In: Hammwöhner, Rainer; Wolff, Christian; Womser-Hacker, Christa (Hg.): Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002), Regensburg, 8. – 11. Oktober 2002. Konstanz: UVK Verlagsgesellschaft mbH, 2002. S. 181 – 199

Automatische Thesauruserstellung und Query Expansion in einer E-Commerce-Anwendung

Libo Chen, Ulrich Thiel, Marcello L'Abbate

Fraunhofer Institut
Integrierte Publikations- und Informationssysteme
(IPSI)
Dolivostraße 25
D-64293 Darmstadt
{chen,thiel,labbate}@ipsi.fhg.de

Zusammenfassung

Diese Arbeit beschreibt eine Methode für den automatischen Aufbau eines Thesaurus auf der Basis von bereits existierenden Kategorien von Dokumenten. Ein neues Clusteringverfahren, das „Layer-Seeds Verfahren", wird in der Arbeit vorgestellt. Es bearbeitet die Terme in einer bestimmten Kategorie und organisiert sie zu einer baumähnlichen hierarchischen Struktur. Anhand dieser Struktur kann in einem interaktiven Retrievalsystem auf der Basis eines intelligenten Agenten die Benutzeranfrage expandiert werden, um bessere Suchergebnisse zu erzielen. Eine Pilotanwendung des Verfahrens wurde im Rahmen eines E-Commerce-Systems entwickelt.

Abstract

This work describes a method for the automatic construction of a thesaurus based on existing categories of documents. A clustering algorithm, "the layer seeds method", is introduced, which facilitates the automatic generation of a thesaurus reflecting the specific vocabulary occurring in a given collection of documents. We assume that the collection is partitioned into document categories. The clustering works on terms extracted from the documents in a certain category and organizes them in a tree-like hierarchical structure. In an interactive retrieval system based on an intelligent agent, we used an automatically generated thesaurus to expand the user's queries, in order to obtain better search results. A pilot application of the method was integrated into an e-commerce website.

1 Einleitung

Online-Shopsysteme mit integrierten Suchfunktionen werden mittlerweile bei den meisten E-Commerce-Anwendungen in Kundengeschäft eingesetzt. Die Qualität der Informationen, die von solchen Online-shops angeboten werden, spielt für den erfolgreichen Geschäftsablauf eine entscheidende Rolle. Im betrachteten Anwendungsfall ist daher die Vermeidung von irrelevanten Angeboten ein wichtiges Ziel. Aber auch die Bereitstellung einer sinnvollen Auswahl von Produkten aus einer Kategorie muss gewährleistet sein. Beide Ziele sind i.a. erreichbar durch die Formulierung elaborierter Datenanfragen, die - zu einer aktuellen Kategorie von Produkten - eine Menge von Begriffen umfassen, welche die potentiell relevanten Objekte charakterisieren. Mit anderen Worten, um IR-Systeme zur Suche von Produktangeboten einsetzen zu können, benötigen wir ein Verfahren, das die üblicherweise von Benutzern ad-hoc eingegebenen Anfragen automatisch erweitert, und so die Folgen von zu vieldeutigen oder einseitig fokussierten Benutzeranfragen unter Verwendung des Kontextes der aktuellen Kategorie kompensiert. Query Expansion zählt zu einer der wichtigsten Methoden, dieses Problem zu lösen. Diese Methode versucht, die von Benutzern eingegebenen Anfragen „richtiger" zu formulieren, damit die Benutzerwünsche möglichst unter Verwendung des Vokabulars der Informationsquelle repräsentiert werden. Dazu werden normalerweise Fachwörterbücher oder Thesauri verwendet. Bei Online-Shopping Systemen wurde bisher noch wenig relevante Arbeit in dem Bereich Query Expansion geleistet. Im Vergleich zu anderen Informationssystemen weist ein Online-Shopping System jedoch einige positive Eigenschaften auf, z. B. gut strukturierte Informationen über Produkte (Produktkategorien) oder praktische Tools zum Auffinden der Produkte (eingebaute Volltext-Suchmaschine). Die Ausnutzung dieser Eigenschaften kann die Aufgabe des Information Retrieval erleichtern.

Das im Folgenden beschriebene Verfahren zielt darauf ab, Anfragen der Benutzer automatisch zu expandieren, damit befriedigendere Ergebnisse mit einer der Situation angemessenen Anzahl von Produktangeboten ausgegeben werden. Dazu wird ein Thesaurus auf der Basis von bereits existierenden Produktkategorien durch ein spezielles Clusteringverfahren („Layer-Seeds Verfahren") automatisch aufgebaut und zur automatischen Expansion der Benutzeranfragen verwendet. Somit lässt sich die Performance des gesamten E-Commerce Systems deutlich verbessern.

In diesem Paper stellen wir zunächst einige verwandte Ansätze vor. Dann präsentieren wir den Aufbau des Thesaurus anhand des „Layer-Seeds Verfahren"

und dessen Anwendung im E-Commerce Projekt COGITO. Die Anwendung wird anschließend experimentell evaluiert. Der Ausblick auf geplante Weiterentwicklungen beschließt diesen Beitrag.

2 Relevante Forschungsansätze

In den letzten Jahren wurden zahlreiche Forschungsprojekte im Bereich Query Expansion auf der Basis von Thesauri durchgeführt. Die automatische Query Expansion ist relativ einfach zu konzipieren bzw. zu implementieren. Dabei nehmen Benutzer nicht an dem Prozess teil. Der Expansionsvorgang läuft im Hintergrund. Die Ergebnisse werden dem Benutzer erst am Ende des Prozesses geliefert. Zur automatischen Query Expansion zählen z.B. die Arbeit von Spark-Jones [Spa71] und Crouch & Yong [CrY92]. Minker et al. [Mink72] zweifelten jedoch an dem Ergebnis von Spark-Jones. Ihre Forschung zeigte, dass in manchen Fällen die Termklassifikation und die darauf basierende Query Expansion zu einer Effizienzminderung des Information Retrieval führen könnten. Salton [Salton73], in einem Review über die Arbeit von Minker et al., hat ihren experimentellen Ansatz in Frage gestellt, somit blieb die Problematik der Effektivität der automatischen Query Expansion immer noch offen und muss bis heute im Einzelfall empirisch nachgewiesen werden. Eine Alternative bieten interaktive Verfahren wie das Relevance Feedback, bei dem Benutzer ihre Anfragen angeben und entscheiden, welche Dokumenten unter den Ergebnissen relevant sind. In diesen relevanten Dokumenten wird dann nach Termen gesucht, die später für Query Expansion verwendet werden [Rocchio71]. Eine andere Möglichkeit ist, dass den Benutzern eine Liste von Termkandidaten geliefert wird und sie treffen die Entscheidung, welche Terme in die Query aufgenommen werden sollen [Ekm92].

Grob gesehen ist ein Thesaurus eine Menge von Termen plus einer Menge von Relationen zwischen diesen Termen. Bisher werden die meisten Thesauri manuell, in einer zunehmenden Anzahl von Projekten aber bereits automatisch aufgebaut. Die manuell erstellten Thesauri lassen sich wiederum in zwei Typen unterteilen. Der erste umfasst die Universalthesauri wie Roget's und WordNet, die Bedeutungsrelationen wie Antonymie und Synonymie für ein einzelnes Thesauruselement enthalten. Sie werden selten in IR-Systemen verwendet. Der zweite betrifft IR-orientierte Thesauri wie INSPEC, die normalerweise Relationen unter Elementen im Thesaurus enthalten, wie z.B. BT (Broader Term), NT (Narrow Term), UF (Used For), and RT (Related To). Diese Art Thesaurus wird allgemein in kommerziellen Systemen verwendet. Ein Thesaurus kann auch automatisch berechnet werden, indem man in großen Dokumentsammlungen aus dem gemeinsamen Auftreten von Wörtern in

Dokumenten oder Dokumentteilen assoziative Beziehungen zwischen ihnen ableitet [Spa71; CrY92; Qiu93]. Die beiden Aufbaumethoden haben ihre Vor- und Nachteile. Die manuell erstellten Thesauri weisen zwar sehr gute Elemente und Struktur auf, sind aber sehr aufwendig zu konstruieren bzw. zu pflegen. Die automatisch generierten Thesauri dagegen haben zwar nicht immer die richtigen Elemente und meist eine Struktur, die deren semantische Beziehungen nur ungenau und fehlerhaft reflektiert, dafür sind sie aber einfacher und schneller zu erzeugen. Daher können sie aktueller sein und sowohl umfassend als auch für spezifische Fachgebiete bereitgestellt werden. Im folgenden Abschnitt wird versucht, eine Kombination der beiden Thesaurusaufbaumethoden zu finden, die die Vorteile der beiden Methoden aufweist.

Panyr [Panyr86] hat in seiner Arbeit den Steinadler-Ansatz vorgestellt, bei dem sowohl Terme als auch Dokumente automatisch klassifiziert werden. Der Vorgang läuft in drei Phasen ab. Zuerst werden Terme anhand ihrer absoluten Dokumentshäufigkeiten im ganzen Textcorpus in mehreren Prioritätsklassen eingeteilt. Anschließend wird in der einzelnen Klasse eine Clusteranalyse durchgeführt, damit Terme und Dokumente in Clusters organisiert werden. Dabei ist irrelevant, welches Clusterverfahren angewandt wird. Zum Schluss werden die benachbarten Prioritätsklassen anhand der gemeinsamen Dokumentenclusters miteinander verglichen und verknüpft.

Ein ganz anderer Ansatz geht in die Richtung automatische hierarchische Text-Kategorisierung [Göv99; Jicr94; From01]. Ein beliebiges Dokument wird unter Berücksichtigung der Struktur von existierenden Kategorien einer zu ihm am besten passenden Kategorie zugeordnet. Dabei wird eine Kategorie als Megadokument betrachtet und weiterhin als Termvektor repräsentiert, der mit dem Termvektor des zu klassifizierenden Dokuments verglichen wird. Die hierarchische Struktur von Kategorien wird dadurch berücksichtigt, dass die Beziehungen unter Kategorien gewichtet und in die Gewichtung der Ähnlichkeit zwischen dem Dokument und den Kategorien einbezogen werden. Ein Clustering von Termen oder Dokumenten findet bei dieser Methode nicht statt.

3 Automatischer Thesaurusaufbau auf der Basis des „Layer-Seeds Verfahren“

Der automatische Vorgang des Thesaurusaufbaus basiert auf manuell erstellten bzw. gepflegten Kategorien. Sie sind Repräsentation des menschlichen Wissens der Anwendungsexperten, die durch die Auswahl und Zusammen-

stellung eines der Anwendung angepassten Kategoriensystems den semantischen und z.T. auch pragmatischen Kontext vorgeben. So ist es nicht überraschend, wenn z.B. die konkreten Kategorien in einer Buchhandelsanwendung von den im Bibliothekswesen gebräuchlichen Klassifikationen etwas abweichen. Wir gehen davon aus, dass die Dokumentkollektion anhand dieser Kategorien sortiert ist, wodurch bereits eine Zuordnung zwischen einer Kategorie und den Termen der in ihr enthaltenen Texte induziert wird. Der im Folgenden beschriebene Vorgang wird nun jeweils nur auf Terme aus Texten einer spezifischen Kategorie d.h. auf eine viel kleinere Datenmenge als die Gesamtmenge angewendet. Dies hat zur Folge, dass die Präzision der Ergebnisse stark erhöht und der Rechenaufwand enorm reduziert wird. Manche komplizierte Clustering Methoden werden erst dadurch realisierbar.

Der eigentliche Vorgang des Aufbaus lässt sich hauptsächlich in 3 aufeinanderfolgenden Phasen untergliedern: Kollektion bzw. Auswahl der Thesauruselemente, Beziehungsfeststellung zwischen den Thesauruselementen und Klassifizierung der Thesauruselemente anhand der Beziehungen. Zu der letzten Phase wird in dieser Arbeit eine Methode mit dem Namen „Layer-Seeds Verfahren“ konzipiert.

Um die Elemente des Thesaurus bestimmen zu können, werden zuerst Terme aus allen Daten unter einer bestimmten Kategorie gesammelt. Sie werden anschließend normalisiert. Die Normalisierung der Terme umfasst zwei Schritte: Stoppworteliminierung und Grundformreduktion. Dabei sind die Vorgehensweisen je nach Sprache unterschiedlich. In dieser Arbeit sind die Normalisierungsmethoden auf die deutsche Sprache ausgerichtet. Nach der Normalisierung werden noch bestimmt, welche übriggebliebenen Terme in den Thesaurus einbezogen werden müssen. Die Entscheidung geht davon aus, dass, wenn ein Term in eine Kategorie miteinbezogen werden soll, er auch kategoriespezifisch sein muss. Die kategorie-übergreifenden Terme sollen somit gelöscht werden. Es wird ein Kriterium eingesetzt, um zu überprüfen, ob ein Term kategoriespezifisch ist oder nicht. Das Kriterium lautet: Wenn sich ein Term gleichverteilt in verschiedenen Kategorien befindet, ist dieser Term *nicht* kategorienspezifisch. Um die Verteilungen eines Terms in zwei Kategorien zu vergleichen, werden folgende Formeln verwendet, die in Analogie zum Vorgehen bei der Varianzanalyse entwickelt wurden. Gegeben sei ein Term, der schon normalisiert ist. Es wird angenommen, dass der Term in m Kategorien vorkommt mit den Häufigkeiten hk_1, hk_2,..., hk_m.

Summe der Häufigkeiten: S= $\sum_{i=1}^{m} hk_i$ **Mittelwert: MW=** $\frac{S}{m}$

Verteilungsgrad= $\sum_{i=1}^{m} \left(\frac{hk_i - MW}{S} \right)^2 = \frac{1}{S^2} \sum_{i=1}^{m} \left(hk_i - MW \right)^2$

Der Verteilungsgrad misst die Unterschiede zwischen den Häufigkeiten, mit denen der Term in jeweils verschiedenen Kategorien auftritt. Je kleiner der Wert, umso mehr ist der Term gleichverteilt. Wenn der Wert größer als ein bestimmter Schwellenwert ist (Schw1), so ist der Term nicht in allen Kategorien gleich verteilt und kann in den Thesaurus miteinbezogen werden. Dieser Verteilungsgrad unterscheidet sich von der Varianz dadurch, dass hier das Quadrat der Häufigkeitssumme S statt der Anzahl der Kategorien m als Nenner genommen wird. Somit werden die Häufigkeiten der Terme normalisiert, so dass ein einheitlicher Schwellenwert gefunden werden kann, um mit dem Verteilungsgrad die Ähnlichkeit der Verteilungen für alle Terme zu bestimmen.

Nachdem die Elemente für den Thesaurus bestimmt sind, werden die Beziehungen zwischen den Elementen festgestellt. Bei dem in dieser Arbeit zu erstellenden Thesaurus sind zwei Termbeziehungen zu betrachten, nämlich die „Thematische Spezialisierung" und die „Relevant"-Beziehung. Alle anderen Terme, die keine dieser zwei Beziehungen besitzen, sind in Bezug auf einander irrelevant. Um festzustellen, welche Beziehung zwischen zwei konkreten Thesauruselementen besteht, werden zuerst die Terme mit Termvektoren repräsentiert. Das i-te Element im jeweiligen Termvektor ist die Häufigkeit des Terms im i-ten Dokument. Die Dimension des Vektors ist die Anzahl der Dokumente in einer Kategorie.

Das Kosinusmaß wird eingesetzt, um die Relevanz zwischen zwei Termen zu berechnen, wobei V_x, V_y zwei Termvektoren und X_k und Y_k das k-te Element im jeweiligen Termvektor sind. Das Kosinusmaß hat den Vorteil, nur den Winkel zwischen beiden Vektoren zu berücksichtigen. Dadurch wird der Einfluss des Längenunterschieds von Dokumenten eliminiert und die absolute Häufigkeit der Terme ist nicht mehr relevant.

$$Cos\ (Vx, Vy) = \frac{Vx \cdot Vy}{|Vx| \cdot |Vy|} = \frac{\sum_{k=1}^{n} X_k Y_k}{\sqrt{\sum_{k=1}^{n} X_k^2} \cdot \sqrt{\sum_{k=1}^{n} Y_k^2}}$$

Es wird noch ein Schwellenwert (Schw4) eingeführt. Zwei Terme, deren Kosinusmaß größer als dieser Schwellenwert ist, werden dann als wechselseitig

relevant betrachtet. Bei dieser Vorgehensweise wird jedoch eine mögliche Beziehung ignoriert. Wenn ein Term immer mit einem anderen Term zusammen in demselben Dokument in einer bestimmten Kategorie vorkommt, der zweite Term aber nur in einem viel geringeren Maß mit dem ersten Term zusammen auftritt, so ließe sich intuitiv behaupten, dass der zweite Term allgemeinere Bedeutung hat als der erste. Das heißt, es besteht eine „Thematische Spezialisierung“-Beziehung zwischen den beiden Termen. Diese Beziehung ist als unabhängig von der mit dem Kosinusmaß bestimmten „Relevant“-Beziehung anzusehen.

Um das Prinzip näher zu erläuten, werden hier Similaritätsgrade eingeführt. t1 und t2 seien Terme, V(t1) ist der Termvektor von Term1, V(t2) ist der Termvektor von Term2

Similarität 0: s0 = cos (V(t1), V(t2))

$$\textbf{Similarität 1: s1=} \frac{\textbf{Anzahl der Dokumente, die t1 und t2 enthalten}}{\textbf{Anzahl der Dokumente, die t1 enthalten}}$$

$$\textbf{Similarität 2: s2=} \frac{\textbf{Anzahl der Dokumente, die t1 und t2 enthalten}}{\textbf{Anzahl der Dokumente, die t2 enthalten}}$$

Mit einer großen Similarität s0 ist die „Relevant“-Beziehung festzustellen. Eine kleine Similarität s1 mit einer viel größeren s2 bedeutet, dass t2 relativ häufig mit t1 zusammen vorkommt und t1 nur wenig mit t2 zusammen vorkommt. t2 ist daher wahrscheinlich der Spezialisierungsbegriff von t1. Wie klein s1 bzw. wie groß s2 sein soll, um die „Thematische Spezialisierung“ behaupten zu können, muss vorher genau angegeben werden. Zwei zusätzliche Schwellenwerte (Schw2 und Schw3) werden daher eingeführt.

Mit dieser drei Similaritäten lassen sich Beziehungen zwischen Termen feststellen.

„Thematische Spezialisierung“:

$$\text{Wenn} \quad \begin{cases} s1 \leq Schw2 \\ s2 \geq Schw3 \end{cases}$$

dann ist t2 ein Spezialisierungsbegriff von t1. Diese Beziehung zeigt an, dass ein Thema – gegeben durch Begriff t1 - durch den Begriff t2 spezialisiert oder

eingeschränkt werden kann. Dies ist z.B. der Fall, wenn t2 ein Unterbegriff, eine Eigenschaft, ein Teil, oder auch nur ein stark assoziierter Begriff in Bezug auf t1 ist.

„Relevant"-Beziehung:

$$\text{Wenn} \quad \begin{cases} s1 > Schw2 \\ s0 \geq Schw4 \end{cases} \quad \text{O-der} \quad \begin{cases} s2 < Schw3 \\ s0 \geq Schw4 \end{cases}$$

dann ist t1 für t2 relevant.

„Irrelevant":

$$\text{Wenn} \quad \begin{cases} s1 > Schw2 \\ s0 < Schwe4 \end{cases} \quad \text{oder} \quad \begin{cases} s2 < Schw3 \\ s0 < Schw4 \end{cases}$$

dann sind t1 und t2 irrelevant.

Die Feststellung der Schwellenwerte muss durch Experimente empirisch vorgenommen werden. Dabei sollen verschiedene Schwellenwerte ausprobiert werden, um zu überprüfen, bei welchen die beste Thesaurusstruktur geliefert werden kann. Wir nehmen $P_{element}$(Schw1) als die Gütewahrscheinlichkeit für die Elementauswahl unter der Einsetzung von Schwellenwert1 und $P_{beziehung}$(Schw2, Schw3, Schw4) als die Gütewahrscheinlichkeit, die darstellt, wie gut die Struktur des Thesaurus unter der Einsetzung von den Schwellenwerten2-4 ist.

$$\mathbf{P_{element}(Schw1)=} \frac{\textbf{Anzahl der beobachteten Terme, die richtig im Thesaurus sind}}{\textbf{Anzahl der gesamten beobachteten Terme}}$$

$$\mathbf{P_{beziehung}(Schw2,Schw3,Schw4)=} \frac{\textbf{Anzahl der beobachteten Beziehungen, die richtig im Thesaurus sind}}{\textbf{Anzahl der gesamten beobachteten Beziehungen}}$$

Die Richtigkeit der Terme und der Beziehungen wird von Menschen rein empirisch beurteilt. Die Schwellenwerte 1-4 müssen dann diejenige Werte einnehmen, die die höchste Gütewahrscheinlichkeit liefern, wobei sie mindestens 50% sein sollte.

Anhand der Beziehungen werden Terme zum Schluss mit dem „Layer-Seeds Verfahren" klassifiziert. Dabei wird zuerst jedem Term eine Gewichtung zugewiesen. Die Gewichtung wird nach folgender Formel berechnet:

K-TF-DF Gewichtung: $G(t, K_i)$ = $THK(t, K_i) * DHK(t, K_i)$

Wobei t ein Term, $THK(t, K_i)$ die Häufigkeit von t in der Kategorie K_i und $DHK(t, K_i)$ die Anzahl der Dokumente in der Kategorie K_i, die den Term t enthalten. Diese Formel besagt, je häufiger ein Term in der Kategorie vorkommt, desto wichtiger ist er; je breiter der Term in der Kategorie verteilt ist, desto wichtiger ist er. Nach der Termnormalisierung und der Löschung der Kategorie-unspezifischen Terme sind nur diejenigen Terme geblieben, die wichtig bzw. Kategorie-spezifisch sind. Je größer die Gewichtung eines solchen Terms ist, desto besser kann er die Kategorie vertreten. Die Wichtigkeit (Gewichtung) eines Terms liegt also nicht mehr darin, wie gut er die Dokumente unterscheidet, sondern darin, wie gut er eine Kategorie vertritt, wobei die DHK eine wichtigere Rolle spielt als THK. Anhand der Gewichtungsformel wird jedes Mal ein Term als „Samen“ in der Kategorie ausgewählt, dessen Gewichtung am größten unter den Termen ist, die noch nicht Samen sind. Diese Samen werden dann als Ausgangspunkte der späteren Prozedur dienen.

Der Algorithmus vom „Layer-Seeds Verfahren“ wird im Folgenden aufgeführt:

a. Am Anfang existiert nur eine Schicht, die alle Terme enthält.
b. Ein Term in der aktuellen Schicht wird als Samen gewählt, der noch kein Samen und dessen K-TF-DF Gewichtung am größten in dieser Schicht ist.
c. Die drei Similaritäten zwischen dem Samen und allen anderen Termen in derselben Schicht, die keine Samen sind, werden berechnet und deren Beziehungen werden somit festgestellt. Diejenigen Terme, die eine „Thematische Spezialisierung“-Beziehung mit dem Samen haben, werden als Spezialisierungsbegriffe in der nächsten Schicht diesem Samen zugeordnet. Diese Beziehungen werden als „Direkte Thematische Spezialisierung„ bezeichnet. Es wird noch überprüft, ob die Spezialisierungsbegriffe auch „Thematische Spezialisierung“-Beziehungen mit den Vorfahren dieses Samens haben; wenn ja, werden die Beziehungen mit den Vorfahren als „Indirekte Thematische Spezialisierung“ bezeichnet.
   Diejenigen Terme, die eine „Relevant“-Beziehung mit dem Samen haben, werden nicht bearbeitet, die Beziehung wird jedoch protokolliert.
   Diejenigen Terme, die „irrelevant“ bzgl. des Samens sind, bleiben in derselben Schicht unbearbeitet.

d. Falls noch Terme, die noch keine Samen sind, in der aktuellen Schicht gefunden werden können, wird Schritt b wiederholt. Im anderen Fall und wenn es noch untere Schichten gibt, wird die Schichtnummer um eins erhöht und der Schritt b wiederholt; ansonsten wird der Prozess beendet.

Abbildung 1 stellt den Vorgang des Verfahrens dar:

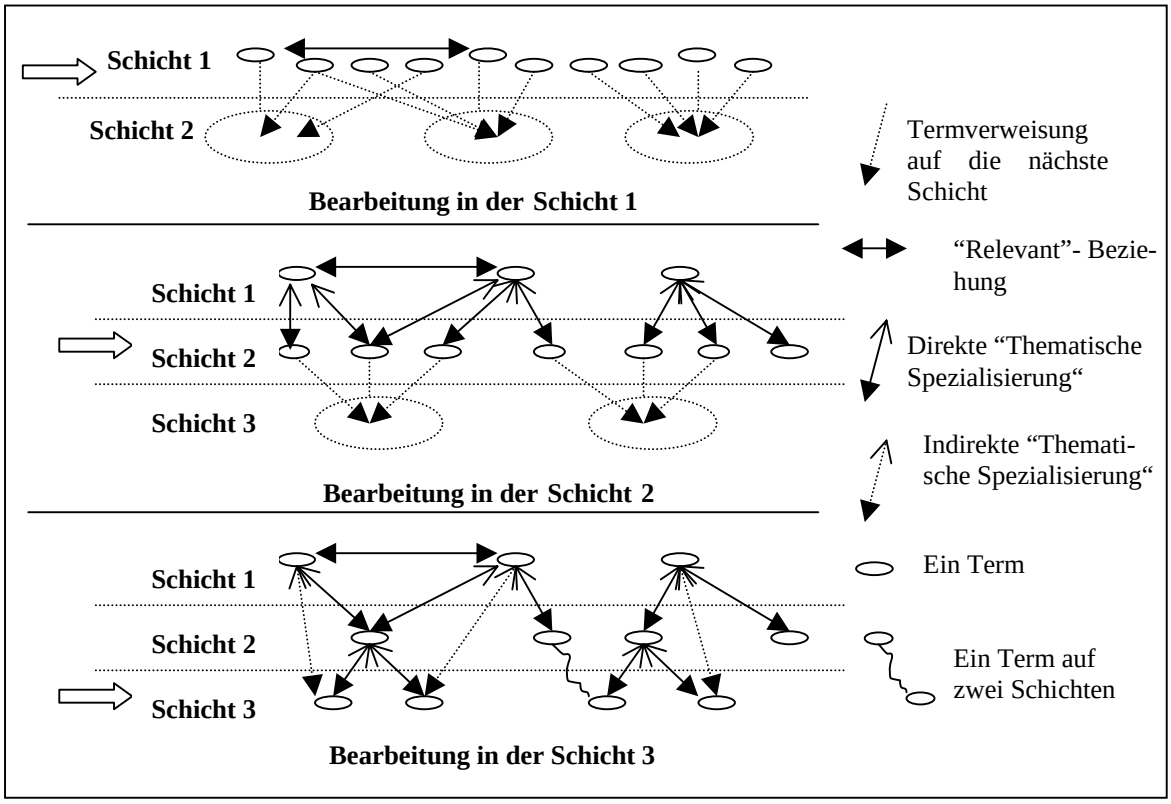


Abbildung 1: Vorgang des Thesaurusaufbaus nach dem „Layer-Seeds Verfahren"

Im Rahmen des COGITO Projektes (siehe Abschnitt 4) wurde das „Layer-Seeds Verfahren" mit verschieden Schwellenwerten ausprobiert. Insgesamt 600 Terme bzw. Termpaare wurden ausgewählt und geprüft. Dabei wurden Kurzbeschreibungen von Computerfachbüchern als Dokumentmenge betrachtet. Tabelle 1 zeigt die Schwellenwerte, die die besten Ergebnisse liefern.

Schwellenwert	Wert	P
Schw1	0,1	0,8
Schw2	0,4	0,69
Schw3	0,8	
Schw4	0,05	

Tabelle 1: Schwellenwertfeststellung

Als Beispiel werden nun von zwei Kategorien jeweils 20 Samen in der ersten Schicht aufgeführt. Die Terme innerhalb einer Kategorie werden nach ihrer Gewichtung geordnet. Es ist ersichtlich, dass diese Terme angemessene Vertreter der jeweiligen Kategorie sind.

Kategorie Nummer: K1 Kategorie Name: Betriebssysteme
Samen: Windows, Linux, Microsoft, Arbeit, Unix, Lernen, Betriebssystem, Netzwerk, Konfiguration, Grundlage, Shell, Programmierung, Mac, Handbuch, Netzwerke, Multimedia, Umgang, einrichten, X, Software

Kategorie Nummer: K3 Kategorie Name: Informatik
Samen: Informatik, Grundlage, Software, Lehrbuch, Algorithmen, Methode, Computer, Sprache, Universität, C, Java, Wirtschaftsinformatik, methods, Multimedia, Übungsaufgaben, Based, Verständnis, Netz, bilden, Data

_1:Java

__2:Netzwerk
__2:Netscape
____3:Browser
____3:Ergebnis
____3:anstellen
____3:leben
____3:VBScript
____3:Stylesheets
____3:Dummy
____3:JavaScripts
____3:Informatio
____3:benutzer-
freundlichen
__2:Mail
__2:Enterprise
__2:Datenstrukturen
__2:Commerce
__2:Ausdrücke
__2:APIs
__2:Grundkonzepte
__2:Corba
__2:classes
__2:Aussagenlogik
__2:Webprogramming
__2:Unicode
__2:Threading

__2:RMI
____3:programming
____3:Security
____3:Objektorientierung
____3:JavaBeans
____3:Zugreifbarkeit
____3:Verteilung
____3:unternehmensweiter
____3:Typumwandlungen
____3:Toolkit
____3:Timecard
____3:Threadprogrammierung
____3:Technology
____3:Technologies
____3:Superklassen
____3:Subklassen
____3:Softwarelösungen
____3:Selection
____3:Resources
____3:Package
____3:OMG
____3:Methodenaufrufe
____3:Klassendeklarationen
____3:IDL
____3:Geltungsbereiche

__2:JDOM
__2:JDK
____3:Security
____3:Klassenbibliotheken
____3:Verteilung
____3:Java2D
____3:IDL
____3:Fibel
____3:Beispielanwendungen
__2:Applets
____3:Threads
____3:Security
____3:frames
____3:EDV
____3:Zugreifbarkeit
____3:Typumwandlungen
____3:Toolkit
____3:Threadprogrammierung
____3:Superklassen
____3:Suns
____3:Subklassen
____3:Methodenaufrufe
____3:Konvention
____3:Klassendeklarationen
____3:JFC

Abbildung 2: Spezialisierungsbegriffe von dem Term „Java“ in der Kategorie „Programmiersprachen“

Bei einem gegebenen Term können in dem Thesaurus seine allgemeineren Begriffe, Spezialisierungsbegriffe und relevante Begriffe nach angegebenen Relevanzstufen gefunden werden. Als Beispiele werden nachfolgend die Spezialisierungsbegriffe von dem Term „Java" aus der Kategorie „Programmiersprachen" genommen. Der Term „Java" befindet sich in der ersten Schicht im Thesaurus. Es handelt sich um einen ziemlich breiten Begriff und er hat daher eine lange Liste von Spezialisierungs-begriffen. In der Abbildung 2 werden kleine Teile davon aufgelistet.

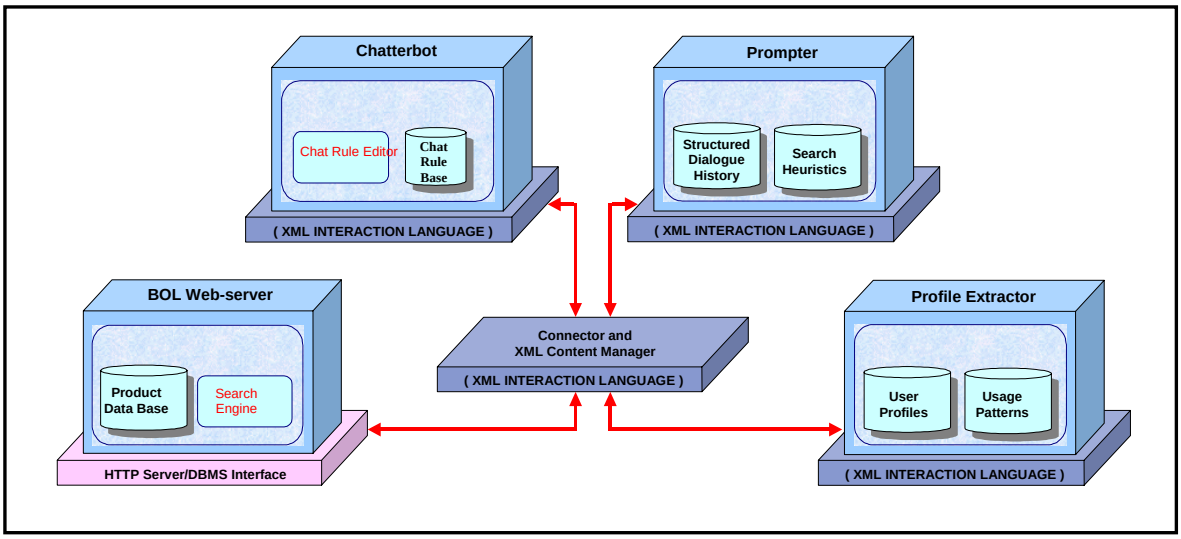
Es ist noch nicht klar, ob es universelle Schwellenwerte für alle Domänen gibt. Die für eine Domäne gültigen Schwellenwerte sind wahrscheinlich nicht auf andere Domäne übertragbar. Experimente müssen dann für jede einzelne Domäne durchgeführt werden.

4 Anwendung des Thesaurus

Ein konkretes Anwendungsbeispiel des Thesaurus und die darauf basierende automatische Query Expansion findet sich im EU-Projekt COGITO (Agenten für den E-Commerce mit personalisierter Interaktion) [Cogito], dessen Ziel die Bewältigung des Kommunikationsproblems im Bereich E-Commerce ist. In Zusammenarbeit mit BOL[1] wurden verschiedene Ansätze untersucht, um die Natürlichkeit, Einfachheit, Personalisierung und Effektivität der Kommunikation zu gewährleisten. Dabei wurde ein intelligenter Agent eingesetzt, der sich mit den Kunden nach vorher bestimmten Regeln in natürlicher Sprache unterhält (Chatterbot). Die Möglichkeiten heutiger Chatterbots werden im Projekt COGITO durch eine innovative Kombination moderner Techniken erweitert.

Dazu wird ein graphisches konversationales Interface implementiert, das nicht bloß auf Benutzereingaben reagiert, sondern in proaktiver Weise eine zielgerichtete Konversation mit dem Benutzer führen kann, um eine evtl. komplexe Aufgabe zu lösen. Unsicherheiten im Verständnis von Zielen/Aufgaben werden ausgeräumt, um den besten Lösungsweg zu finden, wobei die Reaktionen der Benutzer während des Dialogs erfasst und interpretiert werden. Auf dieser Basis kann das Verhalten des Systems adaptiert werden, z.B. durch Übernehmen der Initiative im Dialog, um neue Informationen oder Suchstrategien zu empfehlen. Die Natürlichkeit der Interaktion wird insbesondere für gelegent-

[1] Bertelsmann online. Internationaler Media- und Unterhaltungsshop im Internet http://www.bol.com.



liche Benutzer durch 2D- und 3D-Animationen erhöht, die im Bedarfsfall emotionale Reaktionen des Agenten simulieren.

Selbst „Chatterbots“ mit einem ausgefeilten Repertoire dialogischer Reaktionsmöglichkeiten bleiben jedoch nur auf ihren Unterhaltungswert beschränkt, wenn sie die Benutzer nicht als Individuen mit bestimmten Bedürfnissen, Vorlieben etc. wahrnehmen. In COGITO werden Verfahren zur inhaltsbasierten Filterung, bei denen Benutzerprofile inhaltliche Aspekte relevanter Dokumente oder Angebote erfassen, mit einem Ansatz zur kollaborativen Filterung kombiniert. Dieser Ansatz gruppiert Benutzer in Abhängigkeit ihrer Interessen und Motive und leitet daraus Empfehlungen innerhalb solcher virtuellen Gemeinschaften ab. Weil kollaborative Filterverfahren nur auf den von Benutzern geäußerten Meinungen basieren, sind sie besonders gut in Anwendungsfeldern einsetzbar, in denen der persönliche Geschmack eine große Rolle spielt, wie etwa bei der Auswahl von Belletristik, Spielfilmen oder Fernsehsendungen. Automatische Lernverfahren werden eingesetzt, um Benutzercharakterisierungen mit Vorlieben und Kaufentscheidungen zu assoziieren.

Abbildung 3: Die COGITO Systemarchitektur

Im Allgemeinen wird ein Benutzer nicht auf Anhieb geeignete Einträge in der Produktdatenbank finden. Um langwierige Anfragemodifikationen zu vermeiden, bezieht der Retrievalprozess in COGITO vier Komponenten ein (siehe Abbildung 3): den Chatterbot, das Content-Management der E-Commerce-Site, die Profilgenerierungsmaschine und ein auf den Benutzerprofilen basierendes Erweiterungssystem für Anfragen, „Prompter“ genannt, das dem Agenten Erweiterungsvorschläge für Datenanfragen in der Produktdatenbank „souffliert“. Der Prompter setzt einen Regelinterpreter ein, um die (XML-) Struktur der gesuchten Objekte oder Dokumente ebenso in die Präzisierung

des Suchprozesses einzubeziehen wie die Präferenzen der Benutzergruppen, die dem Profil des aktuellen Benutzers entsprechen.

Der Prompter realisiert außerdem den Zugriff auf den Thesaurus, der auf den BOL Bücherkategorien anhand des Layer-Seeds Verfahren aufgebaut ist, wobei Titel und Beschreibungen der Bücher als Dokumente für den Thesaurusaufbau verwendet werden. Basierend auf den gegenwärtigen Dialogbeiträgen und auf dem sich daraus ergebenden Kontext, wird vom Chatterbot automatisch eine aus Schlüsselwörtern bestehende Datenanfrage generiert. Diese wird vom Prompter expandiert, indem relevante Begriffe aus dem Thesaurus gewonnen werden. Im Folgenden wird dieser Prozess detaillierter aufgeführt.

Es wird angenommen, dass eine Anfrage aus n Stichwörtern besteht; für diese Anfrage ergeben sich m Ergebnisse; k ist die maximale Anzahl von Ergebnissen, die dargestellt werden sollen; k ist kleiner als m.

a. Jedem Stichwort wird eine Kategorie zugewiesen, in der das Stichwort die größte Gewichtung hat.
b. In der jeweiligen Kategorie werden für jedes Stichwort seine „relevanten" Begriffe im Thesaurus gesucht. Falls mehrere Stichwörter gemeinsame „relevante" Begriffe haben, werden die Similaritäten s0 von ihnen zusammenaddiert. Die gefundenen „relevanten" Begriffe werden in der Relevantbegriffliste gespeichert.
c. Der Begriff, der zum Schluss den größten Wert s0 besitzt, wird vom Thesaurus als Expansionswort vorgeschlagen.
d. Dieses Wort wird mit der ursprünglichen Anfrage mit „AND" verknüpft. Wenn die Anzahl der Ergebnisse gleich 0 ist, eignet sich das Wort nicht für die Anfrageexpansion. Es wird von der Relevantbegriffliste gelöscht. Schritt c wird wiederholt. Falls die Anzahl der Ergebnisse größer als 0 und kleiner gleich k ist, wird der Prozess beendet.

Die automatische Anfrageexpansion muss sehr vorsichtig angegangen werden, weil nur die Benutzer selbst genau wissen, welche Produkteigenschaften sie eigentlich suchen. Ohne Interaktion mit dem Benutzer ist es schwierig, unter verschiedenen Expansionsmöglichkeiten zu entscheiden. Die „Thematische Spezialisierung"-Beziehung im Thesaurus ist daher nicht zur automatischen Expansion verwendbar. Nur die Begriffe, wie z.B. der Name der Kategorie oder die relevanten Begriffe, können zur automatischen Expansion eingesetzt werden. Trotzdem konnten interessante Ergebnisse aus der prototypischen Anwendung des COGITO Systems erkannt werden. Diese sind im nachfolgenden Kapitel aufgeführt.

5 Evaluation

Die Evaluation eines Systems umfasst normalerweise drei prozedurale Ebenen:

- die Verifikation, bestehend aus der Überprüfung der Operations-Implementierung, die sich direkt aus den Benutzeranforderungen, sowie aus den operationalen Anforderungen ergibt;
- die Evaluierung, mit dem Ziel der Überprüfung der aus den Benutzeranforderungen und aus den prozeduralen Anforderungen abgeleiteten Funktionalitäten;
- und schließlich die Valutierung, welche eine Kontrolle des Systemnutzens anstrebt. Dabei wird untersucht, ob durch dessen Verwendung eine Steigerung der Effizienz und der Erfolgsrate erzielt werden kann.

Eine übliche Vorhergehensweise für die Evaluation eines Softwaresystems ist die top-down Methode. Valutierung und Evaluierung werden direkt vom interagierenden Benutzer durchgeführt. Falls dieser Test nicht erfolgreich ist, würde im nächsten Schritt nach der bottom-up Methode verfahren, wobei während der Verifikation als erstes die Implementierung der operationalen und dann der funktionalen Eigenschaften überprüft werden würde.

Die Evaluation des COGITO Systems basierte auf der top-down Methode, wobei die Interaktion zwischen Mensch und Maschine im Vordergrund stand [Eva02]. Insbesondere erfolgte die Evaluierung und die Valutierung des COGITO Agenten durch Gruppen von Testpersonen, welchen verschiedene Aufgaben, wie z. B. die Suche nach generischen Informationen oder mehr spezifische Produktinformationen, gestellt wurden. Die Evaluation basierte zum Teil auf quantitativen Maßen, wie z. B. die Länge der eingegebenen Sätze, die stereotypische Nutzung der Textausgaben des Agenten, und die Anzahl der „fallback“ Sätze, also Antworten auf nicht interpretierbare Benutzereingaben. Außerdem war die Evaluation qualitativ abhängig von den subjektiven Benutzermeinungen über die Systembedienung sowie über die Ergebnisse einer Informationssuche. Die Evaluation erfolgte durch direkte Befragung und Ausfüllen detaillierter Fragebögen über verschiedene Aspekte, wie z. B. Eindruck, Kontrolle, Effektivität, Navigationsfähigkeit, Lernfähigkeit, Benutzerfreundlichkeit und Verständlichkeit des Agenten.

Vier Gruppen von je acht Personen wurden eingesetzt für die Testsequenzen, unterteilt in zwei Gruppen von Anfängern und zwei Gruppen von erfahrenen Benutzern, damit die Effektivität des COGITO Prototypen für jede dieser Benutzertypen getestet werden konnte.

Damit eine Referenz für die Evaluation des proaktiven COGITO Agenten möglich war, wurde die BOL Website, ausgestattet mit einem marktüblichen Chatterbot, getestet und verglichen. Dieser Agent hatte ein standardmäßiges Gesprächspotential und wurde in die BOL Website integriert, indem einfache Produktverweise als einzige Verbindung zu der Produktdatenbasis implementiert waren, d.h. dieser Agent hatte keine proaktiven Eigenschaften.

Die Analyse des Konversationsprotokolls diente zur Bewertung des Gesprächverlaufs hinsichtlich der Anzahl korrekter Systemaussagen, „fall-back" Sätze und proaktiver Systemausgaben. (siehe Abbildung 4). Die Angabe „korrekte Systemaussage" basierte auf der manuellen Analyse und Interpretation von erfolgreichen Elementen der Mensch-Maschine Interaktion, bestehend aus einem Benutzer-Eingabesatz und dem entsprechenden Agenten-Ausgabesatz. Der COGITO Agent zeigt eine bessere Performanz als der BOL Agent bezüglich der korrekten Ausgabekategorie (61% vs. 47%). Es scheint, als ob der COGITO Agent über eine bessere Suchtermerkennung verfügt; er hat einen umfangreicheren Wortschatz, unter anderem wegen der Einbindung des Thesaurus. Eine andere Begründung liegt wahrscheinlich in der Fähigkeit proaktiver Aussagen, welche immer dann erfolgen, wenn das System aus Eigeninitiative Vorschläge und Antworten ausweist.

Der BOL Agent stellt selbstverständlich auch Fragen, die allerdings genereller und passiver sind und verwendet nicht die semantische Bedeutung der Benutzereingabe für die Folgenantworten.

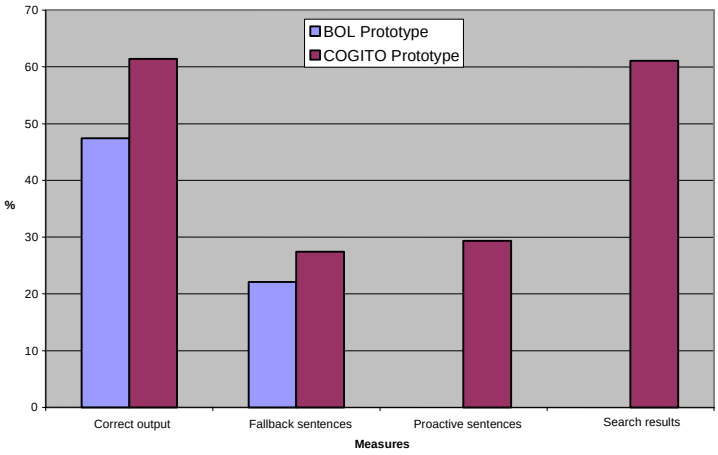


Abbildung 4: Die Ergebnisse aus der Analyse des Gesprächsprotokolls

Aus der Analyse der Suchergebnisse, welche sich aus der automatisch expandierten Anfrage mit Hilfe des Thesaurus ergeben, wird eine Steigerung der Effektivität um insgesamt 61% ersichtlich (siehe letzte Spalte der graphischen

Darstellung in Abbildung 4). Da der konventionelle Chatterbot nur mit den in seinen Regeln vordefinierten „frozen queries“ arbeiten kann, die durch vorher festgelegte Schlüsselbegriffe aktiviert werden, werden hier von den Kunden häufig gebrauchte Anfragen in stereotyper Weise benutzt. Dagegen erweitert der Prompter die als Dialogthemen auftretenden Begriffe unter Verwendung des Thesaurus dynamisch zu einer zur Dialogsituation passenden Anfrage.

6 Ausblick

Der in dieser Arbeit angefertigte Thesaurus befindet sich noch im Prototyp-Stadium. Er lässt sich auf mehrere Arten erweitern. Folgende Veränderungen sind in der Zukunft möglich:

1) Erweiterung der Datenbasis
   Die Datenbasis in dieser Arbeit beschränkt sich auf Deutsch und die BOL-Produktkategorien. Sie kann aber ohne großen Aufwand auf z.B. englische Webseiten umgestellt werden. Ferner besteht die Möglichkeit, Informationen aus allgemeinen kategoriebasierten Informationsportalen (z.B. Yahoo![2]) als Datenbasis für den Thesaurusaufbau hinzuzunehmen, um bessere und allgemeinere Ergebnisse erzielen zu können. Verschiedene Experimente werden durchgeführt, um die Schwellenwerte in den jeweiligen erweiterten Domänen festzustellen. Dabei wird nach universellen Schwellenwerten für alle Domäne gesucht.

2) Verbesserung des Thesaurusaufbaus
   Es gibt noch einige Verbesserungsmöglichkeiten für den eigentlichen Prozess des Thesaurusaufbaus. Dazu zählen z.B.:
   - Unterscheidung der Informationen in Texttitel und Textkörper, damit die Terme nach ihren Positionen gewichtet werden können
   - Erkennung und Einsetzung von Mehrwortbegriffen
   - Verkleinerung der Größe des „Fensters“, in dem die Terme als relevant betrachtet werden. In dieser Arbeit war das „Fenster“ auf das ganze Dokument gestellt

3) Verwendung der „Thematische Spezialisierung“-Beziehung für die automatische Query Expansion
   Durch den Einsatz eines Gesprächprotokolls, welches einen mehrstufigen Dialog zwischen Benutzern und dem System strukturiert aufzeichnet, lassen sich die Benutzerwünsche besser ermitteln. Dadurch wird

[2] http://www.yahoo.com.

nicht nur die „Relevant"-Beziehung, sondern auch die „Thematische Spezialisierung"-Beziehung für die automatische Query Expansion möglich.

4) Verwendung des Thesaurus zur interaktiven Query Expansion
Die interaktive Query Expansion wird durch das Einbeziehen der Benutzer bessere Ergebnisse erzielen als die automatische Query Expansion. Dabei muss die Verbindung zwischen dem Thesaurusmodul und dem Agenten verändert werden, um Benutzern verschiedene Expansionsauswahlen anhand der Hierarchie des Thesaurus anzubieten.

7 Literaturverzeichnis

[Bc92] Belkin, N.J. and Croft, W.B.: Information Filtering and Information Retrieval: Two Sides of the Same Coin? Communications of the ACM, 35(12), pages 29-38, December 1992.

[Bco00] Beschreibung des Projekts COGITO http://www.darmstadt.gmd.de/~cogito/Description.htm.

[Byf93] Frakes, W.B. and Bareza-Yates, R.: Information Retrieval Datastructures & Algorithms, Prentice-Hall, 1993.

[Byrn99] Bareza-Yates, R. and Ribeiro-Neto, B.: Modern Information Retrieval, Addison-Wesley, 1999.

[CrY92] Crouch,C.J., Yong,B., Experiments in automatic statistical thesaurus construction, SIGIR'92, 15th Int. ACM/SIGIR Conf. on R&D in Information Retrieval.

[Eft96] Efthimiadis, E. N.: Query Expansion, Annual Review of Information Systems and Technology(ARIST), v31, pp 121-187, 1996.).

[Ekm92] Ekmekcioglu, F.C., Robertson, A.M., Willett, P.: Effectiveness of query expansion in ranked output document retrieval systems, J. of Information Science. 1992

[Elna94] Elmasri, R., Navathe, S.: Fundamentals of Database Systems. Redwood City: The Benjamin/Cummings Publishing Co.,Inc., 1994.

[Eva02] Project Deliverable report, D7.2 Evaluation of the COGITO system.

[Ferb97] Ferber, R.: Data Mining und Information Retrieval. Vorlesungsskript 1997 http://www.muenster.de/~ferber/.

[From01] Frommholz, I.: Automatische Kategorisierung von Web-Dokumenten. Diplomarbeit, 2001 http://ls6-www.cs.uni-dortmund.de/bib/fulltext/world/Frommholz:01.ps.gz.

[Fuhr96] Fuhr, N.: Information Retrieval. Skriptum zur Vorlesung. Technical Report. Universität Dortmund, Fachbereich Informatik. 1996 http://ls6-www.cs.uni-dortmund.de/ir/teaching/courses/ir/.

[Göv99] Govert, N., Lalmas, M. and Fuhr, N.: A probabilistic description-oriented approach for categorizing web documents. In Susan Gauch and Il-Yeol Soong, editors, Proceedings of the Eighth International Conference on Information and Knowledge Management, pages 475–482, New York, 1999.

[Jicr94] Jing, Y. F. and Croft, W. B.: An Association Thesaurus for Information Retrieval, UMass Technical Report 9417.

[Klas00] Klas, P. and Fuhr, N.: A new effective approach for categorizing web documents. In Proceedings of the 22th BCS-IRSG Colloquium on IR Research, 2000.

[Kow97] Kowalski, G.: Information Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers: Boston/Dordrecht/London, 1997.

[Lath01] L'Abbate, M. & Thiel, U.: Chatterbots and intelligent Information Search. Proceedings of ECIR 2001, Darmstadt, Germany 2001.

[Mink72] Minker, J., Wilson, G.A. and Zimmermann, B. H.: An evaluation of query expansion by the addition of clustered terms for a document retrieval system, Information Storage and Retrieval, 8, 329-348 (1972).

[Panyr86] Panyr, Jiri; Automatische Klassifikation und Information Retrieval: Anwendung und Entwicklung komplexer Verfahren in Information-Retrieval-Systemen und ihre Evaluierung. (Sprache und Information, Bd. 12). Tueringen:Niemeyer, 1986, 416 S.

[Qiu93] Qiu, Y., Frei, H.P.: Concept based query expansion. In Proceedings of ACM SIGIR International Conference on Research and Development in Informaiton Retrieval.

[Rakr99] Robinson, T.; Abberley, D.; Kirby, D. and Renals, S.: Recognition Indexing and Retrieval of British Broadcast News with the THISL System. Proc. Eurospeech-99 Budapest, Hungary, 1067-1070 (1999).

[Rij79] Rijsbergen, C. J. Van: Information Retrieval. Butterworth, London, second edition, 1979.

[Rocchio71] Rocchio, J.Y. : Relevance Feedback in Information Retrieval. The SMART Retrieval System. Engelwood Cliff, N.J.: Prentice Hall, 313-323.

[Salton73] Salton, G., Comment on "an evaluation of query expansion by the addition of clustered terms for a document retrieval system". Computing Reviews, 14, 232 (1973).

[Samc83] Salton, G. and McGill M.: Information Retrieval - Grundlegendes für Informationswissenschaftler. 1983.

[Spa71] Sparck-Jones, K.: Automatic Keyword Classification for Information Retrieval. Butterworth, London 1971.

[Stef96] Steffens, U.: Integration von Information Retrieval Funktionalität in eine offene persistente Programmierumgebung. Informatik Mitteilung FBI-HH-M-257/96, Fachbereich Informatik, Universität Hamburg, Germany, April 1996.

[Thl00] Thiel, U.; Stein, A.; Semeraro, G.; Abbatista, F.; De Candia, L.; Fanizzi, N.; Candela, V.; Lops, P. & Valente, A.: COGITO - E-Commerce with Guiding Agents Based on Personalized Interaction Tools. In Proceedings of the AICA Annual Conference, Taormina, Italy, September 27-30, 2000.

[Xucr96] Xu, J. and Croft, W. B.: Query expansion using local and global document analysis, in Proc. ACM SIGIR, 1996.