Klassifikation

Grundsätzlich versucht man bei der Klassifikation Dokumente mit gemeinsamen Eigenschaften zusammenzufassen um durch eine Vorauswahl der Klasse den Suchraum zu verkleinern.

Beispiel: **Dezimalklassifikation von Dewey (1870)**

Bei der Dezimalklassifikation wird das gesamte menschliche Wissen in 10 Hauptgruppen unterteilt. Jede dieser Hauptgruppen oder Klassen kann dann wiederum 10 Unterklassen besitzen usw.

0 Allgemeine Werke

1 Philosophie

2 Religion, Theologie

3 Sozialwissenschaften
- 30 Soziologie
- 31 Statistik
- 32 Politik
- 33 Volkswirtschaft
- 34 Recht
- 35 Öffentliche Verwaltung
- 36 Sozialverbände, Versicherungen

4 Sprachen

5 Mathematik

6 Technik, Medizin

7 Kunst

8 Literatur

9 Geschichte

Retrieval Verfahren

Retrieval Verfahren sind die Methoden mit denen die gespeicherte Information gefunden bzw. wiedergewonnen wird.
Um eine Suche zu starten, muß eine **Anfrage** an das System gestellt werden. Da die natürliche Sprache in vielen Fällen ungenau oder mehrdeutig ist braucht man wie auch in der Programmierung spezielle Sprachen mit eindeutig definierter Syntax und Semantik. Anfragen an ein Dokumenten-Retrievalsystem werden also in sogenannten **Anfragesprachen** formuliert. Es existieren verschiedene mathematische Modelle als Grundlage für solche Anfragen.

Beim **Booleschen** Retrieval besteht die Anfrage aus einer Menge von Deskriptoren die durch logische Operatoren verknüpft sind. Als Antwort auf eine Anfrage wird die Menge der Dokumente geliefert die mindestens einen gemeinsamen Deskriptor mit der Anfrage hat. Man nennt Sprachen für boolesche Systeme **logisch-deskriptororientiert** oder auch **Deskriptorsprachen.**

Beim **Linearen pseudoboolesches Retrieval** geht man davon aus, daß ein Dokument dem Informationsbedürfnis des Benutzers um so mehr entspricht , je mehr gemeinsame Deskriptoren die Suchfrage mit dem Dokument hat.

Retrieval mit Ähnlichkeitsmaßen
Generell findet hier nicht nur eine Prüfung auf Übereinstimmung des Deskriptors der Suchfrage mit der Dokumentbeschreibung statt sondern zusätzllich eine Prüfung der Wichtigkeit bzw. Relevanz des Deskriptors innerhalb des entsprechenden Dokuments. Anders als beim booleschen Verfahren können beim Retrieval mit Ähnlichkeitsmaßen die Elemente der Dokument-Deskriptormatrix nicht nur die Wahrheitswerte 0 und 1 annehmen sondern beliebige Zahlenwerte sein die generell das Gewicht der einzelnen Deskriptoren in den jeweiligen Dokumenten repräsentieren. Diese Zahlen werden von einer Ordnungsfunktion des Retrievalsytems bewertet. Die Ordnungsfunktion weist jedem Dokument im Dokumentenbestand einen Wert zu der der Ähnlichkeit bzw. der Zugehörigkeit des Dokuments zur Anfrage entspricht. Als Antwort liefert das Retrievalsystem eine nach diesem Ähnlichkeitsmaß sortierte Dokumentenliste.

Vorraussetzung ist dabei das beim Indexieren jeder Deskriptor mit einem Wert versehen wird der seine Wichtigkeit in Bezug auf das Dokument widerspiegelt. Zum Gewichten der Terme werden verschiedene mathematische Modelle verwandt.

(Sei $d=\{x_1,....,xn\} \in$ die Menge aller Dokumentbeschreibungen,

$s=\{y_1,....,yn\} \in$ die Menge aller Suchfragenbeschreibungen)

- **Inneres Produktmaß R_i**

$$R_i\,(d_i,s) = \sum_{j=1}^{n} x_j \cdot y_j$$

(Geometrische Interpretation: Teilung des Dokumentenraums durch Hyperebene)

- **Taniomotomaß R_t:**

$$R_t\,(d_i,s) = \frac{\sum_{j=1}^{n} x_j \cdot y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2 + \sum_{j=1}^{n} x_j \cdot y_j}$$

(Geometrische Interpretation: Alle gefundenen Dokumente liegen innerhalb einer Kugel)

- **Cosinusmaß R_c:**

$$R_c\,(d_j,s) = \frac{\sum_{j=1}^{n} x_j \cdot y_j}{\sqrt{\sum_{j=1}^{n} x_j^2 \cdot \sum_{j=1}^{n} y_j^2}}$$

(Geometrische Interpretation: Alle gefundenen Daten liegen innerhalb eines Kegelmantels)

- **Maß von Bennet und Spiegel**
- **Maß von Maron und Kuhns**
- **Boolesches und modifiziertes boolesches Maß**

Beispiel eines Dokumentenbestandes, seiner Deskriptoren ,einer einfachen Anfragesprache, sowie einer Anfrage inclusive ihrer Auswertung [1]

Dokumentenbestand mit Deskriptoren:

Dokument D_1 besitzt Deskriptoren (SCHLANGE,GIFT)

Dokument D_2 besitzt Deskriptoren (FROSCH,SCHLANGE)

Dokument D_3 besitzt Deskriptoren (SCHLANGE,FROSCH,GIFT)

Dokument D_4 besitzt Deskriptoren (FROSCH)

Dokument D_5 besitzt Deskriptoren (FROSCH)

Dokumentenorientierte Darstellung:

D_1: {SCHLANGE, GIFT}

D_2: {FROSCH, SCHLANGE}

D_3: {FROSCH, SCHLANGE, GIFT}

D_4: {FROSCH}

D_5: {FROSCH}

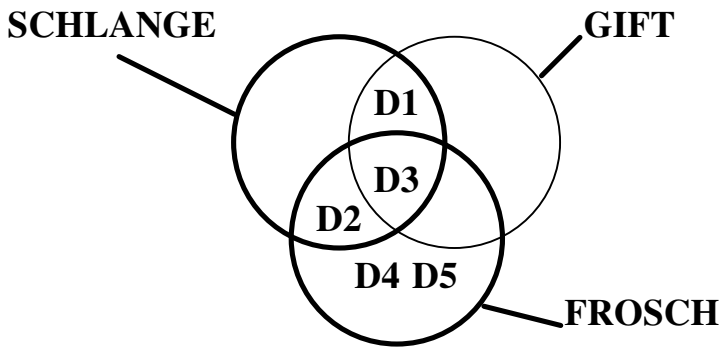
Deskriptororientierte Darstellung:

SCHLANGE = $\{D_1, D_2, D_3\}$

FROSCH = $\{D_2, D_3, D_4, D_5\}$

GIFT = $\{D_1, D_3\}$

Mengenorientierte Darstellung:



Darstellung als Dokument-Deskriptor-Matrix:

	D_1	D_2	D_3	D_4	D_5
SCHLANGE	1	1	1	0	0
FROSCH	0	1	1	1	1
GIFT	1	0	1	0	0

Syntaktische und semantische Definition einer Abfragesprache:

Das Alphabet einer Abfragesprache S bestehe aus:

1. Deskritorsymbolen D_1,D_2,D_3
2. Logischen Symbolen: NICHT,UND,ODER
3. Technischen Symbolen (,)

Rekursiver Aufbau der gültigen Suchfragen:

1. Deskriptoren D_i (i=1,2,3) sind gültige Suchfragen
2. Wenn q_1 eine Anfrage ist dann ist auch (NICHT q_1) eine Anfrage.
3. Wenn q_1 und q_2 Anfragen sind dann ist auch (q_1 UND q_2) eine Anfrage.
4. Wenn q_1 und q_2 Anfragen sind dann ist auch (q_1 ODER q_2) eine Anfrage.
5. Das sind alle Anfragen.

Um den syntaktischen Anfragen eine Bedeutung zuzuweisen wird eine klar definierte Semantik benötigt. Diese stellt die Abbildung dar die den Übergang von nichtsprachlichen zu sprachlichen Elementen herstellt also die Beziehung zwischen Sprache und Realität.

Die syntaktische Anfrage **Q** wird mit Hilfe der Interpretation ϑ eindeutig in die Potenzmenge der Dokumentenmenge **Pot(Δ)** abgebildet:

$$\vartheta : \mathbf{Q} \rightarrow (\Delta)$$

Die Semantik wird rekursiv über den Aufbau der Anfrage definiert:

1. $\vartheta(\mathbf{d_i}) \subseteq \Delta, (i=1,2,3)$
2. $\vartheta(\mathbf{NICHT\ q_1}) \equiv \Delta - \vartheta(\mathbf{q_1})$
3. $\vartheta(\mathbf{q_1\ UND\ q_2}) \equiv \vartheta(\mathbf{q_1}) \cap \vartheta(\mathbf{q_2})$
4. $\vartheta(\mathbf{q_1\ ODER\ q_2}) \equiv \vartheta(\mathbf{q_1}) \cup \vartheta(\mathbf{q_2})$

Die Abbildung ϑ ordnet jedem Deskriptor d_i (i=1,2,3) eindeutig eine Teilmenge von Δ zu.

Interpretation einer Anfrage:

Zusammen mit Operatoren UND, ODER, NICHT und den Deskriptoren können wir nun eine Anfrage syntaktisch formulieren. Beispielsweise:

„ (NICHT GIFT) UND (SCHLANGE ODER FROSCH)“

Die Zeichenkette muß nun vom Dokumentensuchsystem durch semantische Interpretation so abgebildet werden, daß die Antwort eine Dokumentenmenge ist.

ϑ ((NICHT GIFT) UND (SCHLANGE ODER FROSCH))

= ϑ (NICHTGIFT) $\cap$ ϑ(SCHLANGE ODER FROSCH)

= ϑ (NICHT GIFT) $\cap$ (ϑ(SCHLANGE) $\cup$ ϑ(FROSCH))

= $(\Delta - \{D_1,D_3\}) \cap (\{D_1,D_2,D_3\} \cup \{D_2,D_3,D_4,D_5\})$

= $\{D_2,D_4,D_5\} \cap \{D_1,D_2,D_3,D_4,D_5\}$

= $\{D_2,D_4,D_5\}$

Als Ergebnis der Anfrage werden die Dokumente $\{D_2,D_4,D_5\}$ geliefert.

Wer sich über die genannten Modelle und Methoden tiefergehend informieren möchte dem ist vielleicht mit der Diplomarbeit

„Mathematische Modellierung und
funktionellle Beschreibung von
Dokumenten Retrieval Systemen“

von Clemens Süß,Joachim Lechermann
Informatik TU Berlin Juli 1981
geholfen.

[1] Referat „Dokumentensuche“ aus „Suche in Wissensbasierten Systemen“ SS94