First publ. in: Information Engineering, Proceedings des 4. Konstanzer Informationswissenschaftlichen Kolloquiums (1999), pp. 171-182

Die Entwicklung eines Business Intelligence Systems zur Beschaffung von Geschäfts-informationen im WWW

Siegfried Handschuh, Thomas M. Mann, Gabriela Mußler, Harald Reiterer

FG Informatik und Informationswissenschaft
Universität Konstanz
{handschuh, mann, mussler, reiterer}@uni-konstanz.de

Kurzfassung

Nach einem funktionalen Überblick über das Gesamtsystem, wird auf die Konzeption eingegangen. Aufbauend auf dem Rahmenkonzept von Shneiderman wird ein Mulitagentensystem vorgeschlagen, das den Benutzer bei der Informationssuche unterstützen soll. Da der erfolgreiche Einsatz eines Business Intelligence System ganz wesentlich von der Art der Informationspräsentation bestimmt wird, beschäftigt sich ein wesentlicher Teil dieses Beitrags mit ausgewählten Ansätzen zur Visualisierung von Suchergebnissen, die innerhalb des Projektes verfolgt werden.

1 Motivation und Ziel

Im Rahmen des Beitrags wird das Business Intelligence System INSYDER (EU-ESPRIT, Projektnr. 29232) vorgestellt. Ziel dieses Systems ist es Klein- und Mittelunternehmen eine umfassende Unterstützung bei Suche von Geschäftsinformationen im WWW zu bieten.

Die vom INSYDER-System gebotene Funktionalität geht weit über die im Rahmen herkömmlicher Suchmaschinen vorhandene hinaus. Dies rechtfertigt es auch unserer Meinung nach von einem Business Intelligence System zu sprechen. Der Begriff Business Intelligence wurde 1989 von der Gartner Group geprägt und wie folgt definiert: „Business Intelligence is the process of transforming data into information and, through discovery into knowledge.“ [IBM 1999] griff diesen Begriff auf und definierte ein Business Intelligence System folgendermaßen: „A business intelligence system ... provides a set of

technologies and products for supplying users with the information they need to answer business questions, and make tactical and strategic business decisions."

INSYDER unterstützt nicht nur das Auffinden relevanter Information, sondern auch die Verwaltung der Rechercheergebnisse nach verschiedenen Interessensgebieten des Benutzers und bietet damit eine wesentliche Voraussetzung zur Strukturierung und Verteilung des Wissens im gesamten Unternehmen.

Nach einem Überblick über das Gesamtsystem, wird auf die Konzeption der Architektur sowie auf die Methoden zur Unterstützung der Informationssuche detailliert eingegangen. Der Beitrag schließt mit einer Zusammenfassung der wesentlichen Ergebnisse und einem Ausblick zukünftiger Projektaktivitäten.

2 Charakteristika des Business Intelligence Systems INSYDER

INSYDER unterstützt die Prozesse des Sammelns und Analysierens unstrukturierter Daten in Dokumentenform. INSYDER ist in der Lage, in diesen verteilten elektronischen Dokumenten zu suchen, sie zu sammeln und nach benutzerdefinierten Kriterien zu analysieren und zu klassifizieren.

Für die Analyse bedient sich INSYDER eines Thesaurus. Er ermöglicht die inhaltliche semantische Analyse der Dokumente. Das INSYDER-System kann somit Dokumente auch dann finden, wenn diese die Suchbegriffe selbst nicht enthalten, dafür aber ähnliche Konzepte (z.B. Synonyme). Der Thesaurus existiert derzeit in zwei Sprachen, Französisch und Englisch. Dies erlaubt eine mehrsprachige Auswertung der Dokumente: das System kann folglich mit einer englischen Suchanfrage auch französischsprachige Dokumente auffinden und bewerten. Neben Thesauri für verschiedene Sprachen sind auch fachspezifische Thesauri vorgesehen (z.B. für CAD, Pharmazie). Dies ermöglicht die Anpassung von INSYDER an unterschiedliche Unternehmensbedürfnisse.

INSYDER bietet einige Funktionen zur Unterstützung der Informationsbeschaffung, dazu gehören die Sphere-Of-Interest, sowie Search, Watch und News.

Sphere-Of-Interest

Die Sphere-Of-Interest (SOI) ist eine Darstellung und Verwaltung der Interessensbereiche des Benutzers (Abb.1). Der Benutzer kann verschiedene Interessensbereiche definieren, z.B. Technologie, Marketing, Konkurenzbeobachtung. Jeder Interessensbereich wird als Ordner dargestellt. Innerhalb dieser Ordner können verschiedene Suchen (Search) und Beobachtungen (Watch) definiert werden, die dem jeweiligen Interessensbereich zugeordnet sind. Im den Interessensbereichen werden die bisherigen Suchvorgänge sowie die aktuellen Sucheanfragen und deren Zustände dargestellt, d.h. es wird angezeigt, ob die Suche/Beobachtung gerade im Hintergrund durchgeführt wird (Pfeil-Symbol), oder ob sie schon beendet ist (Stop-Symbol). Zudem ist es möglich, vordefinierte SOIs für den Benutzer zu erstellen und INSYDER damit auszuliefern. Weiterhin ist geplant, daß der Benutzer SOIs abonnieren kann, so daß sein INSYDER-System diese SOIs automatisch updated, sobald der Anbieter sie aktualisiert. Die SOIs stellen somit neben den fachspezifischen Thesauri eine weitere Personalisierungsmöglichkeit von INSYDER dar.

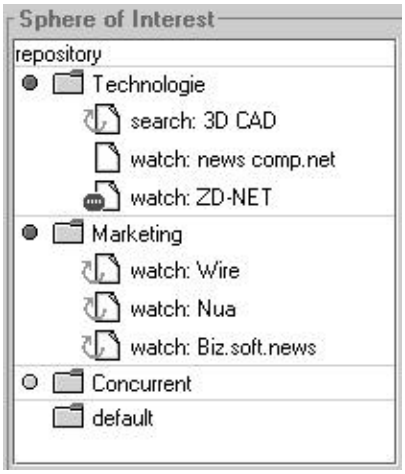

Abb. 1: Sphere-Of-Interest

Search

Der Benutzer formuliert in natürliche Sprache seine Anfrage, wählt die Quellen der Suche aus und definiert gegebenenfalls Parameter wie die Suchdauer oder Suchtiefe. Die Quellen sind in Kategorien geordnet. Diese Kategorien beinhalten vorhandene Suchmaschinen (Altavista, Excite, usw.), Newsgroups, Bookmark-Listen oder Verzeichnisse auf der lokalen Festplatte. Einmal angestoßen, läuft die Suche dann so lange, bis ein benutzerdefiniertes Abbruchkriterum erfüllt ist oder sie manuell gestoppt wird. Standardmäßig werden Suchresultate als eine Liste mit verschiedenen Attributen wie Titel, URL und Relevanz angezeigt (Abb. 2). Andere Darstellungen der Suchergebnisse werden im Abschnitt Visualisierung vorgestellt.

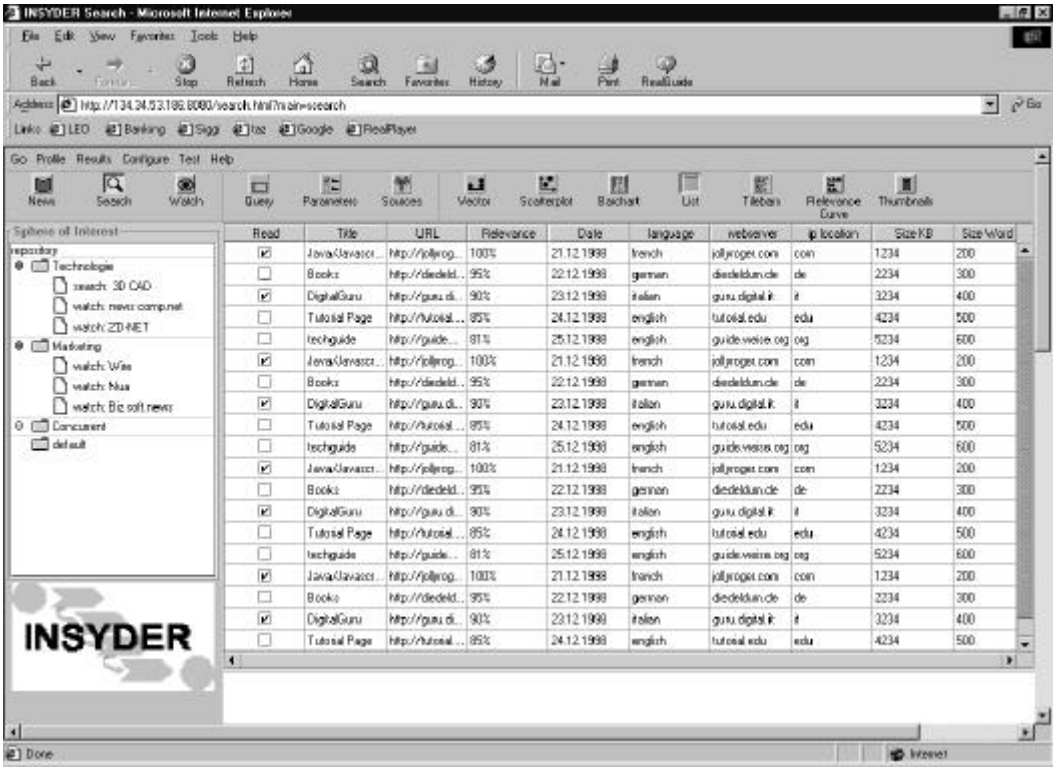

Abb. 2: Trefferliste

Watch
Die Watch-Funktion ermöglicht das Überwachen von Dokumenten, dabei kann eine Veränderung des Dokuments oder das Auftauchen eines benutzerdefinierten Begriffes bemerkt werden. Der Anwender kann damit einen Markt oder Technologien beobachten, um Trends aufzuspüren oder strategische Bewegungen zu entdecken.

News
Die News-Funktion beinhaltet die Idee eines Webportals. Die Informationen werden dem Benutzer auf einer Webseite präsentiert; die Seite besteht aus einer Sammlung von vorgefertigten Links zu nationalen und internationalen täglichen Nachrichten. Als Quelle dient das Internet mit seinen Zeitungsartikeln, Magazinen und Presseagenturen. Des weiteren sieht der Benutzer auf einen Blick die Resultate der noch laufenden Suchen (Search) und Beobachtungen (Watch).

Nach dem funktionalen Überblick über das Gesamtsystem wird im folgenden auf die Konzepte des Multiagentensystems und der alternativen Visualisierungen für die Treffermengen eingegangen.

3 Agenten und Visualisierung der Suchergebnisse

Zunächst wird ein allgemeingültiges Rahmenkonzept der Informationssuche vorgestellt, das auf Arbeiten von Shneiderman [Shneiderman 1998] basiert. Dieses Rahmenkonzept vermittelt einen guten Eindruck hinsichtlich der wesentlichen Aktivitäten bei der Informationssuche und eignet sich sehr gut für die Zuordnung der von uns entwickelten Konzepte der Agenten und Visualisierungen. Es handelt sich dabei um die Phasen:

- *Formulation (Formulierung der Anfrage)*
- *Initiation of action (Initiieren der Suchaktion)*
- *Review of results (Resultatsbetrachtung)*
- *Refinement (Verfeinerung/Bearbeitung der Suchanfrage)*

Diese genannten Phasen bilden sowohl die Grundlage für die Konzeption der Agenten als auch der Visualisierung der Suchergebnisse.

Anhand einer Literaturanalyse konnten eine Reihe von Defiziten bei der Informationssuche ausgemacht werden, vgl. [Mußler 1999]. Die Verbesserung der dort genannten Kritikpunkte an bestehenden Systemen war ein Ziel bei der Konzeption der Agenten und der Visualisierung des INSYDER Systems. Zunächst wird kurz auf die Konzeption der Agenten eingegangen, um anschließend die Visualisierungsansätze innerhalb INSYDER zu erläutern.

3.1 Agenten

Ein Kritikpunkt in der Literatur ist, daß es den Benutzern von Suchsystemen meist nicht möglich ist, ihren Informationsbedarf dem System mitzuteilen. Innerhalb des Projektes INSYDER wird dazu folgender Ansatz verfolgt: Ein **Formulationagent** unterstützt den Benutzer bei der Formulierung seiner Suchanfrage mit Hilfe eines semantischen Netzes, welches auf einem Thesaurus (Knowledge Base) des Systems basiert.

Die Phase der Aktion wird heute bereits durch eine Reihe von bekannten Agenten unterstützt. In der Literatur finden sie sich unter der Bezeichnung Robot, Crawler, Wanderer u.ä. Bezeichnungen wieder. Gemeinsam ist ihnen die Aufgabe, Dokumente aufzuspüren, diese zu indexieren und an die Datenbasis weiterzugeben [Turau 1998]. INSYDER verfolgt einen weitergehenden Ansatz: die Ergebnisse der einzelnen Suchmaschinen bilden nur die Ausgangsbasis für eine weitere Suche nach relevanten Dokumenten, indem die einzelnen Links weiterverfolgt werden. Hierzu ist es notwendig im System

mehrere **Crawlingagenten** zu haben, die diese Aufgabe lösen. Dabei unterteilt sich deren Aufgaben in mehrere Unteraufgaben, bspw. die Bewertung einzelner Seiten und deren Links, mit dem Ziel herauszufinden, ob es sinnvoll ist, diese weiterzuverfolgen.

Diese Phase der Resultatsbetrachtung ist nur bedingt durch den Einsatz von Agenten zu unterstützen. Es wird stets dem Benutzer obliegen, die endgültige Entscheidung zu fällen, ob er ein Dokument, das ihm vom Suchsystem zur Befriedigung seines Informationsbedarfes vorgeschlagen wird, annimmt oder es verwirft. Diese Phase ist aber auch diejenige, die der Benutzer direkt mit der Qualität des Systems verbindet. Eine mögliche Art der Unterstützung in dieser Phase besteht darin, die Ergebnisse dem Benutzer geclustert zu präsentieren. Hier wird daher ein **Clusteringagent** vorgeschlagen, dessen Aufgabe darin besteht, eine gegebene Dokumentenmenge daraufhin zu untersuchen, ob bestimmte Merkmale in mehreren Dokumenten gehäuft vorkommen. Als ein Beispiel für einen Agenten, der diese Aufgabe erfüllt, sei hier auf [Pazzani, et al. 1996] oder [Zamir, et al. 1998] verwiesen.

Die Refinement Phase wird in INSYDER durch ein Relevance Feedback ausgedrückt. Es zeichnet sich dadurch aus, daß eine erneute Suche automatisch vom System erzeugt wird, indem aus der alten Suche und anhand von relevanten Dokumenten eine neue Anfrage berechnet wird [Salton, et al. 1987]. Ein Problem, daß bei der Anwendung dieses Verfahrens besteht, ist, daß der Benutzer explizit sein Feedback zu verschiedenen Dokumenten geben muß: “Ideas like query reformulation, relevance feedback, and query-by-example are all important but do not come naturally to users.” [Nielsen 1997]. Hier kann ein Agent zwar die Aufgabe übernehmen, den Benutzer zu beobachten und daraus zu schließen, welche Dokumente er für wichtig erachtet, letztendlich ist es aber immer die Entscheidung des Benutzers zu sagen, ob ein Dokument relevant ist oder nicht. Im Rahmen des INSYDER Systems wird der **Formulationagent** auch in der Phase des Refinement genutzt.

3.2 Visualisierung von Suchergebnissen

Eine Analyse von Literatur und bestehenden Systemen hinsichtlich möglicher Visualisierungen von Suchergebnissen führte zur Erkenntnis, das hier bereits eine große Menge von Ideen und Konzepten gibt, die verwendet werden könnten. Auch finden sich eine Reihe von Ratschlägen sowie einige Erkenntnisse aus tatsächlichen Umsetzungen der Ideen, die durch Experimente und Untersuchungen abgesichert sind. Die Sichtung der Literatur ergibt eine wichtige Aussage, die sich generell quer durch alle Ansätze zieht: Die optimale Visualisierung für jeden Anwendungsfall gibt es nicht [Mann 1999].

Insgesamt existieren vier Hauptfaktoren, die den Nutzen einer Visualisierung im konkreten Anwendungsfall beeinflussen – nachfolgend als "4T-Umgebung" bezeichnet:

- *Target user group (Zielgruppe),*
- *Type and number of data (Datentyp und -menge),*
- *Task to be done (aktuelle Aufgabe) und*
- *Technical possibilities (technische Möglichkeiten).*

Zielgruppe bedeutet nicht nur die Frage, ob ein Wissenschaftler oder ein Sachbearbeiter vor dem Monitor sitzt. Es bestehen auch interpersonelle Unterschiede in der Informationsaufnahme und –verarbeitung.

Der darzustellende Datentyp und auch die Datenmenge beeinflußt in jedem Fall die Auswahl der graphischen Repräsentation. Wenn z.B. eine Hierarchie innerhalb der Daten besteht, ist es sinnvoll, diesen Umstand für die Visualisierung zu nutzen. 50 Dokumente als Tilebars darzustellen, kann die Ermittlung der relevantesten durchaus unterstützen. Bei 5000 Dokumenten hingegen wird der Benutzer vermutlich zuerst einen Verfeinerungsschritt mit einer anderen Visualisierungsform bevorzugen.

Ein ebenfalls bedeutender Faktor für die Effektivität einer Visualisierung ist die aktuelle Aufgabe des Benutzers. Es gibt eine ganze Reihe von Klassifizierungen oder Einstufungen von Visualisierungen für verschiedene Aufgabenformen, wobei "Aufgaben" hier sehr breit definiert wird. Ein Ansatz auf relativen hohem logischen Aufgabenniveau, ist das bereits erwähnte Vier-Phasen-Modell der Informationssuche nach Shneiderman.

Nicht zuletzt stellen die technischen Gegebenheiten des jeweiligen Szenarios einen entscheidenden Faktor für Einsatz und Erfolg eines Visualisierungskonzeptes dar. Beispiel für einen hier bestimmenden Faktor ist der Einsatz einer Browser-basierten Benutzungsschnittstelle.

Ein Ansatz um für eine gegebene 4T-Umgebung die optimale Lösung zu finden, ist der Einsatz alternativer oder kombinierter Ansätze. So findet sich in [Ahlberg, et al. 1995] die Idee der Kombination von mehreren Visualisierungen wie Karten oder Starfields mit ebenso mehrfachen Anfragemethoden. [Henninger, et al. 1996] schlagen für Retrievalsysteme vor, verschiedene Interaktionsstile zu unterstützen, etwa Browsing oder direkte Suche, mit schnellem Zugriff auf zusätzliche Informationen, und den Einsatz von Feedback-Techniken. [Stenmark 1998] beschreibt einen Ansatz zur Kombination einer visualisierten Cluster-Technik mit Relevanz-Feedback. In [Mann 1999]

wird der innerhalb von INSYDER eingesetzte Ansatz der alternativen synchronisierten Visualisierungen vorgestellt. Nachfolgend werden die sechs unterschiedlichen Visualisierungsansätze beschrieben, die im System getestet werden.

Der **Document-Vector** (Dokumentvektor**)** zielt darauf ab, dem Benutzer einen leicht verständlichen Überblick über eine größere Anzahl vom System gemachter Vorschläge zu geben. Er ist eindimensional ausgelegt. Jedes Dokument wird durch einen schwarzen Punkt dargestellt (die hellen Punkte rühren von bestimmten Benutzeraktionen her und werden weiter unten erklärt). Wenn sich mehr als ein Punkt in einer Spalte der Skala befindet, dann wird das Dokument in einer zweiten Zeile angezeigt und so weiter. Welche Variable angezeigt wird, kann der Benutzer aus einer Liste auswählen. Beispiele sind die Relevanz oder das Datum eines Dokuments.

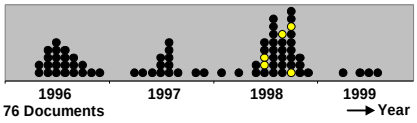

Abb. 3: Document-Vector Jahr (Year)

Im **Scatterplot** (XY-Diagramm**)** können zwei Variablen gleichzeitig angezeigt werden. Auch hier hat der Benutzer selbst die Wahlmöglichkeit, welche Kategorien angezeigt werden. Beispiele sind die Gesamtrelevanz und Jahr, oder die Relevanz von Suchbegriff A und Suchbegriff B, oder andere, die jeweils auf einer der beiden verschiedenen Achsen angezeigt werden.

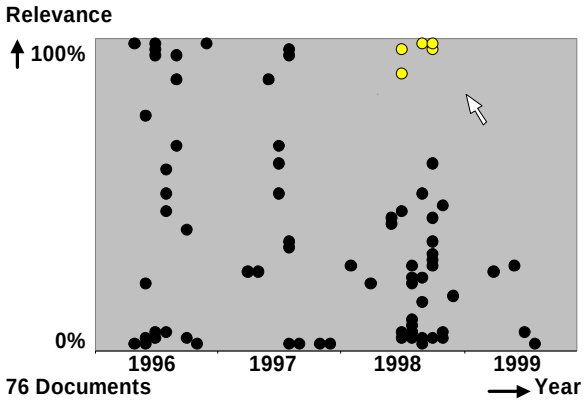

Abb. 4: Scatterplot Relevanz (Relevance) und Jahr (Year)

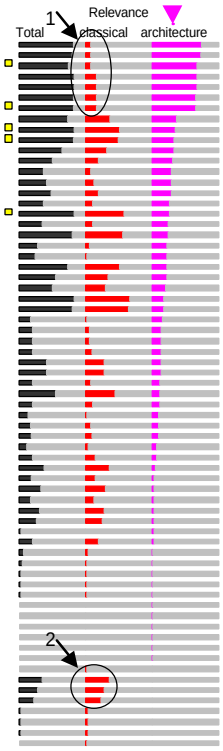

Abb 5. Bargraph

Der **Bargraph** (Balkendiagramm**)** zeigt für jeden Suchbegriff die Gesamt- und die Relevanz im einzelnen an (in Abb. 4 “klassische” und “Architektur”). Die ursprüngliche Idee aus [Veerasamy, et al. 1995] wird hier in mehrere Aspekten modifiziert. Erstens ist die Darstellung um 90° gedreht, um die Dokumente auf dieselbe Art anzuzeigen wie in den übrigen Sichten, bei denen Details der Dokumente angegeben werden: Von oben nach unten statt von rechts nach links. Zweitens wird der Eindruck eines Dokuments als Gesamtheit durch die Anwendung der Gestaltgesetze verstärkt, ohne die generelle Ausrichtung auf die Suchbegriffe allzusehr zu stören.

Tilebars (Kachelreihe) ein Beispiel für Visualisierungen, die mehr Details anzeigen als die gewöhnliche Liste [Hearst 1995]. Abb. 6 zeigt drei Dokumente, unterschiedlicher Länge. Jede Tilebar steht für ein Schlagwort, jede vertikale Gruppe von Tiles (Kacheln) repräsentiert einen Teil des Dokuments. Je dunkler eine Tile ist, umso relevanter ist das Schlagwort für diesen Teil des Dokuments. Im angegebenen Beispiel besteht die Suchanfrage aus zwei Begriffen: “klassische” in der ersten Zeile und “Architektur” in der zweiten. Sieht man sich das dritte Dokument an, so ist deutlich zu sehen, dass es in keinem Teil dieses Dokuments um "Architektur" UND "klassische" geht. Am Anfang des Dokuments wird der Begriff “Architektur” behandelt, der Begriff “klassische” am Ende. Betrachtet man den Dokumententitel “Klassische Archäologie: Griechische Architektur”, ergibt diese Interpretation einen Sinn. Damit der Benutzer die Möglichkeit hat, seine Eindrücke zu bestätigen, bieten die Tilebars die Möglichkeit, durch Klicken auf die entsprechende Tile direkt zum zugehörigen Teil des Dokuments zu springen.

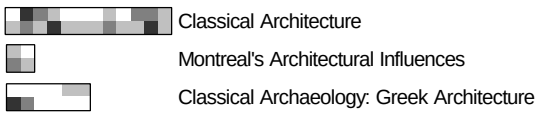

Abb. 6: Tilebars und Dokumententitel

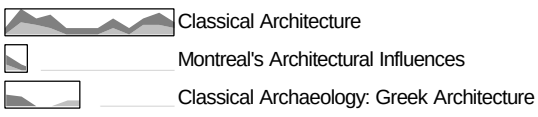

Abb.7: Relevance-Curves und Titel

Relevance-Curves (Relevanzkurven**)** haben dasselbe Ziel wie Tilebars, allerdings mit einer anderen Art der Umsetzung in visuelle Strukturen. Dabei wird die ursprüngliche Idee aus [Arisem S.A. 1999] durch zwei Änderungen erweitert. Die Länge der Kurve repräsentiert die Länge des Dokuments (ursprünglich ist die Kurvenlänge unabhängig von der Dokumentlänge) und es

werden Farbcodes verwendet , um den Einfluß der unterschiedlichen Suchbegriffe anzuzeigen.

Die Hauptidee der Darstellung als **Thumbnail** (Miniaturansicht**)** ist es Benutzern, die oft in den selben Dokumentenräumen suchen, Hinweise auf möglicherweise bereits bekannte Dokumente zu geben. Dabei kann es auch eine Hilfestellung für den ersten Eindruck bei unbekannten Dokumenten geben. Erkanntes Problem der Thumbnails sind die Suchzeiten. Bei allen anderen Visualisierungen und Systemmechanismen genügt es, die HTML-Text-Dateien zu laden. Zur Generierung der Thumbnails müssen jedoch auch alle Bilder in den Dokumenten geladen werden. Der Einsatz von Thumbnails im Internet birgt somit das Problem der Suchzeit. Bei Intranets dagegen ist dies nicht der Fall. Dort allerdings dürften aufgrund von Corporate-Identity-Vorgaben die Dokumente ein einheitliches Erscheinungsbild haben, so daß der Wert von Thumbnails eher gering ist.

Abb. 8: Thumbnail eines Dokuments

4 Resümee und Ausblick

Die Ideen, die in diesem Beitrag diskutiert wurden, beziehen sich auf die Entwicklung eines Business Intelligence Systems, das dem Benutzer eine umfassende Unterstützung bei der Suche nach Geschäftsinformationen im WWW bietet.

Das INSYDER System ist von seiner Konzeption her abgeschlossen und wird gerade implementiert. Eine erste projektinternen Evaluation ist für September vorgesehen. Dabei soll neben der technologischen Machbarkeit auch die Benutzbarkeit des Systems untersucht werden. In der weiteren Folge des Projektes wird dann eine umfassende Pilotstudie mit 30 KMUs aus Frankreich, England und Italien durchgeführt, um die Ergebnisse des Projektes auch unter realen Einsatzbedingungen zu evaluieren.

5 Literatur

[Ahlberg, et al. 1995] C. Ahlberg, E. Wistrand, “IVEE: An information visualization and exploration environment,” In: Proc. IEEE Information Visualization, 1995.

[Arisem S.A. 1999] Arisem S.A. http://www.arisem.com [1999-06-20]

[Hearst 1995] M. A. Hearst, “TileBars: Visualization of Term Distribution Information in Full Text Information Access,” In: Proc. CHI 1995, 1995.

[Henninger, et al. 1996] S. Henninger, N. J. Belkin, “Interface Issues and Interaction Strategies for Information Retrieval Systems,” In: Proc. CHI'96, 1996.

[IBM 1999] IBM. http://www.software.ibm.com/data/pubs/papers/bisolution/index.html

[Mußler 1999] G. Mußler, “Ein Agentensystem zur Unterstützung bei der Informationssuche im WWW,” Im Erscheinen in: Proc. ADI'99 (Agenten, Datenbanken und Information Retrieval), 1999. http://wwwdb.informatik.uni-rostock.de/adi99/

[Nielsen 1997] J. Nielsen, “Search and You May Find". http://www.useit.com/alertbox/9707b.html

[Pazzani, et al. 1996] M. Pazzani, J. Muramatsu, D. Billsus, “Syskill & Webert: Identifying interesting web sites,” In: Proc. National Conference on Artificial Intelligence., 1996. http://www.ics.uci.edu/~pazzani/RTF/AAAI.htm

[Salton, et al. 1987] G. Salton, M. J. McGill, *Information Retrieval: Grundlegendes für Informationswissenschaftler*. Hamburg: McGraw-Hill, 1987.

[Shneiderman 1998] B. Shneiderman, *Designing the user interface: strategies for effective human-computer-interaction*. Reading, Harlow, Menlo Park u.a.: Addison Wesley, 1998.

[Stenmark 1998] D. Stenmark, “To Search is Great, to Find is Greater: a Study of Visualization Tools for the Web". http://w3.informatik.gu.se/~dixi/publ/mdi.htm [1999-01-11]

[Turau 1998] V. Turau, “Web-Roboter,” *Informatik Spektrum*, vol. 21 (3), pp. 159-160, 1998.

[Veerasamy, et al. 1995] A. Veerasamy, S. B. Navathe, “Querying, Navigating and Visualizing a Digital Library Catalog,” In: Proc.

Digital Libraries '95: The Second Annual Conference on the Theory and Practice of Digital Libraries, 1995. http://www.csdl.tamu.edu/DL95/papers/veerasamy/veerasamy.html

[Zamir, et al. 1998] O. Zamir, O. Etzioni, "Web Document Clustering: A Feasibility Demonstration," In: Proc. SIGIR 1998, 1998. http://zhadum.cs.washington.edu/zamir/sigir98.ps