C. Stary (Hrsg.): Mensch & Computer 2005: Kunst und Wissenschaft – Grenzüberschreitungen der interaktiven ART. München: Oldenbourg Verlag. 2005, S. 155-170

Ein integrierter Ansatz zur wissensbasierten Informationsrecherche

Christoph Kunz

Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO), Stuttgart.

Zusammenfassung

In diesem Beitrag wird ein nutzerzentrierter und ganzheitlicher Ansatz beschrieben, um große und vernetzte Informationsräume unter Einbeziehung der menschlichen Nutzung und Interpretation semantisch erschließ- und recherchierbar zu machen. Den Anwendungskontext stellen spezialisierte Dokumentkollektionen wie Intranets, komplexe Portalseiten sowie Wissensdatenbanken dar. Der Ansatz verbindet eine herkömmliche Volltextsuche mit einer semantischen Suche auf Grundlage einer lernenden Themenontologie, welche einen Informationsraum abstrahiert und verdichtet. Neue ontologische Begriffe werden dabei durch Beobachtung des Informationsablageverhaltens der Nutzer bei der Erstellung von Lesezeichen erzeugt. Die verschiedenen damit geschaffenen Arten des Informationszugriffs erfolgen über eine einheitliche Nutzerschnittstelle. Deren Hauptmerkmal ist die Visualisierung der Ontologie. Dies erlaubt die Bereitstellung von Stöbermechanismen auf Strukturebene, die Kontextualisierung von Volltextsuchergebnissen durch Begriffe der Ontologie sowie die visuelle Konstruktion semantischer Abfragen zur präzisen Suche. Mit Hilfe dieser Mechanismen lässt sich ein Informationsbedarf iterativ präzisieren. Eine umfangreiche Benutzbarkeitsanalyse zeigt Leistungsvorteile der Nutzerschnittstelle und belegt deren Nützlichkeit in dem Anwendungsbereich. Innerhalb einer weiteren Studie konnte die Machbarkeit und hohe Qualität der kollaborativen Wissensakquise durch Beobachtung des Ablageverhaltens nachgewiesen werden.

1 Einleitung

Dokumente im Internet als auch in betrieblichen Kontexten (Intranet, Portal, Wissensdatenbank) liegen selten in strukturierter Form vor und besitzen fast keine Metainformationen bezüglich ihres Inhalts. Schon die schiere Menge, aber auch das Nichtvorhandensein einer einheitlichen Zugangsstruktur erschwert sowohl deren menschliche als auch deren maschi-

nelle Verarbeitung und Wahrnehmung, also das Entdecken relevanter Information, deren Zusammenhang und deren Synthese zu einem umfassenden Gesamtbild.

Gerade in Organisationen des Wirtschaftssystems wird es inzwischen als unabdingbar angesehen, durch einen schnellen und reibungslosen Informationszugriff eine Erhöhung organisationaler Lern- und Wandlungsfähigkeit zu fördern, um dadurch auf eine immer turbulenter werdende Umwelt reagieren zu können (Wolf 2003). Dafür werden Instrumente wie Intranets oder Wissensdatenbanken eingesetzt, um vorhandenes (bereits kodifiziertes) Wissen zu konservieren und zugänglich zu machen. Hierzu sind effiziente und nutzerfreundliche Zugriffs- und Recherchemechanismen von besonderer Bedeutung, um eine schnelle Lokalisierung von gerade wichtigen Informationen zu ermöglichen.

2 Ansätze zur Informationsrecherche

Die mit der Anglizisme Information Retrieval (IR) bezeichneten Verfahren der Erschließung und des gezielten Zugriffs auf gespeicherte Informationen lassen sich im Wesentlichen in drei Kategorien unterteilen:

1. Für den menschlichen Zugang zu Informationsräumen existiert der aus dem Bibliothekswesen bekannte Ansatz der Systematik (Hacker 1992), in dem Inhalte nach Themengebieten hierarchisch katalogisiert und nach Autoren, Erscheinungsjahr, etc. sortiert werden. Im übertragenen Sinn und stark vereinfacht folgen auch Portalsysteme (Gurzki et al. 2002) diesem Ansatz. Der Nutzer eines solchen Zugangs erhält eine navigierbare Struktur, mit der er nach dem Prinzip des kontinuierlichen Vordringens bis auf die gewünschte Ressourcenebene vordringen kann. Dies setzt aber hochgradige redaktionelle Vorarbeiten der Systematik voraus. Zudem sind die gebotenen Zugangsmechanismen, wie rollen- und aufgabenbasierte oder hierarchische Kategoriensysteme, zu starr, um flexible und individuelle Sichten auf einzelne Inhalte zu bieten und somit ein Gesamtbild über eine Menge von Inhalten zu vermitteln.

2. Auf Seiten der maschinellen Verarbeitung hat die herkömmliche Volltextsuche (Van Rijsbergen 1979) die weiteste Verbreitung gefunden, obwohl sie der vernetzten Natur von Informationsräumen nur wenig Rechnung trägt. Damit lässt sich fast keine Information über die Struktur und Beziehungen der Inhalte untereinander gewinnen. Aus Sicht der Nutzer sind viele bestehende Information Retrieval Systeme schwer zu nutzen, da Informationsbedürfnisse schwer formulierbar und schwierig in Systemanfragen auszudrücken sind (Belkin 1980). Nicht ungewöhnlich ist die Eingabe von Suchtermen, welche mit den Indextermen der Retrievalsysteme nicht übereinstimmen und daher keine Ergebnisse liefern. Auf zu unspezifische Suchterme reagieren sogar modernste Systeme wie *Google* mit Suchergebnissen im Millionenumfang.

3. Als neuerer Ansatz erhält die Vision des *Semantischen Netzes* (Berners-Lee 1998) derzeit viel Beachtung. Der allgemeine Grundgedanke ist dabei die Anreicherung der dargebotenen Inhalte des bestehenden Internets mit Metadaten, so dass deren inhaltliche Bedeutung

(Semantik) von Automaten versteh- und interpretierbar wird. Bezogen auf IR werden hierbei Informationsräume mittels komplexerer Wissensstrukturen wie Semantischen Netzen oder Ontologien repräsentiert (Farquhar et al. 1996). Ziel dieser, aus dem Gebiet der Künstlichen Intelligenz stammenden Verfahren ist es, Wissen unabhängig von der sprachlichen "Oberflächenstruktur" zu modellieren, aus dem dann automatische Schlussfolgerungen abgeleitet werden können (Nie 1992). Dies eröffnet eine Vielzahl neuer Möglichkeiten bei der maschinellen Verarbeitung von Inhalten. Gerade Wortambiguitäten, das Hauptproblem der volltextbasierten Suche, können durch Einbeziehung des Arbeitskontexts aufgelöst werden (Studer et al. 2001).

Suchmethoden auf Basis von Ontologien versprechen zwar einige Vorteile gegenüber merkmalsbasierten Ansätzen, welche Wörter als bloße Zeichenketten auffassen. Nachteilig wirkt sich jedoch aus, dass die zugrunde liegende Wissensbasis nur schwer automatisiert gebildet werden kann (Maedche&Staab 2000; Heyer et al. 2001), da sie für automatische Schlussfolgerungen widerspruchsfrei sein muss. Zudem muss die Informationsarchitektur sowohl eine hohe Qualität der Struktur als auch gemeinsame Akzeptanz der Nutzer aufweisen und bedarf daher ebenfalls eines erheblichen redaktionellen Aufwands (Uschold&Grüninger 1996, Guarino 1995). Beim direkten Abfragen der Wissensbasis können nur exakte Treffer erzielt werden, da die Suchterme mit der ontologischen Terminologie abgeglichen werden (Sebastiani 1996). Nur wenn die Suchbegriffe in der Wissensbasis enthalten sind, werden Antworten deduziert. Zudem erfolgt keine Sortierung der Ergebnisinformation nach Wichtigkeit, da Ähnlichkeiten und Klassenzugehörigkeiten aufgrund der meistens verwendeten, booleschen Wahrheitsfunktion nicht gewichtet werden können. Daher werden die Begriffe in der Ontologie meist nur zur Verfeinerung einer Suchanfrage genutzt, die an eine herkömmliche Volltextsuchmaschine weitergeleitet wird.

Ein weiteres Problem ergibt sich aus der Komplexität semantischer Ansätze, nämlich das der Gestaltung geeigneter Nutzerschnittstellen, welche die eigentlichen Vorteile erst nutzbar machen und den Informationszugriff ermöglichen. Gestaltungsfragen betreffen die Darstellung der in einer Ontologie formalisierten Informationsstruktur (Furnas 1986, Lamping/Rao 1994) sowie einfache Interaktions- und Visualisierungstechniken zur Konstruktion logischer Abfragen (Jones 1998; Catarci et al. 2003).

3 Ein integrierter Ansatz zur wissensbasierten Informationsrecherche

3.1 Anforderungen

Für die Entwicklung eines geeigneten Ansatzes zur Informationsrecherche ist es essentiell, das Problem des Nutzers bei dem Umgang mit einem Suchsystem zu verstehen. Dieses besteht darin, den Informationsbedarf des Nutzers mit dem Informationsangebot des Systems zu decken. Für das System bedeutet dies, den Bedarf des Nutzers zu erkennen und möglichst

effizient auszugleichen. Im Anwendungsfall sollte auf eine Suchanfrage, durch die der Informationsbedarf ausgedrückt wird, jede relevante und keine überflüssige Information aufgefunden und als Suchergebnis angeboten werden. Überflüssige Information ist hier diejenige, welche nicht zur Deckung des Informationsdefizits benötigt wird.

Der *Informationsbedarf* des Suchenden ist definiert als "die Art Menge und Qualität der Information, die eine Person zur Erfüllung ihrer Aufgabe in einer bestimmten Zeit benötigt" (Picot et al. 2003). Er gliedert sich in einen objektiven und subjektiven Teil. Der so genannte *objektive* Informationsbedarf gibt an, welche Art und Menge von Information tatsächlich zur Erfüllung einer bestimmten Aufgabe benötigt wird. Der *subjektive* Informationsbedarf geht hingegen von der Sichtweise der jeweiligen Person aus und gibt diejenige Information an, welche dieser zur Bearbeitung der Aufgabe als relevant erscheint. Der objektive und subjektive Informationsbedarf sind oft stark unterschiedlich, vor allem bei schwer zu spezifizierenden, komplexen Aufgaben oder bei unbekannten Strukturen und Zusammenhängen des Informationsraums. Die Menge an Information, die letztlich tatsächlich nachgefragt wird, ist wiederum lediglich eine Teilmenge des ursprünglich geäußerten Informationsbedarfs. Demgemäss ist es die Aufgabe des Systems, den subjektiven Informationsbedarf dem objektiven anzunähern. Nur der Bereich, in dem die tatsächliche Informationsnachfrage und das Informationsangebot zusammenfallen, führt schließlich zu einer konkreten Informationsversorgung.

An einen Lösungsansatz sind daher die folgenden Anforderungen zu stellen: Es müssen effektive Zugangs- und Abfragemechanismen geboten werden, welche das gesamte Spektrum von Suchaufgaben unterstützen. Dieses reicht von der allgemeinen Recherche von Themengebieten bis zur gezielten Lokalisierung gewünschter Information. Des Weiteren ist ein möglichst gutes Verständnis des Informationsraumes und des eigenen Informationsbedarfes zu fördern. Erst durch ein inhaltliches Begreifen des Informationsraumes kann es gelingen, Anfragen so zu stellen, dass gewünschte Ergebnisse berechnet werden können. Jedoch die richtigen Anfragen zu stellen, welche zur einer tatsächlichen Informationsversorgung führen, wird nur durch ein Verstehen des eigenen objektiven Informationsbedarfs ermöglicht. Der objektive Informationsbedarf wiederum ist nur im Dialog und durch Rückkopplung unter den Kommunikationspartner zu ermitteln. Daher ist auch die Ergebnismenge einer Suchanfrage so aufzubereiten, dass sie leicht und schnell erfasst sowie in ihrem Inhalt begriffen werden kann. Die Bandbreite der Kommunikation mit dem Suchsystem wird damit erhöht.

3.2 Integrierter Ansatz

Der gewählte Ansatz ist in Abbildung 1 als Übersicht dargestellt. Er integriert eine neuartige Schnittstelle zwischen Mensch und Maschine, einer ebenfalls neuartigen Komponente zur Beobachtung des Ablageverhaltens der Nutzer, einer probabilistischen Deduktionskomponente sowie einer herkömmlichen Volltextsuchmaschine zu einem ganzheitlichen Rahmenwerk.

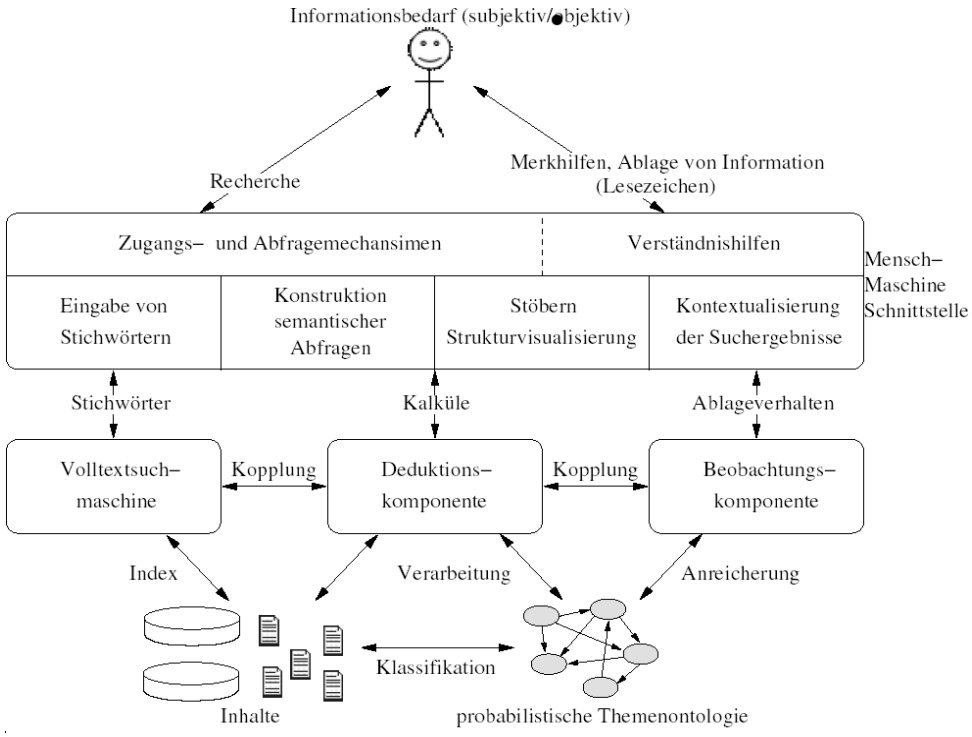


Abbildung 1: Integrierter Ansatz zur Informationsrecherche

Das Hauptmerkmal der Nutzerschnittstelle ist die Visualisierung der semantischen Struktur des Informationsraums zur Unterstützung eines besseren Verständnisses desselben. Im Minimalfall beinhaltet die semantische Struktur ein dem Informationsraum zugehöriges Themennetzwerk sowie bedeutungsvolle Merkmale von Dokumenten (Autor, Publikationsdatum, etc). Zusätzlich lassen sich in dieser Struktur weitere Arten von Inhalten wie zum Beispiel Personaldaten, Organigramme oder Produktdaten einbetten. Eine Formalisierung erfolgt durch eine mit Wahrscheinlichkeiten erweiterte Ontologie. Mittels Wahrscheinlichkeiten können Begriffsüberdeckungen sowie Zugehörigkeitswerte von Inhalten zu Themen bei der Modellbildung der Ontologie und der Abfrage berücksichtigt werden. Die maschinelle Verarbeitung der ontologischen Wissensbasis erfolgt durch eine entsprechend ausgelegte Deduktionskomponente. Eine Volltextsuchmaschine ist mit der Deduktionskomponente derart integriert, dass sich zu einzelnen Suchergebnissen sämtliche semantischen Merkmale aus der Ontologie und andersherum abfragen lassen.

Die Strukturvisualisierung ist mit einer Präsentation von Suchergebnissen als Liste zu einem wählbaren Thema nach dem Prinzip der direkten Manipulation gekoppelt. Zusätzlich sind damit stichwortbasierte und semantische Abfragemöglichkeiten verbunden. Dadurch werden drei Arten von Informationszugängen geschaffen: Auf Strukturebene kann navigational

durch Verbreiterung und Einengung eines Themas sowie der Verfolgung von bedeutungsvollen Querbeziehungen gestöbert und unmittelbar die zugehörigen Inhalte in der Liste eingesehen und recherchiert werden. Eine Volltextsuche ermöglicht eine ungenaue Suche auf Basis von Stichwörtern. Mittels semantischer Abfragen lässt sich Information anhand von Themen und Merkmalen lokalisieren. Semantische Abfragen können visuell konstruiert werden, womit die Erlernung einer logischen Abfragesprache entfällt. Eine gleichzeitige Darstellung von Suchergebnissen und semantischer Struktur ermöglicht zudem eine Kontextualisierung der Ergebnisse in dieser Struktur. Über den Themenbezug lässt sich so interaktiv eine ungenaue Suche weiter verfeinern. Durch diese drei Arten kann ein bestimmter Informationsbedarf iterativ präzisiert werden.

Innerhalb der Nutzerschnittstelle ist, wie in vielen anderen Informationssystemen, die Möglichkeit vorgesehen, Merkhilfen in Form von Lesezeichen zu erstellen. Diese Lesezeichen ermöglichen eine persönliche Klassifikation öfters benötigter, relevanter Inhalte. Zur systemischen Wissensbildung existiert eine weitere Komponente, die das Ablageverhalten von Information der Nutzer beobachtet. Durch die Analyse dieser Lesezeichen werden neue Begriffe und Relationen der Ontologie gelernt beziehungsweise generiert.

3.3 Rechercheschnittstelle

Die prototypische Umsetzung dieser Ideen zeigt Abbildung 2. Das Hauptfenster der Anwendung teilt sich dabei in drei wesentliche Bereich auf: Am oberen Fensterrand befindet sich ein Texteingabefeld mit einem Anfrageknopf zur Formulierung und Durchführung von stichwortbasierten Suchanfragen. Der mittlere Bereich ist für die Visualisierung von Klassen der Ontologie (links) und für die Präsentation der Suchergebnisse (rechts) vorgesehen, während am unteren Rand des Anwendungsfensters ein Feld zur visuellen Konstruktion und Ausführung semantischer Anfragen platziert ist.

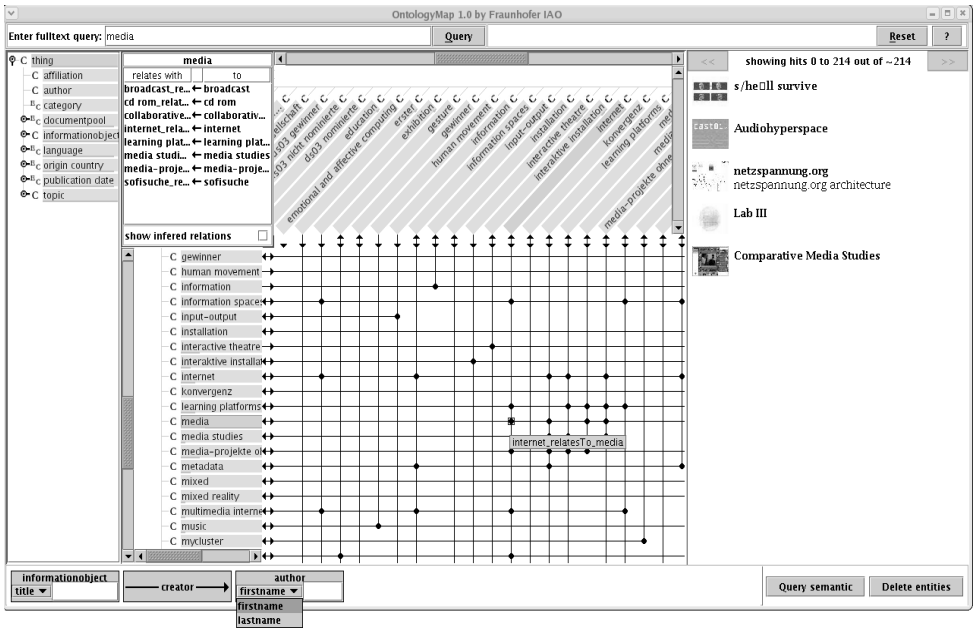


Abbildung 2: Rechercheschnittstelle

3.3.1 Visualisierung von Ontologien mit dem MatrixBrowser

Grundsätzlich lässt sich eine Ontologie als vernetzte Begriffsstruktur mit einem Graph darstellen. Dabei sind die einzelnen Begriffe die Knoten des Graphen, während die semantisch ausgezeichneten Relationen die Kanten des Graphen repräsentieren. Die einem Informationsraum zugrundeliegende Ontologie wird mit Hilfe eines alternativen Graphvisualisierer, dem MatrixBrowser (Ziegler/Kunz/Botsch 2002) dargestellt. Grundidee ist hierbei den Ontologiegraphen auf eine interaktive Adjazenzmatrixdarstellung abzubilden. Die Teile des Ontologiegraphen, welche durch Subklassenbeziehungen verbunden sind, werden dabei mittels Windows-Explorerähnlichen Baummetaphern auf den Achsen der Matrix präsentiert, während die restlichen Beziehungen die Zellen der Matrix bilden. Durch Verwendung von Pfeilen und graphischen Symbolen können sowohl die Richtung der Relation als auch verschiedene syntaktische Typen von Relationen innerhalb der Matrixzellen visualisiert werden. Unterstützend wird zusätzlich die Technik der ToolTips eingesetzt, um die unterschiedlichen textuellen Bezeichnungen der Beziehungstypen zu beschreiben.

Mit dem MatrixBrowser können die Knoten und Teilhierarchien, die an den Achsen dargestellt werden, auf flexible Art aus einer Obermenge ausgewählt und gefiltert werden. So lassen sich z. B. diejenigen Knoten auf einer beliebigen Achse herausfiltern, die nicht durch Relationen verbunden sind. Ebenfalls wird eine Sortierung der Knotenmengen anhand unterschiedlicher Kriterien wie die alphabetische Sortierung ermöglicht.

Zusätzlich befindet sich eine listenbasierte, interaktive Visualisierung in der linken oberen Ecke der Matrix. Diese zeigt alle Relationen und direkten Nachbarn desjenigen Knoten im Netz, welcher in einer der beiden Bäume angewählt wurde. So ist nicht nur der Kontext des jeweiligen Knoten innerhalb einer taxonomischen Substruktur zu erkennen, sondern auch innerhalb des gesamten Netzes.

3.3.2 Kopplung von Matrix- und Listendarstellung

Innerhalb der Rechercheschnittstelle ist eine Trennung von Klassen- und Instanzebene vorgenommen: Während im MatrixBrowser die schematische Ebene dargestellt ist, werden in einer listenbasierten Präsentationsform nur Instanzen der Ontologie visualisiert. Die listenbasierte Form der Präsentation beruht auf einer Studie von Sebrechts et al (Sebrechts et al. 1999). In dieser Studie wurde der tatsächliche Wertvorteil von unterschiedlichen Suchergebnispräsentationen des Information Retrievals untersucht. Verglichen wurden dabei Visualisierungstechniken in allen drei Dimensionen. In der Untersuchung stellte sich heraus, dass die allgemein gebräuchliche flache Textliste den anderen Techniken unter Berücksichtigung des Vorwissens und Erfahrung der Nutzer vorzuziehen ist.

Die auf Schemaebene der Ontologie operierende Matrixvisualisierung und als Liste präsentierte zugehörige Individuenmenge sind nach dem Prinzip der direkten Manipulation gekoppelt. Mit Hilfe einer speziellen Art der Knotenselektion, dem Doppelklick mit der Maus, werden die Individuen der korrespondierenden Begriffsbeschreibungen abgefragt und in der Liste angezeigt.

Der tatsächliche Inhalt der mit der Liste dargestellten Individuen oder Suchergebnisse lässt sich durch die Selektion eines Listeneintrags per Doppelklick mit der Maus einsehen. Es wird dann ein weiteres Fenster geöffnet, welches eine detaillierte Beschreibung des jeweiligen Suchergebnisses zeigt.

3.3.3 Kontextualisierung von Suchergebnissen

Die Kombination aus Strukturvisualisierung und Instanzmenge lässt sich für eine thematische Kontextualisierung von Ergebnissen einer Suchanfrage heranziehen. Dies ist besonders bei stichwortbasierten Suchanfragen vorteilhaft, da dort von der Volltextsuchmaschine, aufgrund von gleichen Stichwörtern, Dokumente unterschiedlicher Themenzugehörigkeit zurückgeliefert werden können.

Wird eine Suchanfrage vom System evaluiert, werden die Ergebnisse der Suche in der Liste dargestellt. Für jedes einzelne Suchergebnis wird gleichzeitig durch Anfrage der Wissensbasis die Menge der Klassen ermittelt, in welchen das Suchergebnis eine Instanz darstellt. Für jede Klasse wird ein jeweiliger Zähler inkrementiert. Nach der Bearbeitung aller Suchergebnisse enthält so der Wurzelknoten des Ontologiegraphen die Anzahl aller Suchergebnisse. Die Knoten, welche den zugehörigen Klassen entsprechen, werden dann in der Matrix mit einem Balken markiert (vgl. Abbildung 2 Matrixachsen). Die Länge des Balkens gibt das Verhältnis von den, in dieser Begriffsbeschreibung gefundenen Treffern, zu der Gesamtzahl

aller Treffer an. Auf diese Weise erhält der Nutzer eine Rückkopplung, wie die erzielten Treffer einer Suche mit der Anfrage korrelieren.

In diesem interaktiven Zustand der Anwendung lässt sich dann die Menge aller Suchergebnisse über den Bezug zur Ontologie einschränken. Wird ein Knoten in der Matrix selektiert, so werden nur diejenigen Treffer in der Liste angezeigt, die Mitglied der Menge sind, die mit dem angewählten Knoten beschriebenen werden.

3.3.4 Visuelle Konstruktion semantischer Suchanfragen

Neben der Visualisierung einer Ontologie kommt der Hilfestellung bei der Konstruktion von logischen Abfragen eine besondere Bedeutung zu. Zur visuellen Konstruktion einer semantischen Anfrage ist eine spezielle Anfrageleiste mit der Matrix- und Listendarstellung so verbunden, dass sämtliche visualisierten Ontologieelemente auf dieser Leiste über den Mechanismus des Ziehen&Fallenlassens platziert werden können. Ist auf eine solche Art die gewünschte Anfrage formuliert, kann die gesamte Anfragebeschreibung mit einem Druckknopf evaluiert werden. Ein weiterer Druckknopf löscht alle in der Anfrageleiste platzierten Objekte wieder heraus.

Das Beispiel in Abbildung 2 zeigt die visuelle Repräsentation der Anfrage "alle Informationsobjekte, die einen Autor besitzen”. Die in der Anfrageleiste platzierten Ontologieelemente werden als Rechtecke dargestellt. Zusätzlich werden die jeweiligen lexikalischen Bezeichner der Ontologieelemente in einer ausgewählten Sprache angezeigt. Da Klassen mit Attributen versehen werden können, entsteht deren visuelle Repräsentation durch ein komplexeres Gebilde: Jedes Attribut bildet durch seinen lexikalischen Bezeichner einen Eintrag in einer Auswahlliste. Innerhalb eines korrespondierenden Texteingabefeldes kann der tatsächliche Wert des Attributs, wie der Nachname eines Autors, eingegeben werden. Dabei hängt der einzugebende Wert von dem jeweiligen Datentyp ab, der mit dem Attribut verknüpft ist.

Zur weiteren Unterscheidung der in der Abfrageleiste platzierten Ontologieelemente werden Instanzen nur durch ihren lexikalischen Bezeichner innerhalb des Rechtecks ausgewiesen, während Relationen zwischen Klassen durch einen zusätzlichen Pfeil visualisiert werden. Die Richtung des Pfeils zeigt dabei auf diejenige Klasse oder Instanz, welche zum Definitionsbereich der jeweiligen Relation gehört. Die Anordnung der Ontologieelemente ist dabei beliebig.

3.4 Kollaborative Wissensakquise

Das hier vorgestellte Verfahren des kollaborativen Indexierens unterstützt die initiale Erstellung und Pflege von Ontologien, in dem neues Systemwissen aus den Artefakten generiert wird, welche in den Kommunikationsprozessen zwischen den Nutzern und dem Recherchesystem entstehen. Diese Artefakte beziehen sich auf Informationsablagestrukturen einzelner Nutzer, in denen relevante Inhalte für diese Nutzer organisiert sind. Die Erzeugung von Ablagestrukturen seitens der Nutzer

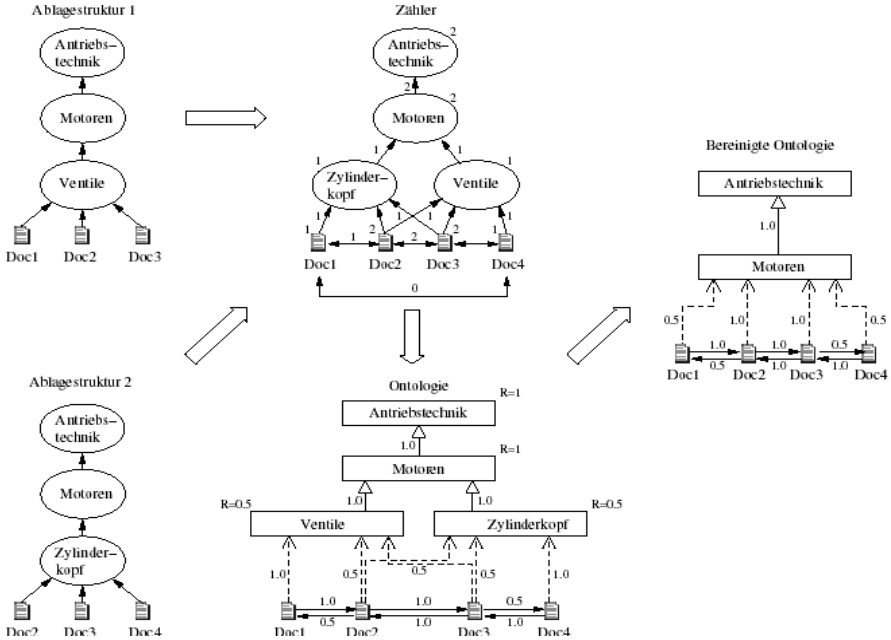


Abbildung 3: Schematischer Ablauf des Algorithmus

stellen Kontexte dar, in denen diese ihr Wissen über vorhandene Inhalte explizieren. Daher ist in der Nutzerschnittstelle eine spezielle Möglichkeit zur Ablage und Strukturierung von Lesezeichen vorgesehen.

Ablagestrukturen manifestieren sich als Lesezeichen. Lesezeichen bestehen aus Referenzen auf bestehende Inhalte eines Informationsraums, welche sich zu einer besseren Wiederauffindbarkeit in thematisch aufgebauten und hierarchisch strukturierten Ordnern zusammenfassen lassen. Somit können Ablagestrukturen als persönliche Klassifikationssysteme betrachtet werden, welche für den jeweiligen Nutzer relevante Informationsobjekte beschreiben und kategorisieren.

Aus der statistischen Analyse solcher Klassifikationssysteme lassen sich neue Aussagen über Inhalte und Begrifflichkeiten sowie deren Ähnlichkeit ableiten. Den prinzipiellen Ablauf des Algorithmus zeigt Abbildung 3. Auf ontologischer Seite ergeben sich aus den Ordnern einer Ablagestruktur Klassenbeschreibungen für die, mit den Verweisen korrespondierenden, Instanzen. Aus den Ordner-Unterordner Hierarchien werden Subsumtionsaxiome für diese Klassen. Die Überlagerung der Ablagestrukturen (Anzahl gemeinsamer Instanzen) liefert entsprechende Beziehungen zwischen Klassen und zwischen einzelnen Individuen. Die bedingten Wahrscheinlichkeiten der Instanz-, Subsumtions- und Querbeziehungen zwischen Instanzen und Klassen werden durch eine statistische Analyse aller Ablagestrukturen ermittelt. Eine Filterung von ungenauen und ungewöhnlichen Begriffen erfolgt auf Basis eines Schwellwerts der Häufigkeit von Ordnungsbezeichnern. Damit ergibt sich die ebenfalls in Abbildung 3 gezeigte bereinigte probabilistische Ontologie. Die einzelnen Verwandtschafts-

beziehungen und -typen der Dokumente untereinander können so direkt zu Suche verwendet werden.

4 Evaluation

Die Evaluation des Ansatzes erfolgte sowohl auf der gestalterischen Ebene der Visualisierungs- und Interaktionswerkzeuge zur Gebrauchtauglichkeit als auch auf der Verfahrensseite zur kollaborativen Wissensaqukise, wobei die Qualität der generierten probabilistischen Ontologien im Hinblick auf die Benutzbarkeit im vorgestellten Rahmenwerk zu bewerten waren.

4.1 Rechercheschnittstelle

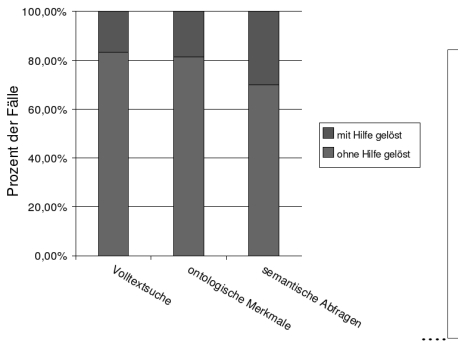
Die Gestaltung des Tests zur Bewertung der Rechercheschnittstelle setzte sich aus einem qualitativen Interview, einem Fragebogen nach DIN EN ISO 9241 "Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten" sowie eigens dafür zusammengestellten Aufgaben zusammen. Bei diesen wurde die Bearbeitungszeit gemessen. Die ersten Aufgaben entfielen auf die Volltextsuche, in denen Themen recherchiert und Mehrdeutigkeiten eines Suchbegriffs aufgelöst werden mussten. Die Lösung dieser Aufgaben erfolgte vergleichend sowohl mit dem Suchwerkzeug Google[1] als auch mit einer Version des Prototypen, dessen Volltextsuche ebenfalls auf Google beruhte. Als Ontologie kam eine semantisch aufbereitete Teilstruktur des Open Directory Projects[2], eines frei verfügbaren Webkatalogs, zum Einsatz. Weitere Aufgaben galten der Bestimmung ontologischer Merkmale wie Attribute (z.B. Titel oder Autor) von Informationseinheiten und Relationen von Klassen. Eine weitere Aufgabe befasste sich mit der Konstruktion einer semantischer Suchabfrage. Insgesamt nahmen zehn Teilnehmer an der Evaluationsstudie teil, welche gleichermaßen in Experten- und Laiennutzer unterteilt waren. Letztere waren zumindest im Umgang mit dem Internet vetraut.

Der prozentuale Anteil der Fälle für die verschiedenen Aufgabengruppen, bei denen die Aufgaben mit dem Prototypen ohne beziehungsweise mit Hilfestellung gelöst wurden, zeigt Abbildung 4 (a). Aus der Grafik wird deutlich, dass zwar alle Aufgaben von den Teilnehmern gelöst wurden, jedoch in einigen Fällen bei manchen Aufgaben Hilfestellung geleistet werden musste. Diese bezogen sich auf die interaktiven Merkmale der Rechercheschnittstelle (Bedeutung der Länge der Balken, Doppelklick auf Knoten), Begrifflichkeiten von Ontologien (Bedeutung von Relationen, Attributen und Klassen) und die Syntax bei der Konstruktion semantischer Suchabfragen (Klassen, Relationen). Innerhalb der Aufgabengruppen ist ein Lerneffekt festzustellen, da meist nur bei der ersten Aufgabe innerhalb einer Gruppe Hilfestellung gegeben werden musste. Ebenfalls ist zu bemerken, dass in der Laiengruppe öfters

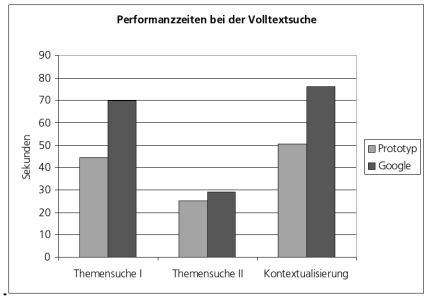
1 http://www.google.de

2 http://www.dmoz.org

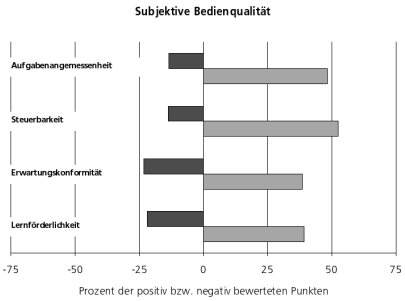
Hinweise gegeben werden mussten als in der Expertengruppe, was zu erwarten war. Der Unterschied ist allerdings nicht signifikant.



(a) Aufgabenlösungen nach Aufgabengruppen



(b) Performanzzeiten Volltextsuche



(c) Subjektive Bedienqualität des Prototypen nach ISO 9241

Abbildung 4: Ergebnisse der Evaluationsstudie des Gesamtkonzepts

Abbildung 4 (b) zeigt die durchschnittliche Bearbeitungszeit der Volltextsuchaufgaben mit dem Prototypen und Google. Dabei zeigen sich teilweise ausgeprägte Performanzvorteile durch den Prototypen. Auch hier ist der Unterschied in den Performanzzeiten zwischen der Laien- und Expertengruppe nicht signifikant.

Die Ergebnisse des Fragebogens zur allgemeinen Benutzungsfreundlichkeit nach ISO 9241 des Prototypen sind in Abbildung 4 (c) dargestellt. Die roten Balken können als relative Schwächen des Produkts und die grünen Balken als relative Stärken des Produkts bewertet werden. Die Bewertung der subjektiven Bedienqualität fällt insgesamt sehr positiv aus. Be-

sonders gut schneiden die Dimensionen Aufgabenangemessenheit und Steuerbarkeit ab. Diese sind für die Beantwortung der Frage nach der Nützlichkeit der Hauptfunktionalitäten des Recherchekonzepts von besonderer Bedeutung, da sie diese belegen. Die überwiegend positiv bewerteten Dimensionen Erwartungskonformität und Erlernbarkeit zeigen die Intuitivität der Benutzungsoberfläche. Dem gegenüber stehen allerdings auch kritische Befunde. Dies begründet sich nach Meinung des Autors in der Komplexität des Ansatzes und der Bedienoberfläche sowie der zum Teil ungewohnten Interaktionskonzepte.

Diese Hypothese wird in den Äußerungen der Probanten im postaktionalen Interview bestätigt. So wurde auf die Frage, wie gut die Probanten bei der Bearbeitung der Aufgaben zurecht gekommen seien, besonders von der Laiengruppe die Vielzahl und Ungewöhnlichkeit der Funktionen des Prototyps angemerkt. Von beiden Gruppen wurde eine fundierte Einführung gefordert. Wurde jedoch Sinn und Zweck des Konzepts von den Probanten erkannt, bestätigen alle Teilnehmer die Nützlichkeit und Bedienbarkeit der Einzelfunktionen, Stöbern in ontologischen Merkmalen und Inhalten, Kontextualisierung von Suchergebnissen und semantischer Abfragen. Es erstaunt, dass auch hier die mittelmäßigen bis schweren Bedienbarkeitsbeurteilungen gleichermaßen auf die Laien- und Expertengruppe verteilt war.

Im Schnitt wurde die Einarbeitungszeit auf ungefähr zwei Stunden eingeschätzt. In dieser subjektiven Einschätzung zeigen sich wiederum Unterschiede zwischen der Laien- und Expertengruppe. In der Laiengruppe betrug die geschätzte mittlere Einarbeitungszeit rund drei Stunden, während die Expertengruppe diese auf ca. eine Stunde schätzte.

4.2 Kollaborative Wissensakquise

Zur Evaluation des Verfahrens zum kollaborativen Indexieren wurden Informationsablagestrukturen von sowohl einer heterogenen Nutzergruppe (24 Ablagestrukturen) für einen offenen Informationsraum (Internet) als auch einer homogenen Nutzergruppe (26 Ablagestrukturen) für einen abgeschlossenen Informationsraum (http://netzspannung.org, Fleischmann et al. 2001) in probabilistische Ontologien überführt. Für jede dieser Ontologien wurde eine bereinigte Ontologie erzeugt und zur Auswertung herangezogen, welche nur Klassen enthielt, die mindestens zweifach als Ordnungsbezeichner in Ablagestrukturen vorkamen.

Im offenen Informationraum konnten 83% der Klassenbezeichner als Themenbegriffe gewertet werden, welche zu 72% aus deutschen (z.B."Mathematik", "Kunst") zu 10% aus englischen Wörtern (z.B. "Games") gebildet wurden. Die restlichen 17% der Klassenbezeichner entfielen auf ungewöhnliche (umgangsprachliche bzw. orthographisch falsche) oder zu unspezifische Worte, welche für eine klar umrissene Klassifikation ungeeignet sind. Die Quote der korrekten Klassifikation von Instanzen betrug von 98%. Von den entstandenen Relationen konnten 99% als semantisch korrekt nachgewiesen werden. Eine Betrachtung der Subsumtionsbeziehungen ergab 98% richtige Hierarchien (z.B. "Institutionen" -> "Universität").

Im geschlossenen Informationsraum konnten 91% (30% deutsche, 20% englische) Themenbegriffe nachgewiesen werden. Die Klassifikationsquote erhöhte sich auf 100%. Ebenfalls konnten alle generierten Rollen als semantisch korrekt eingestuft werden. Subsumtionsbe-

ziehungen konnten hier nicht betrachtet werden, da die verwendete Schnittstelle (Novak et al. 2003) keine hierarchischen Lesezeichen erlaubte.

5 Fazit

Ingesamt zeigte die Evaluationsstudie des Ansatzes dessen überaus positive Aufnahme und Anwendbarkeit. Gerade die Nützlichkeit der drei Hauptmerkmale der Nutzerschnittstelle des Ansatzes wie die Navigation und Exploration von Informationsräumen auf Basis dessen ontologischer Abstraktion, die Möglichkeiten einer Volltext- und semantischen Suche sowie deren Integration in einer einheitlichen Benutzungsschnittstelle konnte belegt werden. Ebenfalls konnte gezeigt werden, dass diese gleichermaßen von Experten und durchschnittlichen Nutzern zu bedienen war, obwohl die letztere Gruppe von der Komplexität und Vielzahl der interaktiven Möglichkeiten zunächst verunsichert wurde. Bei einer wirtschaftlichen Verwertung des Ansatzes besteht daher die Notwendigkeit, Nutzer gründlich in die Basiskonzepte und Bedienung einzuführen. Jedoch scheint der Einlernaufwand relativ gering zu sein, nachdem dieser im Mittel auf ungefähr zwei Stunden geschätzt wurde.

Es sei bemerkt, dass sich zwar Performanzvorteile für die in der Evaluationsstudie gewählten Suchaufgaben gegenüber Google zeigen, diese decken jedoch nicht alle Anwendungsfälle für Google ab. Damit kann kein endgültiges Fazit im Vergleich der beiden Recherchemöglichkeiten gebildet werden.

Die Ergebnisse der Evaluierung des Verfahrens zum kollaborativen Indexieren belegen dessen Anwendbarkeit zur Wissensakquise. Zwar besitzen die mit den Ordnungsbezeichnern vergebenen Begrifflichkeiten nicht durchwegs hohe Verständlichkeit, jedoch besteht die Vermutung, mit einer deutlichen höheren Anzahl von Ablagestrukturen und der statistischen Filterung die Qualität noch steigern zu können. Im Vergleich zwischen offenen und geschlossenen Informationsräumen zeichnet sich eine Tendenz bei geschlossenen Informationsräumen zur allgemeinen besseren Güte der vergebenen Begriffe ab.

Am beeindruckendsten ist jedoch die äußerst hohe Güte der Klassifikation. So sind rund 99% der Individuen korrekt klassifiziert. Dies stellt einen weiten Performanzvorsprung gegenüber rein maschinellen Klassifikationsverfahren (z.B. Support-Vector-Maschine, Naive-Bayes-Klassifikator) dar, bei denen die Klassifikationsquote bei ca. 70% liegt. Ein weiterer Vorteil dieses Verfahrens stellt dessen Möglichkeit dar, hochwertige Begriffsinklusionen (Subsumtionsbeziehungen von Begriffsbeschreibungen) zu generieren. Auch hier kommen automatische Verfahren nicht an die vorliegende Performanzleistung heran.

Als abschließende Bewertung ist festzuhalten, dass auch mit dem hier vorgestellten Verfahren Ontologien nicht vollständig automatisiert erzeugt werden können. Da diese zur automatischen Schlussfolgerung eine hohe Güte besitzen müssen und keine Widersprüche beinhalten dürfen, ist immer eine intellektuelle Überarbeitung durch Menschen erforderlich. Jedoch lässt sich deren Erstellung und Pflege durch das vorgestellte Verfahren aufgrund seiner

hochwertigeren Ergebnisse deutlich vereinfachen. Dies reduziert die Produktionskosten für eine Ontologie und erlaubt eine breitere Einsetzbarkeit semantischer Technologien.

Literaturverzeichnis

Belkin, Nicolas J. (1980): Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science 5, S. 133–143

Berners-Lee, T. (1998): Semantic Web Roadmap. World Wide Web Consortiem, Technischer Bericht URL: http://www.w3.org/DesignIssues/Semantic

Catarci, Tiziana et al. (2003): An Ontology Based Visual Tool for Query Formulation Support. In Proceedings of OTM 2003, On The Move to Meaningful Internet Systems, Springer-Verlag Heidelberg, Oktober 2003, ISBN 3–540–20494–6

Farquhar, A., Fikes, R., Rice, J. (1996): The Ontolingua server: A tool for collaborative ontology construction. Stanford, KSL 96-26 – Technischer Bericht

Fleischmann, M., Strauss, W., Novak, J. (2001): netzspannung.org: an internet media lab for knowledge discovery in mixed realities. In: Proceedings of cast01: living in mixed realities, Fraunhofer IMK 2001, ISSN 1618–1379

Furnas, G.W. (1986): Generalized fisheye views. In: Proceedings of ACM CHI'86, ACM Press, S. 16–23

Guarino, N (1995): Formal ontology, conceptual analysis and knowledge representation.In: International Journal of Human-Computer Studies 43, Nr. 5-6, ISSN 1071–5819, S. 625–640

Gurzki, T., Hinderer, H., Eberhard, C.T., Bullinger, H.-J. (Hrsg.) (2002): Marktübersicht Portal Software - Marktübersicht für Business-, Enterprise-Portale und E-Collaboration. Stuttgart: Fraunhofer IRB, URL: http://www.media-vision.iao.fhg.de/downloads/Portal_Software.pdf

Hacker, Rupert (1992): Bibliothekarisches Grundwissen.München: K.G. Saur

Heyer, G. et al.: Learning Relations using Collocations. In: Proceedings of IJCAI Workshop on Ontology Learning.Seattle/ WA, USA, August

Jones, S. (1998): Graphical query specification and dynamic result previews for a digital library. In: Proceedings of the 11th annual ACM symposium on User interface software and technology, ACM Press, ISBN 1–58113–034–1, S. 143–151

Lamping, J., Rao, R. (1994): Laying Out and Visualizing Large Trees Using a Hyperbolic Space.In: Proceedings of the ACM Symposium on User Interface Software and Technology, ACM Press, S. 13–14

Maedche, A., Staab, S. (2000): Mining Ontologies from Text.In: Proceedings of 12th International Conference on Knowledge Engineering and Knowledge Management, EKAW-2000, Juan-les-Pins, France, Springer

Nie, J.-Y. (1992): Towards a probabilistic modal logic for semantic-based information retrieval.In: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, ISBN 0–89791–523–2, S. 140–151

Novak, J., Kunz, C., Wurst, M. (2003): Entdeckung und Nutzbarmachung von stillem Wissen in heterogenen Expertengemeinschaften.i-com Zeitschrift für interaktive und kooperative Medien, Nr. 3, ISSN 1618–162X, S. 18–26

Picot, Arnold/Reichwald, Ralf/Wigand, Rolf (2003): Die Grenzenlose Unternehmung. Dritte Auflage. Gabler Verlag

Sebastiani, F. (1996): A Note on Logic and Information Retrieval. In: Proceedings of MIRO-95, Workshop on Multimedia Information Retrieval, Glasgow, UK, S. 1–15

Sebrechts, M. et al. (1999): Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces.In: Research and Development in Information Retrieval. 1999, 3–10

Studer, R., Schnurr, H., Nierlich, A. (2001): Semantisches Knowledge Retrieval. Ontoprise, Karlsruhe, Ontoprise Whitepaper Series

Uschold, M., Grüninger, M. (1996): Ontologies: principles, methods, and applications. Knowledge Engineering Review 11, Nr. 2, S. 93–155

Van Rijsbergen, C.J. (1979): Information Retrieval, 2nd edition. Dept. of Computer Science, University of Glasgow, URL: http://citeseer.nj.nec.com/vanrijsbergen79information.html

Wolf, Patricia (2003): Erfolgsmessung der Einführung von Wissensmanagement. Eine Evaluationsstudie im Projekt 'Knowledge Management' der Mercedes-Benz Pkw-Entwicklung der DaimlerChrysler AG. Münster: Verlagshaus Monsenstein und Vannerdat, Managementkompetenz

Ziegler, J., Kunz, C., Botsch, V.: Matrix Browser: Visualisierung und Exploration vernetzter Informationsräume .In: Konferenzband der Mensch und Computer 02. Hamburg, Germany: Teubner Verlag