale del
contenuto autorizzata soltanto con la
citazione della fonte (titolo ed edizione).

Textindexierung durch beispielbasierte Termextraktion

Voltmer/Streiter

Eine sprachunabhängige beispielbasierte Termextraktion wird zur Indexierung von Dokumenten verwendet. Nach einer theoretischen Einordnung des Indexierens als Teilaufgabe des Information Retrieval sowie der möglichen Verfahren (Teil 1) wird in Teil 2 die beispielbasierte Termextraktion beschrieben und in Teil 3 auf die Indexierungsaufgabe angewandt.

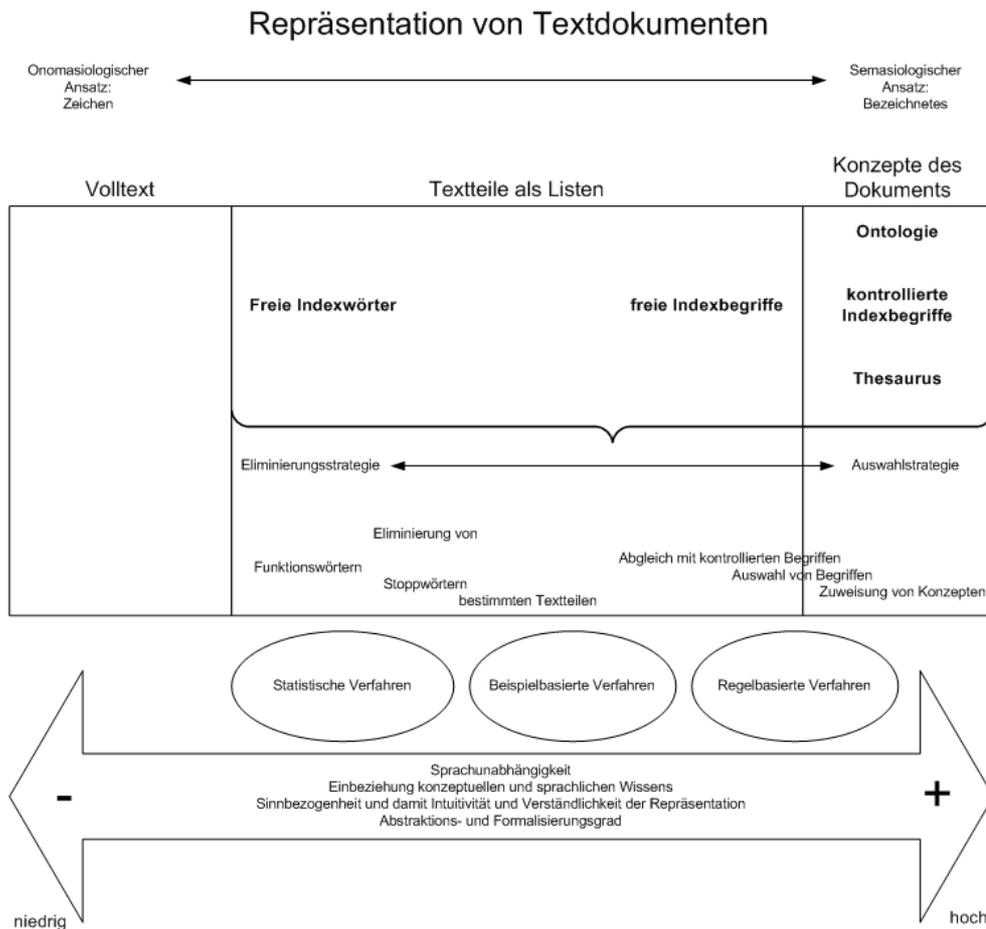
1. Einordnung des wissenschaftlichen Ansatzes

1.1. *Indexieren als Teilaufgabe des Information Retrieval*

Allgemeines Ziel des Information Retrieval (IR, Informationsgewinnung) ist es, ausschließlich alle für den Suchenden relevanten Dokumente auf vage Suchanfragen anzugeben. Dazu müssen die Dokumente und das Informationsbedürfnis in einer Weise repräsentiert werden, dass die Ähnlichkeit der Repräsentationen verglichen werden kann (RIJSBERGEN 1983).

Voraussetzung für die Ähnlichkeitsmessung ist ein gemeinsames Maß von Dokumenten und Informationsbedürfnis. Dokumente sind Zeichenketten, das Informationsbedürfnis menschlicher Benutzer ist wissensbezogen. Die Überbrückung der Kluft zwischen Zeichen und Bezeichnetem ist ein sehr komplexes Problem, mit dem sich ganze Wissenschaftszweige beschäftigen.¹ Im IR gibt es Ansätze, die dieses Problem ganz ausklammern bis hin zu solchen, die dieses Problem komplett zu lösen versuchen.

¹ Die Sprachwissenschaften und die Translationswissenschaften. Das Problem wird besonders deutlich beim Retrieval in mehrsprachigen Umgebungen, also mit Dokumenten oder Suchanfragen in verschiedenen Sprachen.



Dem entsprechend kann man auch beim IR zwischen semasiologischen und onomasiologischen Ansätzen unterscheiden und die Repräsentationen von Dokumenten oder Suchanfragen sind eher zeichen- oder wissensorientiert. Das zeichenorientierte Extrem ist eine Volltextsuche, die das Retrievalproblem wegen der Variabilität von Sprachen oft nicht zufrieden stellend löst. Das wissens-orientierte Extrem ist eine Ontologie², bei der die Bezeichnung einer Wissens-klasse im Dokument selbst kaum auftaucht. Keines dieser IR-Systeme

² Ontologie bedeutet hier die explizite formale Spezifikation einer gemeinsamen Konzeptualisierung.

bearbeitet die in den Dokumenten enthaltenen Zeichen, denn eine Volltextsuche verwendet alle Zeichen, eine Ontologie gar keine.

Viele Textdokumentsammlungen sind neben einer Volltextsuche über ein weiteres, hybrides, zeichenverarbeitendes IR-System zugänglich. Zeichenorientierte Ansätze verwenden eher eine Eliminierungsstrategie, wissensorientierte Ansätze eher eine Auswahlstrategie.

Bei der **Eliminierungsstrategie** werden von allen logisch möglichen Zeichenketten eines Dokuments ausgehend diejenigen Zeichenketten eliminiert, die zum Repräsentationszweck nichts beitragen, etwa weil ihre Unterscheidungskraft gering ist. Ein Beispiel für eine solche Eliminierung sind Stoppwortlisten. In aller Regel bleiben dabei relativ viele Zeichenketten übrig.

Bei der **Auswahlstrategie** werden vorher feststehende Bezeichnungen im Text gesucht oder dem Text zugeordnet. Meist beschränken sich solche Repräsentationen auf wenige Schlagwort-, Thesaurus- oder Indexbegriffe³.

In der Tendenz ist die Eliminierungsstrategie eher recallorientiert und kommt mit weniger explizitem sprachlichen Wissen aus, während die Auswahlstrategie eher precisionorientiert ist und sich eher auf linguistische Verfahren stützt (etwa zur Erstellung eines kontrollierten Vokabulars).

1.2. Textverarbeitende Indexierungsverfahren

Textverarbeitende Textrepräsentationsverfahren können nach der Struktur der Eingabe- und Ausgabedaten in drei Gruppen eingeteilt werden: Bei der Eingabe von Regeln zur Ausgabe von Indextermen ist der Input komplexer als der Output (**regelbasierte Verfahren**). Bei der Eingabe von Beispielbegriffen zur Ausgabe von Indextermen ist der Input von der gleichen Komplexität wie der Output (**beispielbasierte Verfahren**). Bei der Eingabe von Text zur Ausgabe von Indextermen ist der Input von geringerer Komplexität als der Output (**statistische Verfahren**).⁴

Diese Dreiteilung gibt an, wie komplex die Vorarbeiten zur Abstraktion und Formalisierung sind. Den höchsten Grad der Verarbeitung und Abstraktion

³ Ein Thesaurus enthält nach ISO 2788, 1986:2 ein kontrolliertes Indexvokabular, dem Konzepte mit eindeutigen Beziehungen zugrunde liegen.

⁴ Wissenschaftstheoretisch werden zunächst induktiv Regeln erstellt, die dann deduktiv angewandt werden. Der induktive Schritt wird beim regelbasierten Ansatz vor dem Beginn der Datenverarbeitung durch den Linguisten vollzogen, beim statistischen und beispielbasierten Ansatz ist die Induktion Teil der Datenverarbeitung.

erfordern Regeln und Ausnahmen zur Bildung der gesuchten Einheiten. Solche Regeln müssen in programmlesbarer Weise formalisiert sein. Beispielseinheiten erfordern nur relativ einfache Vorarbeiten und die Formalisierung erfolgt vom Programm nach vorgegebenen Parametern. Einfacher Text braucht gar nicht bearbeitet zu werden und enthält noch die volle Komplexität der Sprache.⁵

Nur der statistische und der beispielbasierte Ansatz kommen ohne explizites linguistisches Wissen aus. Der statistische Ansatz benutzt große Textmengen, um die wichtigen Begriffe herauszufinden. Dem beispielbasierten Ansatz dient die Struktur der Beispiele als Ausgangspunkt für die Suche nach Indextermen.

In der wissenschaftlichen **Literatur** werden vorherrschend statistische und regelbasierte Ansätze zur automatischen Indexierung durch Termextraktion⁶ verwendet. JACQUEMIN 2003 gibt den neuesten Stand wieder. JACQUEMIN unterteilt zum einen danach, ob die Suche nach Termen erstmals stattfindet oder ob eine bestehende Liste erweitert wird, und zum anderen danach, ob es einen Numerus Clausus für die Indexterme gibt oder ob grundsätzlich jeder Term in Frage kommt.

	Terme vorhanden	Keine Terme vorhanden
Erwerb von Termkandidaten	erweiternd	originär
Erwerb von Indextermen	Zuordnung von/zu kontrollierten Indextermen	Erstellung eines freien Indexes

Nach Jacquemin 2003

⁵ Aus dem Blickwinkel des IR soll das Kernproblem, die Kluft zwischen Zeichen und Bezeichnetem, mit der Eingabe von linguistischen Regeln gelöst werden, während es bei der Eingabe von Text auf die Textverarbeitung verschoben wird. Durch die Eingabe von Beispielen wird die Kluft nicht überbrückt, sondern es werden die Orte anderer Brücken angegeben, um das Auffinden geeigneter Überbrückungsmöglichkeiten zu erleichtern. Beispiele sind in funktioneller Sicht Text, der sich besonders gut zur Findung von Termbildungsregeln eignet.

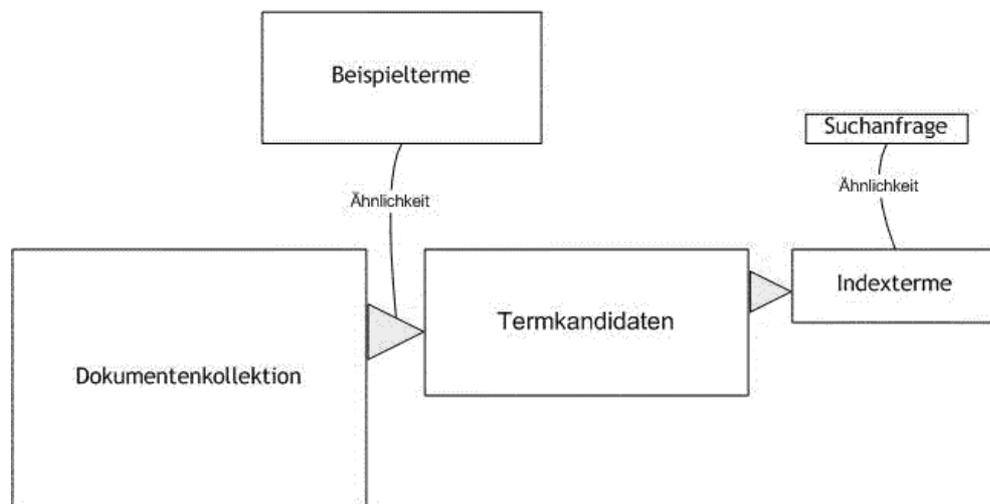
⁶ JACQUEMIN 2003 beschreibt die verschiedenen Ansätze und der einsprachlichen Termextraktionsanwendungen TERMINO, LEXTER, ACABIT, Xtract, ANA sowie der Termerkennungs- und Indexierungsanwendungen FASIT, CLARIT, TTP, COB, COPSY and FASTR.

Textindexierung durch beispielbasierte Termextraktion

In diesem Schema ist die beispielbasierte Indexierung nicht eindeutig einzuordnen. Zwar müssen Beispielterme vorhanden sein, allerdings genügen sehr wenige Beispiele, die manuell erstellt (Die Anforderungen an die Kenntnis der Sprache sind sehr gering) oder aus anderen Projekten (z.B. Internetglossaren) übernommen werden können. Die Beispiele werden außerdem zu Beispielmodellen weiter verarbeitet und kommen nicht in die Liste der Termkandidaten. Der Erwerb von Termkandidaten ist bei der beispielbasierten Termextraktion daher originär.

In der Regel werden die extrahierten Termkandidaten gleich als Indexterme in einen freien Index übernommen. Möglich wäre ein Abgleich mit feststehenden Termen zu einem kontrollierten Index. Dazu wäre die Termextraktion aber nicht nötig, weil die Indexterme sofort im Dokumenttext gesucht und indiziert werden können.

Modell retrievalorientierter Indexierung durch beispielbasierte Termextraktion



Die drei Ansätze können außerdem nach ihren Voraussetzungen, Sprachunabhängigkeit, Kontrollierbarkeit des Verfahrens, Verständlichkeit der Repräsentation und Wiederverwendbarkeit der vorausgesetzten Ressourcen unterschieden werden.

Die Voraussetzung für statistische Verfahren sind große Textmengen in elektronischer Form und in den Sprachen der Dokumente. In den gebräuchlicheren Sprachen stehen genügend große Textkorpora zur freien Verfügung.

Beim beispielbasierten Ansatz kann mit einem bis zwei Beispielen begonnen werden. Gefundene Lösungen können als weiteres Beispiel hinzugefügt werden (Lernschleife), sollten aber vorher verifiziert werden, weil sich sonst die Fehler fortpflanzen. Im unten beschriebenen Versuch ohne Lernschleife waren 100 Beispiele ausreichend.

Bei regelbasierten Verfahren werden von Experten linguistische Regeln ausgewählt. Dies setzt voraus, dass ein muttersprachlicher Linguist konzeptuelles und sprachliches Wissen formalisiert. Damit ist dieser Ansatz bezüglich des eingesetzten Wissens am anspruchsvollsten, bezüglich der empirischen Daten aber am anspruchslosesten.

Der statistische Ansatz ist insoweit sprachabhängig, als er Korpora in der Dokumentsprache erfordert. Der beispielbasierte Ansatz benötigt Beispiele in der Dokumentsprache sowie formale Ansatzpunkte für die Regelgenerierung, z.B. dass Satzzeichen Sinn Grenzen darstellen oder dass die Sprache am Wortende flektiert. Nur der regelbasierte Ansatz ist voll sprachabhängig.

Mit Kontrollierbarkeit des Verfahrens ist die gezielte Manipulierbarkeit der Ergebnisse durch die Eingabe gemeint. Der statistische Ansatz ist sehr manipulationsresistent, weil einzelne Manipulationen in der Menge der Daten untergehen. Der beispielbasierte Ansatz kann zwar leicht manipuliert werden, etwa durch Eingabe anderer Beispiele (siehe im Versuch unten die Eingabe von Fachbegriffen und allgemeinen Wörterbucheinträgen), aber nicht gezielt, da der Effekt kaum voraussehbar ist. Beim regelbasierten Ansatz kann der Experte einzelne Regeln mit spezifischer Funktion hinzufügen oder verändern und das Ergebnis damit gezielt verbessern. Zum einen können solche Manipulationen aber zu einer Überspezialisierung auf den Trainingsdatensatz führen. Zum anderen bilden die intellektuell gefundenen sprachlichen Regeln aufgrund ihrer Anzahl, komplexen Formalisierung und Interferenz ein Geflecht, das letztlich selbst für Experten weder anschaulich noch kontrollierbar bleibt.

Mit der Verständlichkeit textverarbeitender Indexierungsverfahren ist gemeint, wie intuitiv verständlich die Verbindung zwischen dem Index und den Dokumenten ist. Dies hängt davon ab, wie stark sinnbezogen der Index ist.

Regelbasierte Verfahren sollten besser als beispielbasierte sein, die besser als statistische sein sollten. Statistisch kann ein Eigenname ein besonders geeignetes Indexwort sein, an dem aber der Inhalt der damit indexierten Dokumente nicht abzulesen ist.

Unterschiede bestehen auch hinsichtlich der Wiederverwendung der einzelnen Komponenten, also der Eingabedaten und der verwendeten Programme. Die bei statistischen Verfahren verwendeten Texte stammen meist aus Korpora, die von Korpuslinguisten intensiv für die verschiedensten Untersuchungen genutzt werden. Beispiele werden meist Teil des Projektergebnisses, sie sind also weder zusätzliche Arbeit noch wieder verwendbar. Linguistische Regeln werden in der Praxis kaum je wieder verwendet. Programme des statistischen und beispielbasierten Ansatzes können eher wieder verwendet werden als diejenigen des regelbasierten Ansatzes.

Beispielbasierte Ansätze finden insbesondere bei der Textklassifikation und beim Parsen Verwendung. In der Textklassifikation werden bereits klassifizierte Dokumente verwendet, zum Parsen Parsingbäume oder einzelne geparste Einheiten. In Versuchen mit Chinesisch konnten mit einem einzigen Parsingbaum 43% aller Abhängigkeitsbeziehungen erkannt werden (Streiter 2003), was das besonders gute Verhältnis von Aufwand zu Ergebnis unterstreicht.

Beispielbasierte Ansätze eignen sich insbesondere für Minderheitssprachen, die weder Korpora noch muttersprachliche Computerlinguisten einsetzen können (Streiter 2003).

1.3. Beispielbasierte Indexierung

Beispielbasierte Ansätze bei der Verarbeitung natürlicher Sprachen benutzen vorgegebene Lösungen zum Auffinden ähnlicher Lösungen in noch unbearbeiteter Umgebung. Es müssen also nicht nur Lösungen, sondern auch Ähnlichkeitsparameter eingegeben werden. Im Gegensatz zu regelbasierten Ansätzen benötigen beispielbasierte Ansätze zwar keine vorgegebenen Regeln, aber zumindest die Angabe der Parameter, auf die sich die Regeln beziehen sollen. Solche Parameter beschreiben die Zeichenketten intern oder in ihrem Zusammenhang. Die gängigsten internen Parameter beziehen sich darauf, ob die Zeichenketten Satzzeichen, Groß- oder Kleinbuchstaben, Affixe oder Suffixe, Vokale oder Konsonanten, jeweils in bestimmter Anzahl und Position, enthalten.

Die Parameter zum Zusammenhang der Zeichenketten betrachten deren relative und absolute Position im Text.

Durch Lernschleifen können sowohl neue Beispiele als auch neue Aufbaumuster generiert werden. Beispiele und Regeln können mit einem Korpus überprüft werden (QUASTHOFF 2002).

2. Beispielbasierte Termextraktion

Die Europäische Akademie Bozen hat eine Datenbank juristischer Fachausdrücke in italienischer, deutscher, grödnerischer, gadertalerischer und fassanischer Sprache: <http://www.eurac.edu/bistro>. Zum schnelleren Auffinden neuer Begriffe wurde eine automatische, beispielbasierte Termextraktion implementiert. Als Beispiele verwendeten wir in drei Testreihen die fachsprachlichen Begriffe der Datenbank, allgemeinsprachliche Wörterbucheinträge und beide gemischt. Die Evaluierung erfolgte an einem grödner Text, dessen juristische Fachbegriffe vor der automatischen Termextraktion von einem Terminologen festgelegt wurden.

2.1. Regeln und Parameter

Das Programm zerlegt die Beispiele in Kompositionsregeln. Die so gefundenen Regeln und weitere fest vorgegebene werden auf alle Wörter und Wortkombinationen des Textes angewandt, bis nur noch Termkandidaten übrig bleiben.

Als erste feste Regel wurde eingegeben, dass kein Begriff Satzzeichen enthält. Als zweite Regel wurde vorgegeben, dass die in der jeweiligen Sprache häufigsten Wörter nicht an erster oder letzter Stelle eines Terms stehen dürfen. Im Grödnerischen wurden als „Hintergrundkorpus“ zwanzig Texte aus dem Internet benutzt. Hinter dieser Regel steht die Annahme, dass die häufigsten Wörter Funktionswörter sein werden (Merkel 1994). Die dritte Regel legt eine Standarddeviation fest, so dass kein Termkandidat mehr als drei Standardabweichungen von den eingegebenen Beispieltermen haben darf.

Als erster Parameter wird angegeben, dass aus der Groß- und Kleinschreibung der Beispielterme Muster herausgesucht werden sollen. Der Begriff „deliberaziun dl Consëi comunal“ (Gemeinderatsentscheidung) erzeugt das Muster klein-klein-groß-klein.

Als zweiter Parameter wurden die Wortendungen und Funktionswörter zur Musterbildung verwendet. Der Begriff „deliberaziun dl Consëi comunal“ würde in das Muster *n-dl-*i-*l zerlegt, weil „dl“ ein häufiger Artikel ist und daher ganz beibehalten wird.

Die festen Regeln sind:

- Kein Satzzeichen
- Kein häufiges Wort am Anfang oder Ende
- Nicht mehr als drei Wörter Standardabweichung von den Beispielen

Die Parameter zur Regelfindung aus den Beispielen sind:

- Muster der Groß- und Kleinschreibung
- Muster aus Endungen und Funktionswörtern

Termkandidat ist jedes Wort oder Kombination von Wörtern, die mit allen festen Regeln und zusätzlich mit einer der als „TermmodeLL“ generierten Regeln übereinstimmt.

2.2. Ranking

Die extrahierten Termkandidaten müssen nach ihrer Termwahrscheinlichkeit geordnet werden. Das Ranking von Indextermen ist besonders dann ausschlaggebend, wenn man sich auf die ersten Terme als Indexterme beschränkt. Selbst wenn Recall und Precision schlecht sein sollten, könnte so aus einer „schlechten“ Termextraktion mit gutem Ranking noch eine gute Indexierung werden.

Beim Information Retrieval wird zum Ordnen von Ergebnisdokumente nach ihrer Relevanz oft wird das tf.idf-Maß verwendet, also die Häufigkeit im jeweiligen Text dividiert durch die Häufigkeit in allen Dokumenten. Dafür steht hier allerdings nur der kleine Hintergrundkorpus zur Verfügung und insbesondere Mehrworausdrücke kommen dort zu selten vor. Beim tf.idf-Maß stünde dann häufig 0 im Nenner und wäre keine geeignete Vergleichsgröße. Daher wird die einfachere Variante Wortlänge mal Wortfrequenz zum Ordnen verwendet.

2.3. Versuch

Die Termextraktion lief über eine grödner Gemeindeordnung mit 994 Wörtern, in dem ein Terminologe 113 Terme gefunden hatte. Die Extraktion auf Basis von Fachbegriffen ergab wie erwartet eine höhere Präzision, die sie aber

mit niedrigerem recall bezahlen musste.⁷ Den höchsten recall hatte ein Mix aus fachsprachlichen und allgemeinsprachlichen Begriffen.

Methode	Anzahl extrahierter Termkandidaten	recall	precision	mean
Fachbegriffe	299	0.7321	0.284	0.410
Allg. Wörterbuch	322	0.75	0.269	0.396
Mix	390	0.839	0.248	0.386

Es werden also vier von fünf Termen gefunden, aber nur jeder vierte gefundene Termkandidat ist ein Term. Mit anderen Worten wird das Unithoodproblem hervorragend gelöst, während die Termhoodaufgabe noch verbessert werden muss. Wenn die Terminologen die richtigen Termkandidaten direkt auf der Ergebnisseite weiterbearbeiten, dann könnte dies dem Programm rückgemeldet werden und die Präzision würde durch diese Lernschleife vermutlich erhöht werden.

Höchst bemerkenswert sind die **Ranking**qualitäten trotz der einfachen Berechnung. Auch bei weiteren Versuchen waren die ersten fünf Termkandidaten stets Terme, unter den ersten zehn Termkandidaten höchstens ein Nichtterm, dem von den Terminologen im Nachhinein Termqualität zugestanden werden konnte.⁸

In HONG 2001 werden mit Hilfe von manuell erstellten linguistischen Regeln, mit einem statistischen Filter und zusätzlich einer manuellen Stoppwortliste aus einem Text mit 74676 Wörtern 3176 Termkandidaten extrahiert, von denen in der ex-post Bewertung nur 2372, also 75 % Terme waren. Auf den obigen Versuch umgerechnet bedeutet das, dass etwa 12 Termkandidaten extrahiert würden und nur neun Terme wären. Trotz der manuellen Arbeit war das Ergebnis der

⁷ Die höhere Gesamtqualität, erkennbar im mean-Wert, ist statistisch nicht signifikant.

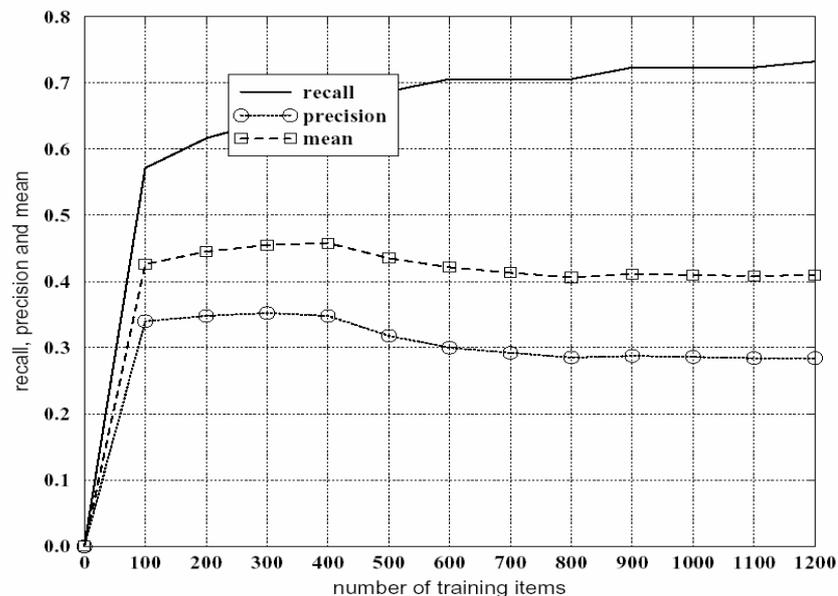
⁸ Ein solcher Term war „Art.“, also die Abkürzung für Artikel; ein Begriff, der für eine Satzung sicher charakteristisch ist und von dem angenommen werden kann, dass er eventuell auch gesucht wird, z.B. als „Art. 1 Gemeindeordnung“.

Extraktion bei HONG 2001, so weit die Versuche vergleichbar sind⁹, nicht besser. Sie entsprechen den Ergebnissen in DAILLE 2000.

Untersucht wurde Recall, Precision und mean-Wert auch in **Abhängigkeit von der Anzahl der Eingabebegriffe**. Der mean-Wert ist ein Maß für die Gesamtqualität durch Kombination von Recall und Precision zu

$$\text{mean} = \frac{2 * \#\{TC \cap T_{doc}\}}{\#\{T_{doc}\} + \#\{TC\}}$$

Recall, Precision und mean-Wert in Abhängigkeit von der Anzahl der Beispiele



Es zeigte sich, dass mit 100 Beispielen bereits die höchste Qualität der Termextraktion erreicht wird, weil bei zunehmenden Beispielen der höhere Recall mit niedrigerer Precision einhergeht.

⁹ Der Vergleich von Termextraktionsversuchen ist schwierig, weil in der Evaluierung meist kein oder nur ein geschätzter Recall angegeben wird, da die Gesamtanzahl der Terme im Text nicht bekannt ist.

Die Erklärung dafür könnte sein, dass die zuerst generierten Regeln mit statistisch höherer Wahrscheinlichkeit die häufig vorkommenden Terme beschreiben, während die später noch hinzukommenden Regeln nur noch die selteneren Termbildungen beschreiben und dabei relativ häufiger auch auf Nichtterme passen. Das „Hintergrundrauschen“ in Form von zufällig auf die Termmodelle passenden Mustern fällt bei selten passenden Termmodellen stärker ins Gewicht. Das Recall steigt zwar weiter an, die Qualität des Gesamtsystems (mean) sinkt aber aufgrund der überproportional hereinkommenden falschen Termkandidaten.

3. Textindexierung durch Termextraktion

Die Hypothese dieser Arbeit ist, dass die automatisch extrahierten Termkandidaten gleichzeitig Indextermkandidaten sind, weil die im Text enthaltenen Begriffe den Text selbst geeignet repräsentieren. Leider kann diese Hypothese hier nicht experimentell überprüft werden, sie kann aber theoretisch untersucht werden.

3.1. Einwände gegen die Verwendung von Termkandidaten als Indexterme

Die wichtigsten Einwände gegen die Übertragbarkeitshypothese sind:

1. Die Indexterme repräsentieren die Dokumente schlecht, wenn Themen nicht ausdrücklich bezeichnet werden und wenn ausdrückliche Bezeichnungen nicht dem Thema entsprechen.

Der erste Teil der Kritik beruht auf der Variabilität der Lexikalisierung von Konzepten. Ein Konzept kann tatsächlich auf verschiedene Weise angesprochen werden, aber auf irgendeine Weise wird es angesprochen werden müssen. Die extrahierten Begriffe müssen ebenso subtil interpretiert werden wie sie im Ausgangstext angesprochen werden. Außerdem kann gerade die Lexikalisierung entscheidender Anhaltspunkt für die Anfrage sein. Beispielsweise könnte nach

„Fuchs Trauben“ gesucht werden, um Äsops Fabel¹⁰ zu finden, auch wenn weder *vulpes* noch *vitis vinifera* mit dem Sinn der Fabel zu tun haben.

Damit trifft dieser Einwand eher Ontologien, die gerade die Sinnrelationen herausarbeiten wollen und Thesauri, die mit kontrolliertem Vokabular arbeiten, und in geringerem Maße Ansätze, die von den vorhandenen Zeichenketten im Text ausgehen.

Der zweite Teil des Einwands ist schwerwiegender. Es besteht die vage Hoffnung, dass die Begriffe, die nicht Thema sind vor allem in Sachtexten weniger spezifisch und weniger frequent sind als die thematisierten. Letztlich bleibt dies aber ein grundsätzliches Problem aller Dokumentrepräsentationen, die von den Zeichenketten ausgehen.

2. Die vielen falschen Indexterme machen den automatisch erzeugten Index unbrauchbar.

Für kleine Textkollektionen können die relevanten Indexterme durch Fachleute ausgewählt werden. Bei sehr großen Textkollektionen könnte man nur diejenigen Indexterme verwenden, die zweimal vorkommen und somit Nichtterme ausfiltern. Falsche Indexterme sind nur dann störend, wenn durch sie falsche Treffer entstehen oder relevante Dokumente nicht gefunden werden. Das passiert dann, wenn zwischen der Repräsentation der Retrievalanfrage und der Repräsentation des Textes aufgrund der falschen Indexterme ein anderes, falsches Ähnlichkeitsmaß errechnet wird. Eine geeignete Ähnlichkeitsermittlung sollte das verhindern können. Bezieht man beispielsweise nur volle Treffer in die Ähnlichkeit ein, dann werden sinnlose Nichtterme niemals Treffer erzeugen. Andererseits können auch Nichtterme Retrievalqualität haben, wenn kein Indexterm mit voller Übereinstimmung vorliegt. Dann wäre das Retrieval ein Mittelweg zwischen Index und Volltext, nämlich anhand textcharakteristischer Textfetzen. Je größer die Textkollektion im Verhältnis zu den möglichen Suchanfragen, umso wahrscheinlicher werden Volltreffer in den Indextermen vorliegen und umso weniger fallen falsche Indexterme ins Gewicht.

¹⁰ Die Fabel vom Fuchs und den Trauben (nach Äsop): Ein Fuchs ging an einer Mauer entlang. Oben stand ein Weinstock, von dem blaue Trauben herabhingen. Der Fuchs sprang in die Höhe, um zu den Trauben zu gelangen. Aber er konnte sie nicht erreichen. Da ging er weiter seines Weges und sagte: „Die Trauben sind mir zu sauer.“

Grundsätzlich stellt sich jedoch die Frage, ob über die Retrievalfunktion hinaus weitere Zwecke mit dem Index verfolgt werden. Für Retrievalzwecke genügen „terminologisch relevante Kollokationen“ (HEID 1999), die aber nicht unbedingt konzeptbeschreibende Terme sind. Ein Index mit vielen Nichttermen ist für Menschen unverständlich und weder manuelle Klassifikation noch die Themenauswahl direkt aus der Indexliste sind dann möglich. Um zu verhindern, dass Nichtterme indexiert werden, müsste die Ergebnisliste rechtzeitig, etwa nach 10 Termen, abgeschnitten werden. Es wäre noch zu untersuchen, wie sich diese verkürzte Indexierungsbreite¹¹ auf die Retrievalqualität auswirkt, denn je feiner die Dokumente unterschieden werden sollen, umso mehr Indexterme sind nötig. Zumindest für große Textkollektionen muss auf weitere Termkandidaten zurückgegriffen werden.¹²

3. Die Indexterme spannen einen n-dimensionalen Raum auf, obwohl Terme mit semantisch unerheblichen Abweichungen (Mehrzahl, Schreibvarianten) in derselben Dimension liegen.

Eine Lösungsmöglichkeit wäre die weitere Bearbeitung der Indexliste, indem Terme mit geringer graphemischer Abweichung unter dem häufigsten Begriff zusammengefasst werden oder alle als Treffer gewertet werden (Ekmekçioglu 96). Dies funktioniert nur bei Abweichungen in wenigen Zeichen und löst nicht das Problem bei Ellipsen und Synonymen.

Bei einer Klassifikationsaufgabe fällt dieses Problem viel weniger ins Gewicht, weil die Ähnlichkeit nicht über einige wenige Terme, sondern über alle Wörter berechnet wird. Je weniger Indexterme also verwendet werden, umso drängender wird das Problem. Es besteht daher ein Zielkonflikt zwischen der Extraktion von möglichst vielen Termen und möglichst präzisen Termen (Einwand 2).

¹¹ Indexierungsbreite ist das Ausmaß, in dem der fachliche Inhalt eines Dokuments von seinen Indexbegriffen abgedeckt wird. Um die Indexierungsbreite in ein Maß fassen zu können, wird davon ausgegangen, dass die Abdeckung proportional mit der Anzahl der Indextermini steigt. Man gibt daher die durchschnittliche Anzahl der Indexbegriffe pro Dokument an. (Knorz 1996).

¹² Eine Möglichkeit, den Index weiterhin menschenverständlich zu halten wäre der Abgleich mit einem Thesaurus bzw. einer umfangreichen Liste zugelassener Indexterme.

4. Lange Texte produzieren mehr Indexterme und ihre Relevanz wird überbewertet.

Eine Lösung wäre es, das Maß der Ähnlichkeit durch die Übereinstimmung in Buchstaben zu berechnen und den Wert des Treffers mit der Anzahl der vorhandenen n-Gramme ins Verhältnis zu setzen. Die überbewerteten langen Texte würden damit zumindest an das Ende der Ergebnisliste gedrückt. Außerdem haben lange Texte das Handicap, dass mehrfach vorkommende Begriffe nur einmal im Index auftauchen. Das wirkt der ungerechtfertigten Höherbewertung sich stark wiederholender Texte entgegen.

Nach dem Zipfschen Gesetz müssten sich die zusätzlich hinzukommenden Begriffe und die Begriffswiederholungen bei Einberechnung der Textlänge ins Ähnlichkeitsmaß die Waage halten.

3.2. Vorteile der Termextraktion beim Indexieren

Der Vorteil der beispielbasierten Indexierung liegt sicherlich in den **geringen sprachlichen und logistischen Voraussetzungen**, die diese Methode stellt. Es genügen zwanzig Texte und hundert Beispielsbegriffe in einer Sprache, um einen Index aufzubauen. Damit eignet sich die Methode besonders für Minderheitensprachen. Es muss sich nicht einmal um eine Sprache handeln, denn die Methode funktioniert auch bei anderen Zeichenketten. Es können also auch Musiknoten, Bilder und überhaupt **alle digitalen Dokumente** indexiert werden, sofern man Indexierungsbeispiele beibringen kann und nötigenfalls einige einfache allgemeine Regeln angibt.

Der **Aufbau eines Indexes nach der Struktur der vorgegebenen Beispiele** bringt weitere Vorteile mit sich. PAIJMANS (1997) zeigt, dass weder durch eine bestimmte Stellung von Wörtern im Text noch durch ihre Nähe zu Signalwörtern (cuewords) ein höherer Informationsgehalt festgestellt werden konnte. PAIJMANS findet aber eine größere Bedeutung von bestimmten Wortarten, etwa von Adjektiven, Verben und Substantiven. Diese Erkenntnis spricht für die beispielbasierte Indexierung, die bestimmte Wortarten bevorzugt. Durch die Modellgenerierung anhand der Beispiele werden nur beispielsähnliche

Zeichenketten indexiert, also bei Eingabe von Adjektiven praktisch nur Adjektive.

Aufgrund dieser **Flexibilität** und des geringen Aufwands bei der Neuerstellung eines Indexes kann zum Import und Export von Dokumenten stets ein neuer, gemeinsamer Index erstellt werden. Das besonders aufwendige Übertragen von Indexkategorien entfällt.

Die Flexibilität der Indexierung kann ausgenutzt werden, um die gleiche Dokumentensammlung gleichzeitig für verschiedene Zwecke verschieden zu indexieren. Eine medizinische Datenbank könnte beispielsweise für Fachleute aufgrund lateinischer Fachbegriffsbeispiele und für Lerner der medizinischen Fachsprache aufgrund von Allgemeinwörterbuchbeispielen erstellt werden. Aufgrund der Beispiele werden jeweils die stärker informationstragenden Terme extrahiert und indiziert.

Nebeneinander bestehende Indexe wären auch für Dokumente mit besonderen Textstrukturen und -eigenschaften geeignet. Dann könnten automatisch verschiedene Indexe für die Zusammenfassungen, den Text und die Bibliographie von wissenschaftlichen Artikeln zur Repräsentierung verwendet werden.

Durch die **Veränderung des Termfrequenzparameters** kann für eine bestimmte Dokumentkollektion oder überhaupt für eine Sprache indexiert werden. Wird für die Gewichtung eines Termkandidaten seine Frequenz im Dokument mit seiner Frequenz in der Dokumentensammlung ins Verhältnis gesetzt, so beschreibt das Maß die Charakteristik eines Dokuments im Gegensatz zu anderen Dokumenten der Sammlung.¹³ Wird die Dokumentfrequenz mit einem Allgemeinkorpus ins Verhältnis gesetzt, so bleibt die Spezifik (z.B. die Fachsprachlichkeit) eines Dokuments auch dann erhalten, wenn er sich in einem Fachkorpus befindet. Die Folge wäre einerseits, dass sich die verschiedenen fachsprachlichen und die allgemeinsprachlichen Dokumente voneinander trennen, andererseits aber, dass viele Dokumente dieselben Indexterme tragen und Suchanfragen stets extrem viele oder extrem wenige Dokumente ergeben. Die Gewichtung sollte also danach gewählt werden, ob die Relevanz der Ergebnisdokumente für eine Suchanfrage im Kontext idealer Information

¹³ Oder genauer: Die Charakteristik des Termkandidaten für dieses Dokument im Gegensatz zu seiner Charakteristik für den Durchschnitt aller anderen Dokumente der Sammlung.

beurteilt wird (absolut) oder ob die Relevanz im Kontext der Dokumentenkollektion (relativ) beurteilt wird.

Schließlich kann die beispielbasierte Indexierung mit statistischen oder regelbasierten Methoden verbunden und verbessert werden, wenn die nötigen Ressourcen vorhanden sind. Das könnte insbesondere bei Sprachen nötig sein, bei denen die Ergebnisse aufgrund vielfältiger Möglichkeiten zur Zusammensetzung, Beugung und Ableitung von Wörtern nicht hinreichend gut sind. Obwohl der Ansatz unabhängig von der Eingabesprache funktioniert, funktioniert er nicht stets gleich gut. Die Extraktion/Indexierung wird bei romanischen Sprachen besser sein als bei germanischen oder slawischen Sprachen.

3.3. Ausblick

Interessant wäre, ob sich die Retrievalqualität steigern ließe, wenn man nicht Wörterbucheinträge oder Fachbegriffe zugrundelegt, sondern tatsächliche Suchanfragen. Zwar kann erst gesucht werden, wenn bereits ein Index aufgebaut ist, aber der Index könnte später mit gespeicherten Suchanfragen noch einmal neu aufgebaut werden. Einerseits ist zu vermuten, dass die Ähnlichkeit, die zwischen Anfrage und Indexbegriffen damit größer wird. Andererseits werden Suchanfragen wohl häufig als Kombinationssuche nach mehreren Begriffen formuliert, die Termextraktion würde die kombinierten Worte aber fälschlicherweise als zusammengehörend interpretieren und unsinnige Muster generieren.

Die Ausnutzung bestehender Termextraktionsmethoden zum Indexieren für Retrievalzwecke erscheint insgesamt als möglicher Ansatz, muss ihre Tauglichkeit aber erst noch experimentell erweisen. Für Versuche steht die Termextraktion unter der Adresse <http://dev.eurac.edu:8080/cgi-bin/index/TermExtract> für die Sprachen Deutsch, Englisch, Französisch, Italienisch, Ladinisch allgemein, Grödnerisch, Gadertalerisch und Fassanisch bei automatischer Spracherkennung zur Verfügung. Die Texte zur Indexierung/Extraktion können als URL oder über ein Textfenster eingegeben werden. Die Sortierung der Termkandidaten kann nach Termfrequenz,

Termfrequenz mal Länge, Ähnlichkeit von n-Grammen, weirdness ratio¹⁴ (Brekke 96) oder mutual information in Tabellen oder Absätzen erfolgen.¹⁵ Wenn beim Indexieren nur einige Begriffe verwendet werden sollen, dann kann man die Ergebnisliste nach der gewünschten Anzahl abschneiden, womit nur die besten Termkandidaten verbleiben. Das Programm gibt die Möglichkeit zur Begrenzung auf eine Anzahl zwischen 5 und 30.000 Termen. Dann kommt dem Ranking besonderes Gewicht zu, das im Versuch gute Ergebnisse lieferte.

¹⁴ Die weirdness ratio verwendet relative Häufigkeiten für die Termfrequenz und die Frequenz in der Dokumentsammlung.

¹⁵ Für die wissenschaftliche Begründung und Berechnung dieser Maße siehe STREITER 2002.

Literaturangaben

Magnar BREKKE, Johan MYKING, Khurshid AHMAD (1996). "Terminology management and lesser-used living languages: A critique of the corpus-based approach.", in: Proceedings des 4. *International Congress on Terminology and Knowledge Engineering (TKE'96)* in Wien. Hg. Peter Sandrini, Innsbruck. Seiten 179-189.

ftp://ftp.ee.surrey.ac.uk/pub/research/AI/TKE.papers/Postscript_versions/Terminology_Management.ps.gz

Béatrice DAILLE, Chantal ENGUEHARD, Christine JACQUIN, Rabaovololona Lucie RAHARINIRINA, Baholisoa Simone RALALAOHERIVONY, und Christian LEHMANN. (2000) "Traitement automatique de la terminologie en langue malgache", in: *Ressources et évaluation en ingénierie des langues*. Hg.: Karim Chibout et al., Actualités scientifiques- Universités Francophones, S. 225-242.

F. Ç. EKMEKÇIOĞLU, M.F. LYNCH, A.M. ROBERTSON, T.M.T. SEMBOK, P. WILLETT (1996). "Comparison of n-gram matching and stemming for term conflation in English, Malay, and Turkish texts", in: *Text Technology*, 6: 1-14, 1996 sowie in: *Information Research*, Vol. 2 No. 2, Oktober 1996.

<http://informationr.net/ir/2-2/paper13.html>

Christian JACQUEMIN, Didier BOURIGAULT, 2003. "Term Extraction and Automatic Indexing", in: *Handbook of Computational Linguistics*. Hg: R. Mitkov, Oxford University Press, Oxford.

<http://www.limsi.fr/Individu/jacquemi/FTP/JacBourHandbookCL.ps.gz>

Ulrich HEID (1999). "Extracting terminologically relevant collocations from German technical Texts", in: *5th International Congress on Terminology and Knowledge Engineering (TKE'99)*.

<http://citeseer.nj.nec.com/heid99extracting.html>

Munpyo HONG, Sisay FISSAHA, Johann HALLER (2001). "Hybrid Filtering for Extraction of Term Candidates from German Technical Texts", in: *Conférence TIA-2001*, Nancy.

Magnus MERKEL, Bernt NILSSON, Lars AHRENER (1994). „A phrase-retrieval system based on recurrence”, in: Proceedings of the *Second Annual Workshop on Very Large Corpora (WVLC-2)*, S. 43-56, Kyoto, 1994.

<http://www.ida.liu.se/~magne/publications/kyoto--94.ps>

Hans PAIJMANS (1997) "Gravity Wells of Meaning: detecting Information-Rich passages in Scientific Texts", in: *Journal of Documentation* 53(5), 1997, S. 520-536.

http://pi0959.kub.nl/Paai/Onderw/V-l/Content/grav_wells.html

Uwe QUASTHOFF, Christian BIEMANN, Christian WOLFF (2002). "Named Entity Learning and Verification: Expectation Maximization in Large Corpora", in: Proceedings of *CoNLL-2002*, *The Sixth Workshop on Computational Language Learning* 31 August and 1 September 2002 in association with Coling 2002 in Taipei, Taiwan.

<http://wortschatz.informatik.uni-leipzig.de/asv/publikationen/Quast-Biem-Conll-2002.pdf>

Gerhard KNORZ (1996). „Indexieren, Klassieren, Extrahieren“, in: Buder/Rehfeld/Seeger/Strauch: *Grundlagen der praktischen Information und Dokumentation*. Saur Verlag 4. Ausgabe 1996.

<http://www.iud.fh-darmstadt.de/iud/wwwmeth/publ/skript/index96/paper1.htm>

Oliver STREITER, Pei-Yun HSUEH (2000). „A case-study on example-based parsing”, in: Proceedings of *ICCLC2000, International Conference on Chinese Language Computing*, Chicago.

<http://dev.eurac.edu:8080/autoren/publs/chicago.pdf>

Oliver STREITER, Daniel ZIELINSKI, Isabella TIES, Leonhard VOLTMER (2002): „Example-based Term Extraction for Minority Languages: A case-study on Ladin”, in: Proceedings of *Soziolinguistica Y Language Planning*, St. Ulrich, 12.-14. Dezember 2002, zur Veröffentlichung durch SPELL.

<http://dev.eurac.edu:8080/autoren/publs/termex5.pdf> .

Oliver STREITER, Ernesto William DE LUCA (2003). „Example-based NLP for Minority Languages: Tasks, Resources and Tools”, in: Proceedings der *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Bats-sur-Mer, Frankreich.

http://dev.eurac.edu:8080/taln/accepted/streiter_de_luca.pdf

C. J. VANRIJSBERGEN (1983) “Information Retrieval” Chapter 1. Hg: Butterworths London.

<http://www.dcs.gla.ac.uk/Keith/pdf/Chapter1.pdf>

Alle Internetquellen zuletzt überprüft Mitte April 2003.