

Stefan Wagner

Barrierefreie und thesaurusbasierte
Suchfunktion für das Webportal der Stadt
Nürnberg

Diplomarbeit

Bibliografische Information der Deutschen Nationalbibliothek:

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de/> abrufbar.

Dieses Werk sowie alle darin enthaltenen einzelnen Beiträge und Abbildungen sind urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsschutz zugelassen ist, bedarf der vorherigen Zustimmung des Verlanges. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen, Auswertungen durch Datenbanken und für die Einspeicherung und Verarbeitung in elektronische Systeme. Alle Rechte, auch die des auszugsweisen Nachdrucks, der fotomechanischen Wiedergabe (einschließlich Mikrokopie) sowie der Auswertung durch Datenbanken oder ähnliche Einrichtungen, vorbehalten.

Copyright © 2007 Diplomica Verlag GmbH
ISBN: 9783836607612

Stefan Wagner

**Barrierefreie und thesaurusbasierte Suchfunktion für
das Webportal der Stadt Nürnberg**

Stefan Wagner

Barrierefreie und thesaurusbasierte Suchfunktion für das Webportal der Stadt Nürnberg

Stefan Wagner

**Barrierefreie und thesaurusbasierte Suchfunktion für das Webportal der Stadt
Nürnberg**

ISBN: 978-3-8366-0761-2

Druck Diplomica® Verlag GmbH, Hamburg, 2008

Zugl. Georg-Simon-Ohm-Fachhochschule Nürnberg, Nürnberg, Deutschland,
Diplomarbeit, 2007

Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses Werkes ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland in der jeweils geltenden Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtes.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Die Informationen in diesem Werk wurden mit Sorgfalt erarbeitet. Dennoch können Fehler nicht vollständig ausgeschlossen werden, und die Diplomarbeiten Agentur, die Autoren oder Übersetzer übernehmen keine juristische Verantwortung oder irgendeine Haftung für evtl. verbliebene fehlerhafte Angaben und deren Folgen.

© Diplomica Verlag GmbH

<http://www.diplomica.de>, Hamburg 2008

Printed in Germany

Kurzfassung

Im Internetportal der Stadt Nürnberg wurde in einer vorausgehenden Diplomarbeit eine Suchmaschine auf Basis des Produktes e:IAS der Fa. empolis GmbH realisiert. Diese Lösung soll in verschiedenen Bereichen verbessert und erweitert werden.

Es sollen aussagekräftige Logfiles generiert und ausgewertet werden, insbesondere sollen die Auswertungen mit denen der vorhergehenden Suchlösung vergleichbar sein.

Bei der Ergebnispräsentation sollen die Erfordernisse der Barrierefreiheit beachtet werden und die vorhandenen Templates entsprechende Anpassung erfahren.

Die Lösung soll um Ansätze semantischer Suche erweitert werden. Es ist angedacht die vorhandene Synonymverwendung auszubauen und um Taxonomien zu einem Thesaurus zu erweitern. Dabei sollen verschiedene Möglichkeiten untersucht werden und eine Möglichkeit, mindestens prototypisch, integriert werden.

Inhaltsverzeichnis

Kurzfassung.....	1
Inhaltsverzeichnis	3
Abbildungsverzeichnis	7
Tabellenverzeichnis	9
Formelverzeichnis.....	9
1 Motivation	11
2 Grundlagen	13
2.1 Textbasierte Suche	13
2.2 Taxonomien und Thesauri	16
2.2.1 Was sind Taxonomien und Thesauri	16
2.2.2 Semantische Suche mittels Thesauri	22
2.2.3 Taxonomiebasierte Ähnlichkeitsmaße	22
2.2.3.1 Pfadlänge	23
2.2.3.2 Normalisierte Pfadlänge	23
2.2.3.3 Dichte des Zweigs	24
2.2.3.4 Extended gloss overlaps measure.....	24
2.2.3.5 Maß basierend auf Informationsgehaltswert des Konzepts ...	25
2.2.3.6 Maß basierend auf knoten- und kantenbasierten Techniken .	25
2.2.3.7 Maß abgeleitet aus der Informationstheorie	26
2.2.3.8 Vergleich.....	27
2.2.4 RDF-basierte Thesaurusrepräsentation: SKOS.....	27
2.3 Barrierefreiheit von Webanwendungen.....	34
2.3.1 Allgemeine Regelungen	35
2.3.2 Rechtliche Regelungen	37
3 Suchlösung der Stadt Nürnberg – der Ist-Stand	39
3.1 Abacho.....	40
3.2 E:IAS.....	41
3.2.1 Systemaufbau.....	41
3.2.1.1 Indexierung.....	42
3.2.1.2 Ergebnissauslieferung	42
3.2.2 Konfiguration	43
3.2.3 Such- und Indexierungsablauf.....	43

3.2.4	Verbesserungspotentiale	50
4	Thesaurusbasierte Suche	53
4.1	Ist-Stand.....	53
4.2	Realisierte Systemerweiterungen.....	53
4.2.1	Ähnlichkeitsmaße in e:IAS	53
4.2.1.1	Taxonomieähnlichkeitsmaß: Taxonomie.....	54
4.2.1.2	Taxonomieähnlichkeitsmaß TaxonomiePfad	58
4.2.2	Mögliche Thesauri und Datenquellen.....	60
4.2.2.1	WikiSaurus in Wiktionary	60
4.2.2.2	OmegaWiki	60
4.2.2.3	OpenThesaurus	61
4.2.2.4	Getty Thesaurus of Geographic Names.....	61
4.2.2.5	Projekt Deutscher Wortschatz.....	61
4.2.2.6	HUGO	62
4.2.2.7	GEMET Thesaurus	64
4.2.2.8	Eurovoc Thesaurus.....	64
4.2.3	Beispielhafte Einbindung von Thesauri.....	65
4.2.3.1	Ein XML-Thesaurus - Eurovoc.....	67
4.2.3.2	Ein SKOS-Thesaurus – GEMET	68
4.2.4	Gewichtung von Attributen und Synonymen	69
4.2.4.1	Gewichtswerte der Taxonomieähnlichkeit.....	70
4.2.4.2	Änderung der Gewichtungsfaktoren.....	70
4.2.4.3	Berechnung der globalen Ähnlichkeit.....	71
4.3	Analyse der Suchergebnisse.....	74
5	Barrierefreie Präsentation der Suchergebnisse.....	77
5.1	Ist-Stand.....	77
5.2	Realisierte Systemerweiterung.....	78
5.2.1	Zugriff auf die Ergebnisdaten	78
5.2.2	Aufbau der GUI	79
5.2.2.1	Verfügbare e:Script Tags	79
5.2.2.2	Gliederung der Seite	79
5.2.2.3	Umsetzung der Navigation.....	81
5.2.3	Verwandte Links	86
6	Logdateiauswertung.....	91
6.1	Ist-Stand.....	91
6.1.1	Analyse des Benutzerverhaltens.....	91
6.1.2	Anforderungen an die Logdateiauswertung	95
6.2	Realisierte Systemerweiterung.....	95

6.2.1 Erzeugung der Logdateien in e:IAS	95
6.2.2 Datenschutz.....	99
6.2.3 Umwandlung der Logdatei mit Shell-Skripten.....	99
6.2.4 Prototyp einer Logdateianalyse in Perl	100
6.2.4.1 Sicherheit der Anwendung.....	101
6.2.4.2 Datenbankstruktur	101
6.2.4.3 Einlesen der Logdateien	103
6.2.4.4 Auswertung der Logdaten.....	104
6.2.5 Stresstest.....	109
7 Ausblick	111
Literaturverzeichnis	113
Anhang.....	117

Abbildungsverzeichnis

Abbildung 1: Auszug aus dem Eurovc-Thesaurus.....	19
Abbildung 2: Beziehung BF und BS	20
Abbildung 3: Beziehung UB und OB	20
Abbildung 4: Beziehung VB.....	21
Abbildung 5: Beispiel für Maß basierend auf knoten- und kantenbasierten Techniken	26
Abbildung 6: RDF-Graph	28
Abbildung 7: Beispiel aus dem UK Archival Thesaurus (UKAT).....	29
Abbildung 8: Darstellung der SKOS Relationen	29
Abbildung 9: Darstellung von „skos:Concept“.....	30
Abbildung 10: RDF-Beispiel zu „skos:Concept“.....	30
Abbildung 11: Graph zu prefLabel, altLabel und Sprachkennzeichnung	31
Abbildung 12: RDF-Syntax zu prefLabel, altLabel und Sprachkennzeichnung	31
Abbildung 13: „skos-changeNote“ mit Verzweigung.....	32
Abbildung 14: Ober- und Unterbegriffe mit SKOS.....	33
Abbildung 15: RDF-Repräsentation der Ober- und Unterbegriffe.....	33
Abbildung 16: Systemaufbau e:IAS.....	41
Abbildung 17: Grafischer Editor „Creator“	43
Abbildung 18: DataPipeline und InsertCasePipeline - Einfügen von Fällen in den Index (Insert).....	44
Abbildung 19: Auszug aus dem Index.....	48
Abbildung 20: SuchPipeline - Passende Dokumente zu Anfragen finden (Retrieval)	48
Abbildung 21: Ordnungsbaum im Model Manager des Creators.....	54
Abbildung 22: Ähnlichkeitsmaß Taxonomie - Anfrage: Optimistisch, Fall: Pesimistisch.....	57
Abbildung 23: Ähnlichkeitsmaß Taxonomie – links: Anfrage: Pessimistisch, Fall: Pesimistisch; rechts: Anfrage: Optimistisch, Fall: Optimistisch	57
Abbildung 24: Ähnlichkeitsmaß TaxonomiePfad - Ähnlichkeiten zu Knoten K41 59	
Abbildung 25: Beispielhafter Pfad der HUGO Navigation.....	63
Abbildung 26: Analysemöglichkeit 1	65
Abbildung 27: Analysemöglichkeit 2.....	66
Abbildung 28: Benutzeroberfläche von Twinkle.....	69
Abbildung 29: Ähnlichkeiten in der Eurovoc-Taxonomie	70
Abbildung 30: Vergleich der Suchergebnisse (Ausschnitt).....	75
Abbildung 31: Gliederung der Ergebnisseite	80
Abbildung 32: Ausgabe verschiedener Ergebnisse	81

Abbildung 33: Aufbau der Navigation (Auszug für 4 Seiten).....	83
Abbildung 34: Navigationslinks (ohne JavaScript).....	84
Abbildung 35: Quelltextauszug für die dritte Navigationsseite	85
Abbildung 36: Navigationslinks (JavaScript).....	86
Abbildung 37: Verwandte Links	87
Abbildung 38: Verwandte Links – Initialisierungsregel.....	88
Abbildung 39: Verwandte Links - Altstadtfreunde (gekürzt)	89
Abbildung 40: Ausgabe der Verwandten Links	89
Abbildung 41: Verteilung der gesamten Anfragen pro Monat. Die Werte sind normalisiert.....	91
Abbildung 42: Altes Layout des Webportals der Stadt Nürnberg.....	92
Abbildung 43: Neues Layout des Webportals der Stadt Nürnberg.....	92
Abbildung 44: Suchpipeline - links: ursprüngliche Anordnung, rechts: neue Anordnung.....	96
Abbildung 45: Regelsatz 1 (Vervollständigungsregeln)	97
Abbildung 46: Regelsatz 2 (Anpassungsregeln).....	97
Abbildung 47: Logdatei von e:IAS.....	98
Abbildung 48: Umgewandeltes Logfile.....	100
Abbildung 49: Datenbankstruktur.....	103
Abbildung 50: e:IAS Logfile-Auswertung.....	105
Abbildung 51: Ausgabe der Suchwörter	106
Abbildung 52: Ausgabe der Suchphrasen.....	107
Abbildung 53: Ausgabe von Daten zu Anfragehäufigkeit.....	108

Tabellenverzeichnis

Tabelle 1: IDF-Werte	15
Tabelle 2: TF-Wert	15
Tabelle 3: TF/IDF-Werte.....	15
Tabelle 4: Kürzel und Bezeichnungen in Thesauri	21
Tabelle 5: Datentypen der Attribute	46
Tabelle 6: Attribute	47
Tabelle 7: Beispielhafte Anfragen und ihre Ähnlichkeitswerte	56
Tabelle 8: Schlagwörter aus der HUGO Navigation	63
Tabelle 9: Dateien des eurovoc-Thesaurus.....	67
Tabelle 10: Dateien des GEMET-Thesaurus.....	68
Tabelle 11: Attribute, Gewichte und Ähnlichkeitsmaß	70
Tabelle 12: Lokale und Globale Ähnlichkeit (Relevanz)	73
Tabelle 13: Vergleich von Abacho und Empolis	76
Tabelle 14: Zusammenfassung der BIENE-Kriterien Auswertung	77
Tabelle 15: e:Script Tags	79
Tabelle 16: Top-Suchanfragen	94
Tabelle 17: Suchwörter mit wenigen Ergebnissen (<10 Ergebnisse im Durchschnitt).....	94

Formelverzeichnis

Formel 1: Inverse Dokumenthäufigkeit des Terms i	14
Formel 2: Termfrequenz des Terms i im Dokument d	14
Formel 3: Pfadlänge	23
Formel 4: Normalisierte Pfadlänge	23
Formel 5: Extended gloss overlaps measure	24
Formel 6: Ähnlichkeitsmaß basierend auf Informationsgehalt	25
Formel 7: Ähnlichkeitsmaß basierend auf Knoten und Kanten.....	25
Formel 8: Ähnlichkeitsmaß abgeleitet aus der Informationstheorie	26
Formel 9: Berechnung des globalen Maximums	72
Formel 10: Globales Maximum: Euklidischer Abstand nach Dokumentation....	72
Formel 11: Globales Maximum: Euklidischer Abstand	73

1 Motivation

Das Schwerste: Immer wieder entdecken, was man ohnehin weiß.¹

In dem Webportal der Stadt Nürnberg ist Wissen zu vielen verschiedenen Themen auf unzähligen Seiten gespeichert. Doch dieses Wissen ist nutzlos, wenn man es nicht findet. Diese Diplomarbeit soll dazu beitragen, dass die Informationen, die ein Besucher des Webportals sucht, von ihm auch gefunden werden.

Diese Arbeit baut auf der Diplomarbeit von Marek Ertel² auf und führt dessen Thema weiter. Neben der produktiven Inbetriebnahme der Suchmaschine auf Basis des Produktes e:IAS der Fa. empolis GmbH besteht die Arbeit aus drei Themen:

- Es sollen aussagekräftige Logfiles generiert und ausgewertet werden.
- Bei der Ergebnispräsentation sollen die Erfordernisse der Barrierefreiheit beachtet werden.
- Thesauri sollen die Lösung um Ansätze semantischer Suche erweitern.

¹ Elias Canetti (1905 - 1994), Schriftsteller spanisch-jüdischer Herkunft

² Siehe [Ertel2006]

2 Grundlagen

Dieses Kapitel soll einige Grundlagen klären, die für die vorliegende Diplomarbeit benötigt werden.

2.1 Textbasierte Suche

Die einfachste Suchmöglichkeit, um passende Dokumente zu finden, ist der Vergleich der Wörter in der Anfrage mit den Wörtern im Dokument; je mehr Wörter der Anfrage im Dokument vorhanden sind, desto relevanter ist es für den Benutzer. Um mehr Ergebnisse zu erzielen, können die Wörter vorher auf ihre Grundformen zurückgeführt werden, so wird z. B. „ging“ zu „gehen“ und „Häuser“ zu „Haus“. Diesen Vorgang nennt man Stemming.

Allerdings ist diese Methode zunächst nicht sehr gut geeignet, um die Relevanz eines Dokuments zu bestimmen, da beispielsweise nicht berücksichtigt wird, wie häufig das gesuchte Wort allgemein im Sprachgebrauch vorkommt. So ist z. B. „nicht“ das 16-häufigste Wort im Deutschen³, es wird also in vielen Dokumenten der Suchbasis vorkommen und ist somit als Suchbegriff wesentlich schlechter geeignet als ein Wort, das nur selten verwendet wird. Wie relevant ein Dokument als Ergebnis einer Suche ist, hängt weiterhin sicher davon ab, wie häufig ein Suchbegriff in dem Dokument enthalten ist.

Beide Überlegungen werden mit dem Suchverfahren TF/IDF (Term Frequency / Inverted Document Frequency) verfolgt. Über die inverse Dokumenthäufigkeit (engl. IDF) bekommt ein Term, also ein Wort, das nur in wenigen Dokumenten der Dokumentenbasis vorkommt, einen höheren Wert als ein Wort, das in vielen Dokumenten der Dokumentenbasis erscheint. Formel 1 ist die dazugehörige Berechnungsfunktion, sie kann reelle Werte größer Null annehmen.

³ Vgl. [Wortschatz]

$$idf_i = \log \frac{N}{n_i}$$

mit N Anzahl aller Dokumente

n_i Anzahl der Dokumente die Term i beinhalten

Formel 1: Inverse Dokumenthäufigkeit des Terms i

Die Termfrequenz (TF) gibt die relative Häufigkeit eines Wortes bzw. Terms in einem bestimmten Dokument an (siehe Formel 2). Sie kann Werte zwischen Null und Eins annehmen.⁴

$$tf_{i,d} = \frac{freq_{i,d}}{\max_l freq_{l,d}}$$

mit $freq_{i,d}$ Häufigkeit des Terms i im Dokument i

$\max_l freq_{l,d}$ Häufigkeit des häufigsten Terms l im Dokument i

Formel 2: Termfrequenz des Terms i im Dokument d

TF/IDF ist das Produkt aus Termfrequenz und inverser Dokumenthäufigkeit und somit kann die Relevanz eines Dokuments zu einer Suchanfrage berechnet werden.

Folgendes Beispiel soll die Berechnung erläutern:

Das Dokument d_1 enthält die Wörter „Stadt Nürnberg“, das zweite Dokument d_2 „Stadt Schwabach“ und das dritte (d_3) „Landkreis Fürth“. Zuerst werden die Terme in Kleinbuchstaben umgewandelt und Umlaute durch ihre Umschreibung ersetzt. Danach erfolgt die Berechnung der IDF-Werte (siehe Tabelle 1).

⁴ Vgl. [Ertel2006], Kapitel 2.4.2, S. 14 f

i	idf_i
stadt	$\log \frac{3}{2} = 0,176$
schwabach	$\log \frac{3}{2} = 0,176$
fuerth	$\log \frac{3}{1} = 0,477$
nuernberg	$\log \frac{3}{1} = 0,477$
landkreis	$\log \frac{3}{1} = 0,477$

Tabelle 1: IDF-Werte

Ebenso werden die TF-Werte aller Terme berechnet (siehe Tabelle 2).

i \ d	d1	d2	d3
stadt	1	1	0
schwabach	0	1	0
fuerth	0	0	1
nuernberg	1	0	0
landkreis	0	0	1

Tabelle 2: TF-Wert

Aus diesen Werten lassen sich dann die TF/IDF-Werte berechnen (siehe Tabelle 3), die einzelnen Spalten der Tabelle lassen sich nun auch als Gewichtsvektor des jeweiligen Dokuments lesen.

i \ d	d₁	d₂	d₃
stadt	0,176	0,176	0
schwabach	0	0,477	0
fuerth	0	0	0,477
nuernberg	0,477	0	0
landkreis	0	0	0,477

Tabelle 3: TF/IDF-Werte

Wird nun eine Anfrage q mit dem Term „Stadt Nürnberg“ übermittelt, wird wieder wie oben der TF/IDF-Wert berechnet und als Anfragevektor bekommt man $q = (0,176, 0,477, 0,0,0)$. Durch den direkten Vergleich sieht man, dass das Dokument d_2 exakt der Anfrage entspricht, die Ähnlichkeit also 1 ist, zum Dokument d_3 besteht gar keine Übereinstimmung, also eine Ähnlichkeit von 0 und zum Dokument d_1 besteht nur eine

teilweise Ähnlichkeit. Würde man die Werte in eine Berechnungsfunktion für die Größe der Ähnlichkeit einsetzen, auf die hier nicht weiter eingegangen wird (eine abgewandelte Form des Kosinusmaßes), würde man eine 35-prozentige Ähnlichkeit erhalten.⁵

Diese Form der Relevanzbestimmung wurde in einer Vorgängerdiplomarbeit in die e:IAS Suche integriert, die vorliegende Arbeit wird die Einbindung und Verwendung von Thesauri zur Relevanzbestimmung untersuchen.

2.2 Taxonomien und Thesauri

2.2.1 Was sind Taxonomien und Thesauri

Ein Thesaurus, im Sinne der Information und Dokumentation, ist nach DIN 1463-1⁶ (bzw. ISO 2788) Teil eines Informationssystems. Seine wesentlichen Anwendungen lassen sich wie folgt darstellen:

Die wesentlichen Inhalte einer Wissensquelle werden mit einem Thesaurus beschrieben (erschlossen), dies geschieht durch die sogenannte Indexierung. Das Indexierungsergebnis ist eine Liste natürlichsprachiger Wörter, die nicht frei wählbar sind, sondern nach bestimmten Regeln einem Thesaurus entnommen werden müssen.

Bei dem Information Retrieval (Informationswiedergewinnung) dient der Thesaurus der Suche nach relevanten Wissensquellen (Dokumenten), indem sich der Nutzer der indexierten Wörter bedient. Da der Thesaurus diese Wörter gleichzeitig auch miteinander in Beziehung setzt, kann dieses Beziehungsgeflecht auch als Suchhilfe (Pfad) genutzt werden.

Thesauri werden immer auf Basis einer Wissenssammlung erstellt. Weltweit dürfte es mehrere Tausend Thesauri geben, die alle auf bestimmte Fachgebiete ausgerichtet sind. Der „Thesaurus Guide“ verzeichnete 1993 rund 600 aktiv genutzte Thesauri in unterschiedlichen natürlichen Sprachen⁷.

Ein Thesaurus enthält ein „kontrolliertes Vokabular“, also eine eindeutige Benennung für jeden Begriff (Deskriptor oder Schlagwort) – diese Benennung kann, wenn der Thesaurus elektronisch verarbeitet wird, auch vollkommen abstrakt sein (z. B. eine

⁵ Vgl. [Ertel2006], Kapitel 2.4.2 S. 16 f

⁶ [DIN1463-1]

⁷ [Eurobrokers1992]

laufende Nummer), deswegen spricht man hier auch von einem Konzept. Oft nimmt man aber dennoch eine natürlichsprachige Vorzugsbezeichnung, den Deskriptor. Da eine Eindeutigkeit in der natürlichen Sprache jedoch nicht gegeben ist, werden außerdem Äquivalenzrelationen eingefügt. So ist die Synonymie die Gleichheit oder auch nur große Ähnlichkeit der Bedeutung von unterschiedlichen Wörtern. Zu einem Deskriptor können also beliebig viele Synonyme in Beziehung gebracht werden. Gleiches gilt für Wörter, die unterschiedliche Schreibweisen besitzen, hier werden alle Möglichkeiten als Synonym-Beziehung angegeben; wichtig ist das insbesondere, wenn man eine Wissensbasis indexieren will, in der Dokumente in neuer und alter deutscher Rechtschreibung vorhanden sind. Auch Abkürzungen und eventuell Übersetzungen können so behandelt werden.

Schwieriger wird es mit Homonymen oder Polysemen, also Wörtern, die verschiedene Bedeutungen besitzen.⁸ Hier werden die Wörter mehreren Deskriptoren zugeordnet und gleichzeitig markiert, um ihre Mehrdeutigkeit anzuzeigen. Zur richtigen Einordnung eines Dokuments muss dann der Kontext angeschaut werden, was bei der automatischen Verarbeitung Schwierigkeiten macht.

Daneben existieren noch hierarchische Relationen, um auf Hyponyme (Unterbegriffe) und Hyperonyme (Oberbegriffe) zu verweisen. DIN 1463-1 unterscheidet dabei noch zwischen generischer Relation, was als „eine hierarchische Relation zwischen zwei Begriffen, von denen der untergeordnete Begriffe (Unterbegriff) alle Merkmale des übergeordneten Begriffs (Oberbegriff) besitzt und zusätzlich mindestens ein weiteres spezifizierendes Merkmal“⁹ definiert wird und partitiver Relation, was als „eine hierarchische Relation zwischen zwei Begriffen, von denen der übergeordnete (weitere) Begriff (Verbandsbegriff) einem Ganzen entspricht und der untergeordnete (engere) Begriff (Teilbegriff) einen der Bestandteile dieses Ganzen repräsentiert“¹⁰ beschrieben wird.

Soll eine Beziehung beschrieben werden, die nicht den bisherigen Definitionen entspricht, existiert noch die Assoziationsrelation, sie ist eine „zwischen Begriffen bzw. ihren Bezeichnungen als wichtig erscheinende Relation, die weder eindeutig hierar-

⁸ Besitzen beide Wörter die selben etymologische Wurzeln, spricht man von Polysemie, diese Wörter haben eine ähnliche Bedeutung (z. B. Pferd als Tier und Turngerät), andernfalls von Homonymie (z. B. Bank als Sitzmöbel oder Kreditinstitut).

⁹ [DIN1463-1], Teil 1

¹⁰ ebenda

chischer Natur ist, noch als äquivalent angesehen werden kann.“¹¹ Diese schwammige Definition ist auch die Problematik dieser Beziehung. Sie kann zu einem Sammelbecken geraten, in das alles hineingenommen wird, was in einem sehr weiten Sinn mit dem Ausgangsbegriff zu tun hat. Oft entstehen dadurch sehr lange Reihen solcher „verwandten Begriffe“. Ein Thesaurus sollte aber nicht versuchen, alle möglichen Zusammenhänge auszuweisen, in denen ein Begriff vorkommen kann. Der Sinn dieser Relation ist vielmehr, „zusätzlich zur hierarchischen Struktur Querbeziehungen zu anderen, für die Formulierung des Sachverhaltes möglicherweise geeigneten Deskriptoren anzubieten“¹².

¹¹ ebenda

¹² [Burkart2004], Kapitel B 2.1.4.4, Seite 149