

# **Seminar aus Informationswirtschaft WS 2000/2001**

## **Thema: Automatische Indexierung**

**Bearbeiter: Bernd Zöchling (9351041)**

### **Übersicht:**

#### **1. Einleitung**

#### **2. Manuelle versus automatische Indexierung**

- **Manuelle Indexierung**
- **Automatische Indexierung**
- **Mischform**

#### **3. Verschiedene Suchanfragen**

- **Exact-Match**
- **Best-Match**
- **Assoziative Suche**

#### **4. Besonderheiten bei deutschsprachigen Dokumenten**

#### **5. Verschiedene Indexierungsverfahren**

- **Freitextverfahren**
- **Linguistische Verfahren**
- **Statistische Verfahren**
- **Begriffsorientierte Verfahren**

#### **6. Retrievaltests**

- **Precision**
- **Recall**
- **Einheitsmaß nach Rijsberg**

#### **7. Ein umfassendes Indexierungs-Projekt mit autom. Indexierung**

#### **8. Literaturliste**

# 1. Einleitung

Die Indexierung von Dokumenten ist eine zentrale Operation in einem Informationssystem. Es werden dabei Dokumente analysiert, um daraus Terme zur Inhaltserschließung abzuleiten. Die Güte des Indexierungsprozesses bestimmt letztendlich die Effektivität des Informationssystems bei der Recherche.

Diese Seminararbeit soll einen Überblick über die verschiedenen Möglichkeiten zur Indexierung von Dokumenten schaffen wobei nicht nur Verfahren für englischsprachige Texte, sondern auch für das Deutsche angeführt werden.

## 2. Manuelle Indexierung versus Automatische Indexierung

Es können zwei Arten von Indexierung unterschieden werden:

- **Manuelle Indexierung**
- **Automatische Indexierung**

Bei der manuellen Indexierung werden die Inhalte der Dokumente intellektuell erschlossen. Bei großem Dokumentenumfang ist diese Methode jedoch sehr zeitaufwändig. Weiters ist die Subjektivität des Indexierungsergebnisses ein Problem, wenn mehrere Indexierer einen Dokumentenbestand bearbeiten, da jeder den vorliegenden Text unterschiedlich deutet und unterschiedliche Gewichte den Termen zuordnet.

Bei der automatischen Inhaltserschließung wird der Inhalt ohne menschlichen Eingriff durch einen automatischen Indexierungsalgorithmus erschlossen. Bei manueller Indexierung können aus Kosten und Zeitgründen meist nur Titeldaten und Inhaltsverzeichnisse erfaßt werden wohingegen bei automatischem Vorgehen auch Kurzfassungen bzw. ganze Texte indexiert werden können. Es ist damit möglich, einen umfangreichen Dokumentenbestand zu indexieren. Weiters fällt das Problem der Subjektivität des Indexierungsergebnisses weg.

(Vgl. Krause 1999 S. 6-12)

Es sei hier noch erwähnt, daß neben den oben dargestellten Verfahren auch Mischformen möglich sind. Es können z.B. die Dokumente zuerst automatisch indexiert werden, die damit gewonnenen Terme dienen dann als Vorschlag für die intellektuelle Indexierung. Ein solches kombiniertes Verfahren findet beim Darmstädter Indexierungsansatz AIR/X Anwendung. (Vgl. Hennings 1994)

## 3. Verschiedene Suchanfragen

Hengartner unterscheidet in seinem Buch zwei Arten von Suchanfragen an das Informationssystem:

- **Formalisierte Suchanfragen:** (besteht aus exakt formulierten Suchbegriffen die durch Operatoren logisch verknüpft werden können z.B. 'Basel' AND 'Bern') (Exact-Match)
- **Freiformulierte Suchanfragen:** (Hier wird die Suchanfrage in einer natürlichsprachlichen Form an das System übermittelt z.B. eine frei formulierte Frage oder Dokumententitel) (Best-Match)

Jede dieser Suchanfragen benötigt bestimmte Indexierungsverfahren. Während bei der formalisierten Suchanfrage es ausreicht, Stichwörter aus den Dokumenten zu extrahieren, ist

es bei der freiformulierten Suche notwendig, auch die Bedeutung des Textes im Index zu erfassen (z.B. mit Methoden der Informationsstatistik).

Hengartner unterscheidet zwei Indexarten bei automatischen Indexierung:

- **String Index** bei exaktem Suchen (Exakt-Match)
- **Vager Index** bei vagen Suchanfragen (Best-Match: Suche nach Dokumenten die der Suchanfrage ähnlich sind)

#### **String-Index:**

Wird für den Zugriff auf Zeichenpositionen einer sehr langen Zeichenkette verwendet. Ein solcher Index kann leicht aufgebaut werden (z.B. PAT-System).

#### **Vager-Index:**

Basiert auf statistischen Methoden. Die meisten Ansätze bauen auf der Beobachtung auf, daß die Häufigkeit einzelner Wörter in der natürlichen Sprache mit der Bedeutsamkeit dieser Wörter zur Inhaltsbeschreibung korreliert. Es wird damit indirekt versucht – ohne den Inhalt zu verstehen – Dokumenteninhalte zu bestimmen.

Problem: Der Index kann sehr Umfangreich werden (Speicherbedarf, Zugriffsgeschwindigkeit)

(Vgl. Hengartner1997 S. 116-117, 128-135)

### **Assoziative Suche**

Im Gegensatz zur booleschen Suche mit exakten Suchbegriffen arbeitet die assoziative Suche mit ganzen Dokumenten als Suchargumente. Der Nutzer benötigt daher zu Beginn des Suchvorganges ein Musterdokument.

Es können zwei Varianten unterschieden werden:

- **Suche nach Schlagwörtern**
- **Suche nach Zitaten**

Bei der ersten Variante wird das Musterdokument automatisch indexiert, die Stichwörter werden zu Suchargumenten. Danach erfolgt der Suchvorgang.

Bei der zweiten Variante werden Zitate aus dem Musterdokument als Suchargumente herangezogen (nur bei Dokumenten mit Zitaten möglich).

Die Grundidee liegt darin, daß zwei Dokumente (Musterdokument und Suchergebnis), die gleiche Zitate enthalten, miteinander verwandt sind. Die Sortierung der gefundenen Dokumente nach der Relevanz ergibt sich aus der Anzahl der gemeinsam vorkommenden Zitate. Probleme entstehen jedoch, wenn Autoren nach unterschiedlichem Schema zitieren. (Vgl. Stock 2000 S. 166-168)

## **4. Probleme bei der autom. Indexierung deutschsprachiger Texte**

Werden bei der Inhaltserschließung einfach alle Wörter, die aus dem Text extrahiert wurden, indexiert (=Freitextverfahren) dann ergeben sich keine Probleme bei den verschiedenen Sprachen. Sollen jedoch linguistische Verfahren (Wortstammanalyse, Phrasenerkennung, Mehrwortbegriffszerlegung, ...) zur Anwendung kommen, dann muß auf die grammatikalischen Eigenschaften der Sprachen Rücksicht genommen werden.

Die englische Sprache hat eine einfache Struktur, die Flexionsformen können einfach durch Regeln beschrieben werden (mittels Suffixliste).

Die deutsche Sprache hat jedoch viele Ausnahmen und Eigenheiten, die Flexionsformen können kaum durch Regeln beschrieben werden.

Die meisten Systeme automatischer Indexierung arbeiten mit englischsprachigen Dokumenten. Die Realisierung von Systemen zur automatischen Indexierung von deutschsprachigen Texten ist wesentlich komplizierter. Während erstere Systeme auf Regeln aufsetzen werden bei letzteren wörterbuchbasierte Verfahren angewandt.

## 5.1 Freitextverfahren (Vgl. Buder 1996)

Bei diesem Verfahren werden alle im Text vorkommenden Wortformen unabhängig von der Flexionsform und Sprache (ausgenommen Funktionswörter aus Stoppwortlisten) erschlossen. Dieses Verfahren schafft oft die indexierte Dokumentenbasis für Suchanfragen mit Boole'schen Operatoren (AND, OR, ..).

## 5.2 Linguistische Verfahren (Vgl. Stock 2000 S. 149-157)

Linguistische Verfahren können in 2 Gruppen geteilt werden: (Vgl. Hennings 1994)

- Linguistisch-wortorientierte Verfahren
- Linguistisch-syntaxorientierte Verfahren auf Satz- u. Phrasenebene

Das erste Verfahren arbeitet auf Wortebene und behandelt Probleme der Wortformenvarianten und der Wortbildung (Komposita). Die genaue Funktion eines Systems für deutschsprachige Texte wird weiter unten erklärt (PASSAT).

Das zweite Verfahren arbeitet auf Satzebene. Es wird hier versucht, mit Hilfe einer syntaktischen Analyse Beziehungen zwischen den einzelnen Wörtern herauszufinden und für die Indexierung zu verwenden. Es werden Mehrwortgruppen durch Verbindung von Adjektiv + Substantiv oder Adjektiv + Präposition + Substantiv gebildet (Bsp: ... die Diplomarbeit wurde begutachtet => begutachtete Diplomarbeit). Die Universität des Saarlandes hat dazu den Ansatz CTX (Computergestützte Texterschließung) entwickelt.

Informationslinguistische Methoden haben die Aufgabe, nicht sinntragende Wörter zu eliminieren, grammatikalische Flexionsformen auf eine Grundform zu bringen, aus mehreren Termen bestehende Phrasen zu erkennen und Pronomina den jeweiligen Nomen zuzuordnen.

Zu Beginn müssen die einzelnen Wörter bzw. Zeichenfolgen mit n Elementen aus dem Dokument isoliert werden. Die Wörter können mit Hilfe von Leerzeichen und Satzzeichen isoliert werden. Zeichenfolgen werden in sgn. „**N-Gramme**“ (Tupel) mit einer bestimmten Anzahl von Zeichen zerlegt.

Bsp: „Widerspruchsfreiheit“ mit n=5

Wider  
 iders  
 dersp  
 erspr  
 ... reihe ...  
 bewei  
 ews

Viele der Tupel sind irrelevant. Der Suchbegriff „Freiheit“ (Tupel „frei“) findet das Wort „Freiheitsbeweis“ genauso wie der Suchbegriff „Beweis“ (Tupel „bewei“). Das Verfahren ist jedoch auch fehleranfällig: So wird bei der Suche nach „Reihe“ (Tupel „reihe“) das Wort „Freiheitsbeweis“ ebenfalls als Treffer betrachtet.

## Stoppwörter

Nachdem die Wörter des Dokumentes isoliert worden sind, werden mit Hilfe einer Stoppwortliste häufig vorkommende Funktionswörter eliminiert, die zur Inhaltserschließung keinen Nutzen bringen (~50% des Dokuments).

a	amongst	becomes
about	an	becoming
after	and	been
afterwards	another	before
again	any	beforehand
against	anyhow	behind
all	anyone	being
almost	around	between
alone	as	both
along	at	...

## Stoppwortliste (Ausschnitt)

## Wortstammanalyse

Im zweiten Schritt werden durch Reduktionsalgorithmen Endungen entfernt, um damit zur Grundform zu gelangen. Dazu findet eine Suffixliste Anwendung.

e	em	est
eable	ence	et
eal	ency	eta
ectual	eness	etion
ed	ening	etic
edly	ent	
edness	entia	
ee		
eer		...

## Suffixliste (Ausschnitt: Anfangsbuchstabe E)

Beispiel: listened → listen  
 Problem: ringing → ring → r  
 Lösung: eine bestimmte Mindestlänge darf nicht unterschritten werden

## Phrasenerkennung

Durch die Phrasenerkennung sollen Wörter, die zusammen gehören, als Einheit betrachtet werden. Das ist z.B. bei Personennamen („Wolfgang Clement“) und Organisationsnamen („WU Wien“) notwendig.

Das geschieht durch zwei Methoden:

- durch Abgleich mit Listen
- durch Analyse des gemeinsamen Auftretens innerhalb der Dokumente

Die Listen können bereits die Phrasen enthalten, oder sie enthalten Indikatorbegriffe, mit deren Hilfe sich die Phrasen aufspüren lassen.

Indikatorbegriffe für Unternehmensnamen		
Co	Corp.	Ltd
Co.	Corporation	Ltd.
Company	Inc	
Corp	Inc.	...
Indikatorbegriffe für Namen von Organisationen		
Agency	Center	College
Association	Club	Commission
Board		...

### Indikatorbegriffe

Wird beispielsweise ein Vorname, der in der Indikatorliste enthalten ist, im Dokument gefunden, dann wird der Vorname gemeinsam mit dem folgenden Wort (=Nachname) als Phrase betrachtet.

Die zweite Möglichkeit der Phrasenerkennung ist eine Erweiterung der Stoppwortliste um Adverbien, Hilfsverben und weitere Verben (z.B. „what“, „is“, „newly“, „by“, ...)

## Identifikation von Textklumpen:

**Citing** what is **called** newly **conciliatory comments** by the **leader** of the **Irish Republican Army's political wing**, the **Clinton Administration** announced **today** that it would issue him a **visa** to attend a **conference** on **Northern Ireland** in **Manhattan** on **Tuesday**.

Die zwischen den Stoppwörtern stehenden Terme sind durch Fettdruck markiert. Für die Indexierung sind jedoch nur die fettgedruckten Mehrworttextklumpen relevant (einzelne Wörter fallen weg). Wortkombinationen werden bei mehrmaligem Vorkommen als Phrase betrachtet.

## Synonyme

Sind Wörter mit unterschiedlicher Bezeichnung jedoch mit gleicher Bedeutung.

- Bsp:
- verschiedene Schreibvarianten
  - unterschiedliche Schreibweise („ten“ -- „10“)
  - Abkürzungen und Vollformen („N.J.“ -- „New Jersey“)
  - Echte Synonyme („Samstag“ -- „Sonntag“)

Bei der Suche dem Term „Samstag“, müssen als Ergebnis auch jene Dokumente aufgelistet werden, die den Term „Sonntag“ enthalten. Hier findet ein Thesaurus als Synonymwörterbuch Anwendung.

## Pronomina-Analysen

Pronomina, die an die Stelle ihrer Nomen rücken, müssen beachtet werden, da sie für statistische Berechnungen (Worthäufigkeit) notwendig sind.

„The president has a girl friend, but he doesn't love her“.

In diesem Fall wird „he“ dem Wort „präsident“ und „her“ der Phrase „girl friend“ zugeordnet. Präsident und girl friend kommen daher jeweils zweimal vor.

## PASSAT – Vertreter des Linguistisch-wortorientierte Verfahrens

(Vgl. Hennings 1994)

PASSAT ist ein von der Firma Siemens entwickeltes wörterbuchbasiertes Verfahren. Es unterstützt mehrere Sprachen, ist hauptsächlich für das Deutsche gedacht.

### Funktionalität:

- Reduzierung von Wortformen auf Grundformen
- Kompositazerlegung in Komponenten
- Elimination von Nicht-Deskriptoren
- Erkennen einfacher Formen von Phrasen
- Rechtschreibfehlererkennung

Da PASSAT auch für deutschsprachige Texte geeignet ist, verwendet es – wie zuvor erwähnt – kein Regelwerk sondern ein Wörterbuch zur Herleitung der Stammform. Im Wörterbuch wird entschieden, wie mit den einzelnen Wörtern umgegangen werden soll. Es können jedoch nur Wörter bearbeitet werden, die im Wörterbuch enthalten sind. Beispieleintrag im Wörterbuch:

- Stammwort:	Antrag	... bei allen Wortformen gemeinsamer Wortanfang
- Substitutionstyp:	K	... Festlegung ob Substitut statt/zusätzlich zur Grundform oder nur die Grundform Ergebnis der Indexierung ist
- Umlautung:	U	... Soll Umlautung bei Plural zurückgenommen werden
- Bindungsliste:	[ - , s ]	... Liste von Fugenmorphemen bei Kompositazerlegung
- Endungsliste:	[-,e,en,es,s]	... Liste aller Endungen die sich an die Stammform anschließen können
- Substitute:		... Liste von mgl. Substituten

### Arbeitsweise Bsp.1:

Beispieltext: ... den Anträgen ...

1. Umlaut zurücknehmen (=>Antragen)
2. Überprüfen ob Wort mit einer Stammform beginnt (Ja)

3. Prüfen ob Endung in Endungsliste enthalten (en => Ja)
4. Abtrennen der Endung (=>Antrag)
5. Anfügen der ersten Endung aus Endungsliste (=> -)

Ergebnis für Indexierung: Antrag

Punkt 4 und 5 sind deshalb notwendig, um die korrekte Endung jeder Stammform zu erhalten (z.B. Küchen – [en] aus Endungsliste => Küch (Umlaut hier anders behandelt als bei Anträge) + [e] aus Endungsliste für Küche => Küche).

### Arbeitsweise Bsp.2:

Beispieltext: ... vielen Antragsformularen ...

Anmerkung: dieses Wort hätte bei Bsp.1 kein Ergebnis gebracht da das Wort Antragsformular nicht im Wörterbuch als Stammform enthalten ist. Es kann jedoch das zusammengesetzte Wort durch eine Kompositazerlegung aufgespalten werden und in der Folge die einzelnen Wörter wie in Bsp.1 indexiert werden.

1. Wort zerlegen (s aus Bindungsliste [ - , s ])
2. Mit einzelnen Wörtern erneuter Durchlauf wie in Bsp.1
3. Bei Mißerfolg => Eintrag in Liste unbekannter Zeichenfolgen

Ergebnis für Indexierung: Antrag, Formular

## 5.3 Statistische Verfahren

**Allgemeines:** (Vgl. Stock 2000 S. 149-157)

Informationsstatistik ermittelt Worthäufigkeiten in Dokumenten und ermittelt mit diesen Daten Wortgewichtungen, Dokumentenhäufigkeiten und daraus abgeleitet ein Ranking der Dokumente nach Relevanz. Der Begriff Wort kann dabei enger oder weiter gesehen werden: jedes vorkommende Wort, nur der Wortstamm einzelner Wörter, Berücksichtigung von Phrasen, zusätzlich Zählung der zugehörigen Pronomina.

### Worthäufigkeiten:

Alle Wörter im einzelnen Dokument und die Summe der Wörter in der gesamten Datenbasis werden gezählt. Die ermittelten Werte sind Basis für weitere Berechnungen.

Annahmen:

L	Anzahl der Wörter im Dokument j
n	Anzahl der Dokumente, in denen ein bestimmtes Wort i vorkommt
N	Gesamtzahl der Dokumente in der Datenbasis
i	bestimmtes Wort im Text
j	bestimmtes Dokument
WDF	Within-document frequency weight (dokument.spezifisches Wortgewicht)
IDF	Inverse Dokumenthäufigkeit

Wenn nur einfache Worthäufigkeiten zur Bestimmung der Relevanz herangezogen werden würden, dann käme es zu einer Bevorzugung von längeren Texten. Daher müssen Wortgewichtungen berechnet werden.



### **Dokumentenspezifische Wortgewichtung:**

Hier werden die relativen Häufigkeiten von Textwörtern bestimmt (Quotient aus Häufigkeit eines Wortes und der Gesamtmenge der Wörter in einem Dokument).

Berechnungsformel nach Donna Harman:

$$WDF(i) = (Id[Freq(i,j) + 1] / Id L$$

Je häufiger das Wort i im Dokument j vorkommt, desto größer ist WDF.

### **Gewichtung nach Position im Text**

Hier werden die Wörter nach der Position im Text (im Titel, Abstract, am Anfang/Ende im Dokument, ...) bewertet. Z.B. könnte ein Wort, das im Titel vorkommt, höher bewertet werden als ein Wort, das im Text enthalten ist.

### **Inverse Dokumenthäufigkeit:**

Bei der inversen Dokumenthäufigkeit ist ein Wort umso wichtiger, je weniger Dokumente dazu in der Datenbasis vorhanden sind.

$$IDF(i) = (\log_2 N/n) + 1$$

### **Wortabstand:**

Wenn in der Suchanfrage mehrere Wörter enthalten sind dann kann anhand des Wortabstandes dieser Wörter im Dokument ein Ranking erstellt werden. Kommen Wortpaare mehrmals vor, dann wird ein arithmetisches Mittel berechnet.

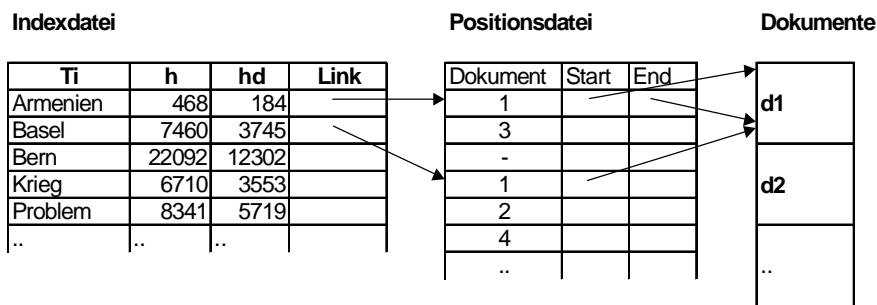
### **Ranking nach Relevanz:**

Durch Multiplikation der dokumentenspezifischen Wortgewichtung mit der inversen Dokumentenhäufigkeit ( und bei Berücksichtigung der Position im Text auch mit dem entsprechenden Gewichtungsfaktor P) ergibt sich ein Gewichtungswert für das Dokument j, mit dessen Hilfe ein Ranking der Suchergebnisse realisiert werden kann. Je höher das Gewicht, desto höher ist das Ranking.

$$\text{Gewicht}(i,j) = WDF(i) * IDF(i) (* P)$$

Diese Ranking kann bei der zuvor beschriebenen Best-Match-Methode zur Sortierung der Suchergebnisse nach Relevanz verwendet werden.

Beispiel eines Index mit enthaltenen Worthäufigkeiten: (Vgl. Hengartner 1997)



h ... Häufigkeit in allen Dokumenten

hd ... Häufigkeit des Begriffen im Dokument d

Gewicht  $w = hd/h$

Jedem Indexterm wird die Häufigkeit in der gesamten Datenbasis und die Häufigkeit in den einzelnen Dokumenten d zugeordnet. Weiters wird in der Positionsdatei das Vorkommen des Wortes in den einzelnen Dokumenten erfaßt.

## Anwendung statistischer Methoden im Vektormodell

Das von Salton entwickelte Vektormodell basiert auf einen n-dimensionalen Vektorraum.

Jeder Indexterm entspricht dabei einer Dimension. Jedes Dokument (= Summe aus einzelnen Indextermen) repräsentiert einen Punkt in diesem Raum.

Jeder Indexterm des Dokumentes wird nach Häufigkeit gewichtet:  $gt = \text{Häufigkeit des Begriffes } t \text{ im Dokument} / \text{Anzahl Dokumente mit } t$ .

Jede Suchanfrage wird in ihre einzelnen Suchbegriffe aufgespaltet und kann dadurch ebenfalls als Punkt im Raum dargestellt werden.

Als Suchergebnis sind alle Dokumente relevant, deren Punkte in der Nähe des Punktes der Suchanfrage liegen. Der Abstand zu diesem Punkt kann als Relevanzmaß betrachtet werden.

## 5.4 Begriffsorientierte Verfahren

Hier werden, im Unterschied zu normalen Stichwortverfahren, die Wörter aus den Dokumenten nicht nur selektiert, normiert und gewichtet sondern es wird auch versucht, die Bedeutung der Wörter zu berücksichtigen. Ziel ist daher das Textverstehen (Wort => Begriff).

### Praktische Realisation:

Grundsätzlich kann dies durch manuelle Indexierung erreicht werden da der Indexierer Zusammenhänge im Text erkennen kann.

Bei automatischer Indexierung behilft man sich durch die Anwendung von Methoden der künstlichen Intelligenz.

In der Folge werde ich zwei Ansätze zur begriffsorientierten automatischen Texterschließung beschreiben.

## TCS: Text Categorization Shell

Wurde von der Carnegie Group entwickelt und ist dreistufig aufgebaut:

1. Begriffsidentifikation (Regelbasis)
2. Deskriptor-Hypothesenbildung
3. Indexierungsentscheidung

### Das System:

TCS besteht aus einem Laufzeitsystem das zur Indexierung dient und einer Entwicklungsumgebung zur Erzeugung/Änderung/Testen der Regeln.

Damit das System überhaupt anwendbar ist, muß zuvor die Regelbasis geschaffen werden. Es wird hier für jeden einzelnen Begriff eine Regel aufgestellt. (siehe Punkt 1). Es werden auch Beziehungen zwischen diesen einzelnen Begriffs-Regeln aufgestellt (siehe Punkt 3).

### Regelentwicklung:

Zu Beginn wird manuell indexiert, danach wird automatisch (mit der Regelbasis) indexiert, um Abweichungen festzustellen und die Regeln überarbeiten zu können. Dieser Vorgang wird solange wiederholt, bis das Ergebnis den Vorstellungen entspricht.

#### ad 1) Begriffsidentifikation (Regelbasis)

Begriffe werden als komplexe Muster von Wörtern definiert.

Bsp.: Identifizierung des Begriffes Gold als Metall / nicht als Einheit des Goldmarktes

((not((non-)! stockpiled)) **gold** (not(&skip; 4 (index !! stock +N !! reserve +N))))

**gold** steht für den Begriff Gold als Metall wenn nicht "non" oder "stockpiled" dem Wort **gold** im Text vorausgehen und innerhalb der folgenden 4 Wörter nach dem Wort **gold** im Text die Wörter "index", "stock", "reserve" nicht folgen.

#### ad 2) Deskriptor-Hypothesenbildung

Regeln stellen die zuvor identifizierten Begriffe als Hypothese für eine Deskriptor-Zuteilung auf:

(hypothese **gold** (or[**gold** scope: head 1][**gold** 3] (and[**gold** 2][precious-metal metal 2])))

der Deskriptor gold kommt dann in Frage wenn ...

- der Begriff **gold** im Titel mit Gewicht von mindestens 1 vorkommt, oder
- mit einem Gewicht von min. 3 irgendwo im Text vorkommt, oder
- nur mit Gewicht von min. 2 im Text vorkommt, jedoch zusätzlich precious-metal oder metal mit dem Gewicht von min. 2 vorkommt

#### ad 3) Indexierungsentscheidung

Die zuvor aufgestellten Hypothesen werden durch eine weitere Klasse von Regeln verworfen oder bestätigt (Deskriptor/kein Deskriptor)

(when-hypothesized gold	
(if actions:	(disconfirm gold)
test	[not gold 2]
actions	(disconfirm gold)
test	(hypothesized-cats money-fx)
actions	(disconfirm gold)
default-actions	(confirm gold)
	(confirm precious-metal)

Der Deskriptor gold wird verworfen wenn der Begriff gold nicht das Gewicht von 2 erreicht ((disconfirm gold) wenn [not gold 2])

Wenn eine Hypothese für die Deskriptorzuteilung für den Begriff "money-fx" besteht dann wird der Deskriptor gold ebenfalls verworfen ((disconfirm gold) wenn (hypothesized-cats money-fx))

Wenn die oben beschriebenen Fälle nicht eintreten (gold ablehnen) dann kommt die Standardaktion (default-actions) zur Anwendung wobei die Begriffe gold und precious-metal zur als Deskriptor gewählt werden ((confirm gold), (confirm precious-metal)).

## **Darmstädter Ansatz: AIR/X**

### **Gemeinsamkeiten mit TCS:**

- benötigt ebenfalls manuelle Indexierung
- ist ebenfalls Mehrstufig
- benützt Regeln und Hypothesen für Deskriptorenzuteilungen

### **Unterschied:**

- automatisches Vorgehen bei Regelentwicklung
- daher nur einfach strukturierte Regeln möglich
- benötigt umfangreichere manuell indexierte Datenbasis (400.000 Dokumente, 20.000 Deskriptoren)
- aber Zeit und Kostenersparnis gegenüber TCS

### **Einsatz:**

Das System wird zur Produktion der Datenbasis PHYS (für englischsprachige Dokumente im Bereich Physik mit ca. 10.000 Dokumente/Monat). Es wird jedoch als kombiniertes Verfahren eingesetzt (automatische Indexierung macht Deskriptorvorschlag, Auswahl erfolgt manuell).

## **6. Retrievaltests**

Um die Qualität der automatischen Indexierung beurteilen zu können bzw. verschiedene Verfahren untereinander vergleichen zu können werden sog. Retrievaltests durchgeführt.

Im Artikel "Automatische Indexierung für Online-Kataloge: Ergebnisse eines Retrievaltests" von Lepsky wird die Durchführung eines Retrievaltest beschrieben:

Es wurde dazu ein Testdatenbestand aus Titeldaten automatisch indexiert (40.000 Titeln). Dabei kamen Grundformerzeugung, Dekomposition, Derivation und teilweise semantische Relationen zur Anwendung. Bei der Testdurchführung wurden mehrere Indizes (reiner

Stichwortindex und reiner Stichwortindex ergänzt um Ergebnisse der autom. Indexierung) angelegt.

Die Ergebnisse wurden anhand Precision, Recall und dem Einheitsmass nach Rijsbergen gegenübergestellt.

**Precision** ist ein Maß für die Genauigkeit einer Recherche, es wird dabei die Zahl der gefundenen relevanten Dokumente durch die Zahl der überhaupt gefundenen Dokumente dividiert (ist daher Maß für die Treffgenauigkeit)

$PRECISION = (\text{gefundene relevante D.}) / (\text{gefundene relevante D.} + n. \text{relevante D.})$

**Recall** ist das Verhältnis zwischen der Menge der gefundenen relevanten Dokumente und der Gesamtzahl der relevanten Dokumente in der Datenbasis (ist daher Maß für den quantitativen Erfolg der Suche).

$RECALL = (\text{gefundene relevante D.}) / (\text{relevante D. in Datenbasis})$

Das allgemeine Problem bei der Indexierung ist der Tradeoff zwischen Precision und Recall.

Van Rijsberg hat eine Formel entwickelt, die Precision und Recall vereinigt und darüber hinaus auch eine Gewichtung der Maße ermöglicht (mit Beta).

**Einheitsmaß nach Rijbergen:**  $e = ((\text{Beta}^2 + 1) * p * r) / (\text{Beta}^2 * p * r)$

Die ermittelten Einheitsmaße bewegen sich zwischen 0 und 1.

## 7. KASCADE – MILOS

**K**atalogerweiterung durch **S**canning und **A**utomatische **D**okument**E**rschließung ist ein an der UNI Düsseldorf durchgeführtes Projekt mit dem Ziel, die bis zu diesem Zeitpunkt auf reine Titeldaten beschränkten Datenbasis der Bibliothek um inhaltsbezogene Daten zu erweitern. Damit sollen bessere Suchmöglichkeiten eröffnet werden. Die Datenbasis KASCADE beinhaltet Zeitschriften aus dem Fachgebiet Jura.

Es wird dabei OCR-Texterkennung für nicht maschinenlesbare Texte verwendet.

Das System baut auf das linguistisch-syntaxbezogenen Verfahren IDX/MILOS, das von Prof. Zimmermann an der Universität des Saarlandes entwickelt wurde, auf und basiert auf einem Wörterbuch (deutschsprachige Texte möglich). MILOS ist eine bibliothekspezifische Erweiterung des Indexierungssystems IDX.

Es beinhaltet ebenfalls statistische Verfahren (Wortgewichtung) zur Relevanzbestimmung von Suchergebnissen

Weiters beinhaltet es semantische Verfahren die es ermöglichen, durch lexikalische Einträge im Wörterbuch die Bedeutung der Wörter zu berücksichtigen und damit Dokumente automatisch verschiedenen Themenbereichen zuordnen zu können (Themen-Aspekt-Identifizierung).