namics
a PubliGroupe Company

Enterprise Information Retrieval

Moderne Suchtechnologien und deren Nutzen für Unternehmen

namics Whitepaper
17. Februar 2004

namics ag
Teufenerstrasse 19
CH–9000 St.Gallen

t [+41] 71 228 67 77
f [+41] 71 228 67 88
info@namics.com

Offices in:
Frankfurt, Hamburg,
Bern, St. Gallen, Zug, Zürich

Übersicht

1	Einführung	5
2	Exkurs: „Klassisches“ Information Retrieval	11
3	Enterprise Information Retrieval Systeme (EIRS)	19
4	Marktüberblick: Anbieter von IR-Systemen	32
5	Leistungsangebot	45
6	Ressourcen	48

Inhaltsverzeichnis

1	Einführung	5
1.1	Welchen Nutzen bringt Information Retrieval für Unternehmen?	5
1.2	Begriffsdefinitionen	7
1.2.1	Information Retrieval	7
1.2.2	Content	8
1.2.3	Information	9
1.2.4	Informationsarbeit	9
1.2.5	Information Worker	10
1.2.6	Informationeller Mehrwert	10
2	Exkurs: „Klassisches“ Information Retrieval	11
2.1	Grundbegriffe	11
2.2	Fachinformationsmarkt und Online-Datenbanken	12
2.3	Information Retrieval Modelle	14
2.3.1	Modelle mit exakter Übereinstimmung	15
2.3.2	Modelle mit bestmöglicher Übereinstimmung	15
2.3.3	Übersicht über die IR-Modelle	16
2.4	Wissensrepräsentation und Anfragesprache	17
3	Enterprise Information Retrieval Systeme (EIRS)	19
3.1	Abgrenzung zum klassischen Information Retrieval	20
3.1.1	Besonderheiten des Internet	21
3.1.2	Besonderheiten des Intranet	23
3.2	„Big Picture“ eines Enterprise Information Retrieval Systems	24
3.2.1	Architektur	24
3.2.2	Laufzeitverhalten	26
3.3	Relevante Aspekte von Enterprise IRS	27
3.3.1	Einsatz von Suchmaschinen	27
3.3.2	Metasuchmaschinen	27
3.3.3	Einsatz von Katalogen	28
3.3.4	Informationsassistenten	28
3.3.5	Suchmaschinenmarketing	29
3.3.6	Einsatz und Bedeutung von Metadaten	30
4	Marktüberblick: Anbieter von IR-Systemen	32
4.1	Search Applikationen	33

4.1.1 Europsider – relevancy 6.0 33
4.1.2 Autonomy – Content Infrastructure 35
4.1.3 Vivisimo – Clustering Engine 36
4.1.4 Verity – UltraSeek 37
4.2 Kompakte Searchprodukte 38
4.2.1 Google – Google Search Appliance 38
4.2.2 Atomz – Atomz Search 39
4.2.3 e-serve – e-serve®Business 40
4.3 Integrierte Searchprodukte 41
4.3.1 Microsoft – Sharepoint Portal Server 41
4.3.2 IBM – WebFountain 42
4.4 OpenSource Systeme 43
4.5 Auswahl eines EIR-Systems 44

5 Leistungsangebot 45
5.1 IR-Assessment 45
5.2 Produkte-Evaluation 46
5.3 Beratung und Implementierung 47
5.4 Kontakt und weiterführende Informationen 47

6 Ressourcen 48

1 Einführung

1.1 Welchen Nutzen bringt Information Retrieval für Unternehmen?

„Die richtigen Informationen zum richtigen Zeitpunkt am richtigen Ort“ – dies ist heute sowohl für Unternehmen als auch für deren Mitarbeiter zunehmend der **zentrale Erfolgsfaktor** in einer globalisierten (und damit virtualisierten) Informationswirtschaft und -gesellschaft.

Geschäftsprozesse werden heute zunehmend im Sinne von **Business Information Processes** (BIP) interpretiert, Mitarbeiter werden zu **„Information Workern“**. **Information Retrieval Systeme** (IRS) sind die Schnittstelle zwischen Information Workern und BIPs und haben sich (spätestens seit dem Markteintritt von Microsoft[1]) zum **strategischen Thema** der Unternehmens-IT entwickelt.

So verwundert es nicht, dass Mitarbeiter einen Grossteil der Arbeitszeit (und damit der Kosten) heute mit der Suche nach wichtigen Informationen verbringen. Eine Reihe von Studien[2] zeigen, dass Mitarbeiter (je nach Tätigkeit) bis zu **50% Ihrer Arbeitszeit mit der Suche nach Informationen** verbringen, und nur etwa 50% mit der tatsächlichen und nutzbringenden Anwendung dieser Informationen! Dies zeigt das **enorme Potential** welches der Einsatz von Information Retrieval Technologie in Unternehmen birgt.

Diese Erkenntnis spiegelt auch das grundlegende Dilemma wider, welches die **Bedeutung von Information Retrieval für den Geschäftserfolg von Unternehmen** charakterisiert: Das Problem mit der Information ist nicht, dass es Sie im Unternehmen nicht geben würde. Das Problem ist vielmehr, dass durch die **Verwaltung von Content** selbst kein Nutzen generiert werden kann, sondern erst durch die **Anwendung von Content** auf der Grundlage einer konkreten Problemstellung.

---

[1] Information Worker Architecture und Office 2003 als erste Produktsuite auf Basis dieser Architektur

[2] Delphi research et al.

Weiterhin charakteristisch ist in vielen Unternehmen der Effekt der so genannten **„Informations-Inseln[3]“**:

» Über 80% der digitalen Informationen eines Unternehmens befinden sich auf einzelnen Festplatten und in persönlichen Dateien.
» In einem Unternehmen erhalten die Mitarbeiter 50-75% der relevanten Informationen direkt von anderen Personen.
» Einzelpersonen machen das Wissen wirtschaftlich nutzbar; der grösste Teil davon geht beim Austritt aus dem Unternehmen verloren.

IR-Systeme virtualisieren solche Strukturen und machen sie damit kontrolliert nutzbar.

Information Retrieval Systeme sind also Grundlage für **effiziente und effektive Teamarbeit, höhere Mitarbeiterproduktivität** durch einfacheren, schnelleren Zugriff auf Informationen sowie den systematischen **Abbau der gegenwärtigen Informations-„Inseln“** in Unternehmen.

Der gezielte Einsatz von **IR-Technologie birgt enorme Potentiale** für **Kostenreduktion** und **Produktivitätssteigerung** und stellt damit eine exzellente Chance für Unternehmen dar, um sich im Wettbewerb besser zu positionieren.

[3] Siehe auch „The Knowledge Investment Paradox“, Gartner Research 17.7.2002

1.2 Begriffsdefinitionen

In diesem Abschnitt werden nochmals die wichtigsten Begriffe definiert, welche im Dokument durchgängig verwendet werden und für das Verständnis der Argumentationen zentral sind.

1.2.1 Information Retrieval

Der Begriff besteht aus den Wörtern **Information** und **Retrieval.** Retrieval heisst wörtlich wiederfinden oder wiederbekommen, es ist unmittelbar klar was damit gemeint ist. Interessanter ist jedoch der **Begriff der Information**, schliesslich bekommt man als Ergebnis jeder Suchanfrage ja bestimmte Inhalte zurück, aber ist dies gleichzeitig auch immer Information? – wohl kaum.

Im Grunde genommen ist es ganz einfach: Entscheidend für den Wert ist die Handlungsrelevanz der zurückgelieferten Inhalte. Das bedeutet konkret, dass nur die Inhalte welche ein Anwender in Bezug auf eine konkrete Anfrage verwenden kann (welche Ihn also weiterbringen) als Information bezeichnet werden können. Das heisst auch dass ein Inhalt, der grundsätzlich zu einer bestimmten Anfrage passt, den ich aber schon kenne, für mich keine Information mehr darstellt. Etwas, was ich weiss, ist für mich keine Information mehr.

Diese Erkenntnis ist zentral und gleichzeitig auch trivial: Nur Inhalt, der in einem konkreten Zusammenhang angewendet werden kann, hat einen Wert, und nur solche Inhalte sind gleichzeitig auch Informationen.

Dies nennt man auch den „pragmatischen Mehrwert von Information[4]“.

Information Retrieval umfasst somit als Gattungsbegriff alle Technologien, Methoden und Konzepte welche sich mit der Aufbereitung von Content

[4] Streng genommen hat das Konzept des pragmatischen Mehrwerts mehrere Dimensionen, so reicht es z.B. nicht aus, die „richtigen“ Informationen zu bekommen, sondern diese müssen darüberhinaus so aufbereitet sein, dass sie auch (didaktisch wie auch technisch) verarbeitet werden können. Der Kern des Mehrwertgedankens basiert jedoch immer auf dem konkreten Nutzen, und soll deshalb in diesem Dokument vereinfachend so dargestellt werden.

(zum Zwecke der späteren Suche) und der Extraktion einer bestimmten Teilmenge dieses Wissens auf Basis einer formalen Anfrage befassen.

Enterprise Information Retrieval fokussiert speziell auf den Teilbereich des Information Retrieval welcher für den Einsatz in Unternehmen relevant ist.

1.2.2 Content

Content ist die Gesamtheit aller elektronisch verfügbaren Daten. Da Content in der Regel „einen Sinn ergibt“ bzw. mit einer bestimmten Intention erstellt wurde ist er eine (beliebige) Form der Wissensrepräsentation. In den meisten Fällen kann Content mit dem (elektronisch verfügbaren) Wissen eines Unternehmens gleichgesetzt werden.

Aufgrund dieser Diskussion kann man also den Begriff „Content“ (bzw. Wissen) in Beziehung setzen zu den Begriffen der Information und der Datenhaltung. Diese „Aggregatzustände“ unterscheiden in Bezug auf den Wert, den derselbe Inhalt erzeugt, je nachdem auf welche Ebene er materialisiert werden kann.

Dies ist auch ein wesentlicher Aspekt der Abgrenzung von IR-Systemen zu herkömmlichen Suchmaschinen. Während die letzteren hauptsächlich auf der syntaktischen Ebene nach Inhalten suchen (Vergleich von Wörtern oder Zeichenketten) bewegen sich IR-Systeme bevorzugt auf der semantischen oder pragmatischen Ebene.

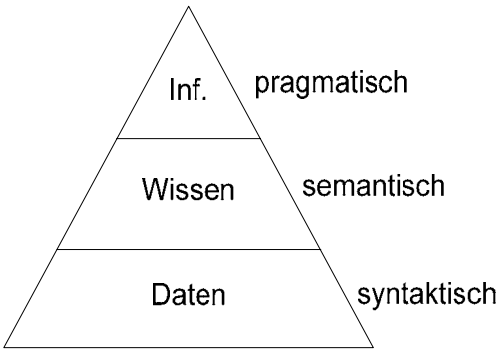

Abbildung 1: Daten - Wissen - Information

1.2.3 Information

Content wird nur unter ganz bestimmten und vor allem kontrollierten Bedingungen zu Information. Zentral ist hier der Aspekt des Neuigkeitswerts und des Nutzens: Nur etwas was man noch nicht weiss und was man gleichzeitig gebrauchen kann ist faktisch Information. D.h. Information hat im Gegensatz zu Wissen einen konkreten Nutzen und damit auch einen (quantifizierbaren) Wert. Dieser Aspekt macht das Konzept der Information zum zentralen Aspekt für den Einsatz von Informationstechnologie in Unternehmen.

Interessanterweise kann man so auch begründen warum der Einsatz effizienter Content Management Systeme allein keinen (oder kaum) operativ messbaren Nutzen bringt[5] (eine Erfahrung, welche wohl schon viele Unternehmen machen mussten) – erst die Kombination von Content Management und Information Retrieval bringt wirklich Nutzen und damit auch geldwerte Vorteile.

1.2.4 Informationsarbeit

Informationen entstehen nicht von selbst. Informationen sind wertvoll. Wie für jeden wertschöpfenden Prozess braucht es auch für die Entstehung von Information systematische Arbeit um diesen Wert zu schaffen. Die Prozesse der Informationsarbeit lassen sich dabei in zwei Hauptkategorien unterteilen. Zum einen die Prozesse, die dazu dienen Inhalte zu erstellen bzw. zu veredeln (also z.B. Klassifikation von Inhalten, Abstracing etc.). Zum anderen die Prozesse, die dazu dienen aus diesen Inhalten Informationen zu gewinnen (also z.B. Formulierung von Suchanfragen, Auswertung der Ergebnisse etc.).

Die Prozesse der Informationsarbeit werden in vielen unternehmensweiten Prozessmodellen weitgehend ignoriert, ein weitreichender Fehler, denn gerade diese Prozesse bergen in der Regel das grösste Optimierungspotential!

[5] In der Regel basieren ROI Betrachtungen für CMS deshalb auch auf der Reduktion des administrativen Aufwands und nicht auf der Steigerung der Produktivität.

1.2.5 Information Worker

Ein Information Worker ist ein Mitarbeiter, der für die Erledigung seiner täglichen Arbeit auf Informationstechnologie angewiesen ist (Wer ist das heutzutage eigentlich nicht?). Dies betrifft sowohl die fachlichen Aspekte (welche Inhalte braucht welcher Mitarbeiter), vor allem aber die pragmatischen (wie kann er effizient mit diesen Inhalten arbeiten).

1.2.6 Informationeller Mehrwert

Der informationelle Mehrwert ist ein sehr vielschichtiges Konzept. Grundsätzlich umschreibt dieser Begriff die Tatsache, dass dieselbe Information (eigentlich: derselbe Content) einen unterschiedlichen Wert haben kann, je nachdem ob er für einen Information Worker in einem konkreten Zusammenhang nützlich ist, oder eben nicht.

Neben diesem Aspekt der Handlungsrelevanz von Information sind natürlich auch die Aspekte der System-Technologie von Bedeutung. So zeigen beispielsweise neuere Studien, dass durch den Einsatz von Tablet PCs in Unternehmen die Produktivität der Information Worker um bis zu 35% gesteigert werden konnte – allein durch den einfacheren und ortsunabhängigen Zugriff auf Informationen. Dies ist ebenfalls ein (systemischer) informationeller Mehrwert.

Weiterhin sind hier auch die Aspekte der (didaktischen bzw. kognitiven) Informationsvermittlung von Bedeutung. Informationen, die an sich relevant sind, können trotzdem nicht nützlich sein, wenn ein Information Worker sie nicht verarbeiten kann. Dies geschieht z.B. wenn bestimmtes Vorwissen fehlt, oder die Informationen in einer Art und Weise aufbereitet sind (z.B. chemische Strukturformeln) die nicht von allen potentiellen Konsumenten (z.B. Marketing) verstanden werden kann.

2 Exkurs: „Klassisches“ Information Retrieval

Wie die Einführung dieses Dokuments veranschaulicht, ist das Auffinden von (business-relevanten) Informationen für den Anwender (Information Worker) mit grossem zeitlichem Aufwand (und damit Kosten) verbunden.

Das Problem ist nicht neu – schliesslich sind im Bereich der Fachinformationen und Online-Datenbanken Techniken des Information Retrieval seit langem etabliert – jedoch ist deren Einsatz in Verbindung mit Internet oder Intranet-Applikationen in Unternehmen bislang noch wenig verbreitet.

Hier kommen meist noch konventionelle Suchmaschinen zum Einsatz, welche aber zunehmend mit dem Mengengerüst und der steigenden Komplexität von Anfragen an Ihre Grenzen stossen.

Um zu verstehen, wo die Grenzen herkömmlicher Suchtechnologien liegen und wie durch gezielten Einsatz von Produkten mit IR-Technologien Verbesserungen zu erzielen sind, werden in diesem Kapitel zunächst einige grundlegende Konzepte aus dem IR-Themenfeld eingeführt und darauf aufbauend dann die Aspekte des Einsatzes dieser Technologien in Intra- und Internet-Szenarien beleuchtet.

In allen klassischen IR-Systemen werden die Dokumente intern vereinfacht bzw. normalisiert repräsentiert. Die Verallgemeinerung beinhaltet im Wesentlichen die Reduzierung und Normalisierung der Inhalte sowie die Indexierung und Kategorisierung basierend auf unterschiedlichen Verfahren. Wesentlich ist jedoch beim klassischen Information Retrieval, dass der Inhalt immer zum Zweck der Suche aufbereitet wird[6].

2.1 Grundbegriffe

» Recall oder Vollständigkeit: Die Vollständigkeit drückt aus, wie viele relevante Dokumente gefunden wurden. Wünschenswert ist ein Wert möglichst nahe bei 1.

[6] Diesen Aspekt hat man im Inter- und Intranet lange Zeit ingnoriert, erst in jüngerer Zeit gewinnt dieses Thema an Akzeptanz und erhält insbesondere im Zusammenhang mit Suchmaschinenmarketing eine zentrale Bedeutung.

$$\text{Recall} = \frac{\text{gefundene relevante Dokumente}}{\text{relevante Dokumente insgesamt}}$$

» Precision oder Trefferquote: Die Trefferquote gibt an, wie viele relevante Dokumente sich in den insgesamt gefundenen Dokumenten befinden. Sie stellt somit ein Mass für die Güte der gefunden Dokumente dar. Auch hier ist ein Wert nahe 1 erstrebenswert.

$$\text{Precision} = \frac{\text{gefundene relevante Dokumente}}{\text{gefundene Dokumente insgesamt}}$$

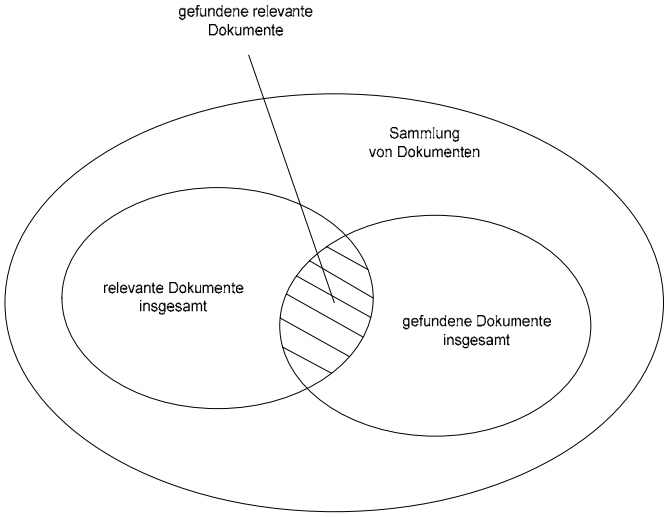

Abbildung 2: Menge der relevanten und gefundenen Dokumente

2.2 Fachinformationsmarkt und Online-Datenbanken

Information (das digitale Gut) kann auch als Ware gesehen werden. Diese Ware wird von Organisationen und personalen Informationsassistenten (z.B. Informationsbroker, Dokumentare u.a.) in Online-Datenbanken aufbereitet gesammelt oder erzeugt und vermarktet. Deswegen wird das Zusammenspiel von spezifischen Informationsanbietern und informationssuchenden Benutzern als Fachinformationsmarkt bezeichnet.

Online-Datenbanken enthalten für ein bestimmtes Fachgebiet (Themengebiet) aufbereitete Daten, die so genannte Fachinformationen. Dabei

kann man grundsätzlich zwischen faktenbasierten-, textbasierten- und hybriden Datenbanken unterscheiden, wie Abbildung 3 veranschaulicht. Benutzer von Online-Datenbanken sind in der Regel Fachleute (z.B. Wissenschaftler) und professionelle Nutzer (z.B. Informationsbroker und Informationsmanager). Um nach den genauen Fachinformationen zu suchen werden von den Anwendern komplexe Anfragesprachen verwendet.

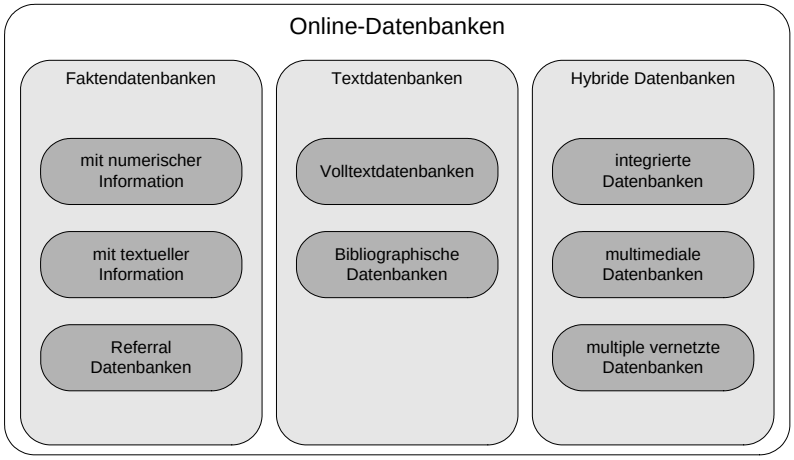

Abbildung 3: Taxonomie der Online-Datenbanken

Beispiele für professionelle und kommerziell genutzte Online-Datenbanken sind die Rechtsprechungsdatenbank des schweizerischen Bundesgerichtes[7] (vgl. Abbildung 4) oder die Datenbanken zum Gesundheitswesen, Medizin und Pflege DIMDI[8]. Wegen ihrer hohen Qualität der Inhalte (bedingt durch redaktionelle Arbeit der Anbieter) sind solche Dienste meistens gebührenpflichtig.

[7] http://www.bger.ch/index/juridiction/jurisdiction-inherit-template/jurisdiction-recht.htm
[8] http://www.dimdi.de/de/db/recherche.htm

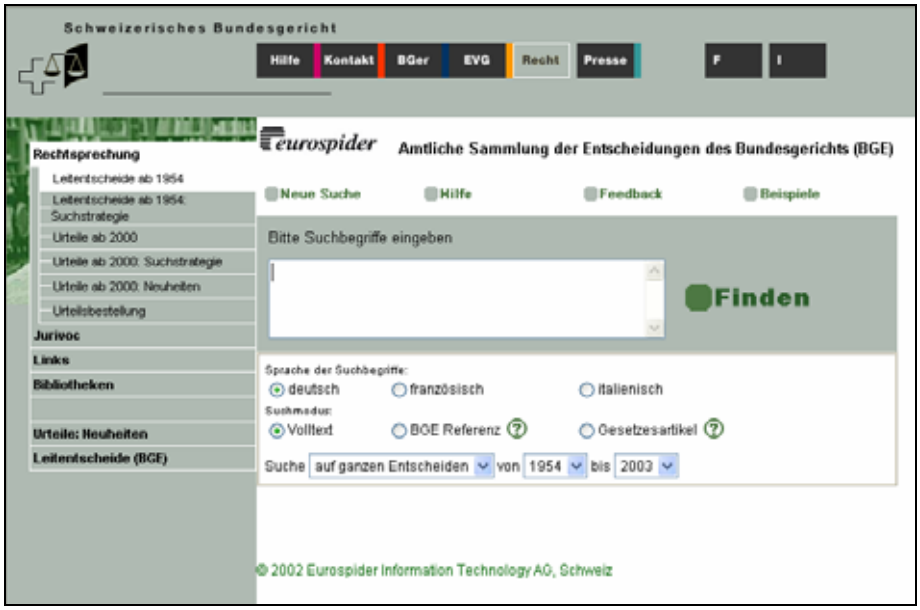

Abbildung 4: Suchmaske beim Bundesgericht; Quelle: http://www.bger.ch/index/juridiction/jurisdiction-inherit-template/jurisdiction-recht.htm

2.3 Information Retrieval Modelle

Die klassischen IR-Systeme werden grob nach zwei Kategorien von Modellen unterschieden. Die erste Kategorie wird von Modellen mit exakter Übereinstimmung gebildet, die zweite von Modellen mit bestmöglicher Übereinstimmung, die sich wiederum noch genauer unterteilen lassen (vgl. Abbildung 5). Gemeinsam haben alle Verfahren, dass die Daten immer im Hinblick auf eine spätere Suche modelliert werden.

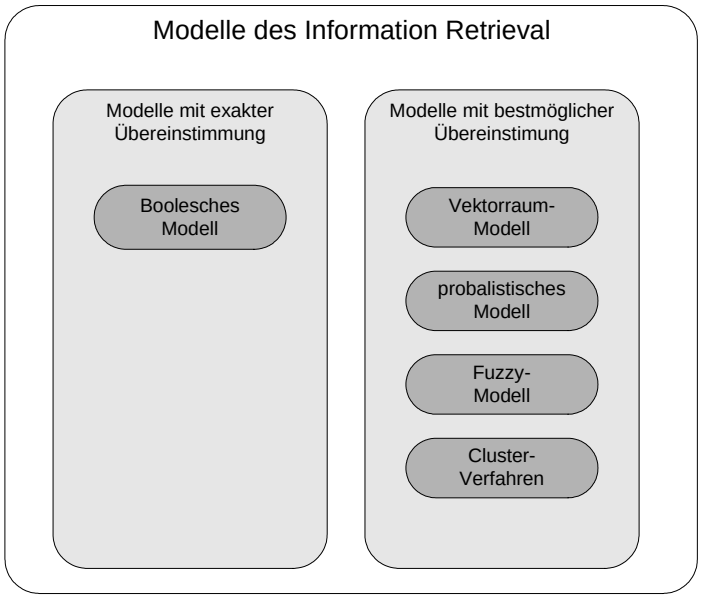

Abbildung 5: Klassifikation der IR-Modelle

2.3.1 Modelle mit exakter Übereinstimmung

IR-Systeme, die einem Modell mit exakter Übereinstimmung folgen, ermitteln für jedes Objekt einen Wahrheitswert, der angibt, ob das Dokument für eine Anfrage Relevanz besitz oder nicht. Modelle mit exakter Übereinstimmung sind mathematisch einfach und schnell in der Ausführung. Eines der bekanntesten Modelle mit exakter Übereinstimmung ist das **Boolesche Modell**. Die Nachteile dieses Modells sind die schlechte Retrievalqualität[9] und dass das Ergebnis nicht nach Relevanz sortiert werden kann. Zudem kann die Erstellung einer Anfrageformulierung den Benutzer überfordern.

2.3.2 Modelle mit bestmöglicher Übereinstimmung

Hier werden den Objekten der Datensammlung keine Wahrheitswerte zugewiesen, sondern Retrieval-Statuswerte, die einen Vergleich der Objekte

[9] es gibt keine partiellen Übereinstimmungen

anhand ihrer Relevanz für eine Anfrage zulassen. Das Ergebnis dieses Vergleichs ist eine Relevanzrangfolge. Die bekanntesten Ausprägungen der Modelle mit bestmöglicher Übereinstimmung sind das Vektorraummodell und das probabilistische Modell.

» Im **Vektorraum-Modell** werden Eigenschaften betrachtet, die von Objekten der Datensammlung erfüllt bzw. von der Anfrage gefordert werden. Das Vektorraummodell ist ein relativ einfaches benutzerfreundliches Modell, da die Anfrage leicht zu stellen ist. Für allgemeine Dokumentensammlungen liefert dieses Modell sehr gute Ergebnisse und findet deswegen wachsende Popularität in Internetsuchmaschinen.

» Im **probabilistischen Modell** erfolgt die Anfrageauswertung mit Hilfe von Wahrscheinlichkeiten, indem das System abschätzt, wie wahrscheinlich es ist, dass ein bestimmtes Dokument für eine Anfrage relevant ist. Diesen Wahrscheinlichkeiten entsprechend stellt das System die gewonnenen Dokumente in eine Rangfolge. Für spezifische Dokumentensammlungen liefert das Modell gute Ergebnisse, die nach der Wahrscheinlichkeit ihrer Relevanz sortiert wird. Ein Nachteil ist, dass die Häufigkeit eines Terms innerhalb eines Dokuments nicht beachtet wird.

» **Fuzzy-Retrieval** basiert auf dem booleschen Modell. Dabei wird zugelassen, dass die Dokumente in den Indexen mit Gewichten versehen werden. Dies ergibt im Gegensatz zum einfachen Booleschen Modell eine Rangordnung der Dokumente in der Antwort zu einer Abfrage. Nachteile sind die umständliche Frageformulierung und die relativ schlechte Retrievalqualität.

» Beim **Clustering** wird hauptsächlich die Ähnlichkeit von Dokumenten genutzt, um relevante Dokumente zu finden. Die Berücksichtigung der Abhängigkeiten ist ein grosser Vorteil, fast alle anderen IR-Modelle nehmen an, dass die Dokumente unabhängig voneinander sind. Ein Nachteil ist jedoch, dass das Clustering-Modell im Vergleich zu den anderen Verfahren eine deutlich schlechtere Retrievalqualität besitzt.

2.3.3 Übersicht über die IR-Modelle

Tabelle 1 gibt einen Überblick der hier behandelten Information Retrieval Modelle.

	Bool.	Fuzzy	Vektor	Prob.	Cluster
Theoretische Basis:					
- boolesche Logik	x				
- Fuzzy-Logik		x			
- Vektoralgebra			x		x
- Wahrsch.-Theorie				x	
Bezug zur Retrievalqualität		x		x	
gewichtete Indexierung		x	x	x	x
gewichtete Frageterms		x	x	x	
Fragestruktur:					
- linear			x	x	
- boolesch	x	x	x	x	
Suchmodus:					
- Suchen	x	x	x	x	
- Browsen					x

Tabelle 1: IR-Modelle; Quelle: [@FUHR], S.59

2.4 Wissensrepräsentation und Anfragesprache

Bei der Wissensrepräsentation geht es um die Abbildung von Wissen (Content) und der Erschliessung des Informationsinhaltes. Das Ziel ist dabei, das Wissen bei Bedarf möglichst exakt zu finden und in die aktuellen Arbeitsabläufe bzw. in die informationelle Absicherung einer Entscheidung einzubeziehen. Zur Erschliessung des Informationsinhaltes werden Methoden wie der Thesaurus oder die Klassifikation benötigt. Klassifikationen dienen dazu, Themen oder Objekte systematisch zu ordnen (nach z.B. der UDC = Universal Decimal Classification). Thesauri hingegen sammeln, ordnen und verknüpfen relevante Begriffe eines Sachgebiets.

Normalerweise ist eine Anfragesprache immer auf eine Modellierungstechnologie zugeschnitten, d.h. Anfragesprache und Wissensrepräsentation basieren auf demselben Konzept. Das Problem im modernen, speziell internet-basierten IR ist, dass die Anfragesprache in der Regel sehr primitiv ist, da die Nutzer nicht ausgebildet werden können. Es werden nur syntaktische Basis-Operatoren unterstützt. Daraus ergibt sich das zweite

grosse Problem: Die Mächtigkeit der Anfragesprache ist ausschlaggebend für die Komplexität der Fragen die mit Ihr gestellt werden können!

Aus diesem Abschnitt ist ersichtlich, dass beim traditionellen Information Retrieval die Daten immer im Hinblick auf die Suche modelliert werden (= gepflegte homogene Sammlung). Die klassischen Ansätze greifen aber nicht mehr (oder nur mit Einschränkungen) bei der Suche im Web, da dort die Daten in den verschiedensten Formaten (Unheitlichkeit) vorliegen. Im folgenden Abschnitt werden deswegen moderne Information Retrieval Modelle für das Web vorgestellt.

3 Enterprise Information Retrieval Systeme (EIRS)

Wie in den vorangegangenen Kapiteln beschrieben, ist Information als zentraler Erfolgsfaktor der internen und externen Unternehmenskommunikation zu sehen.

Die zunehmende Virtualisierung und Globalisierung der Märkte führt zwangsläufig zu einer gewissen Standardisierung von Produkten und Leistungen konkurrenzierender Unternehmen. Dies ist zum einen bedingt durch die Anforderungen der Beschaffung (eProcurement) und der Lieferantenbeziehungen (Supply Chain Management), auf der anderen Seite aber auch getrieben vom Kunden, der Vergleichbarkeit und Kompatibilität von Angeboten und Produkten wünscht.

Entscheidend für den Markterfolg ist deshalb immer weniger das Angebot selbst, sondern vielmehr die Absicherung des Prozesses, welcher der Entscheidung für oder gegen ein konkretes Angebot vorausgeht. IR-Systeme unterstützen sowohl Mitarbeiter als auch Kunden dabei, alle für einen bestimmtem Anwendungsfall (z.B. Kauf) entscheidungsrelevanten Informationen möglichst effizient zu erhalten.

Grundlage dafür sind leistungsfähige (Enterprise) Content Management Systeme, die eine bisher nicht gekannte Verfügbarkeit von Wissen gewährleisten. Die entscheidende Frage dabei ist jedoch: Wie kommt das relevante Wissen in den Kontext einer Transaktion? Die Antwort darauf sind moderne, integrierte IR-Systeme und Technologien.

Im Information Retrieval werden Informationssysteme in Bezug auf ihre Rolle im Prozess des Wissenstransfers vom menschlichen Content-Manager (Wissensproduzenten) zum Informations-Nachfragenden betrachtet. Content Management Systeme, Datenbanken und ERP-Systeme als „Hort des Wissens“ in Unternehmen stellen dabei hohe Anforderungen an die Qualität und die Funktionalität von IR-Systemen, die auf wissensbasierten und computerlinguistischen Verfahren basieren.

3.1 Abgrenzung zum klassischen Information Retrieval

Es gibt wesentliche Unterschiede zwischen dem klassischen IR und dem modernen Web Information Retrieval. Tabelle 2 stellt die Eigenschaften dar.

Merkmal	**Klassisches IR**	**Enterprise IR**
Wissensrepräsentation und Metainformation	Ist auf Suche ausgelegt, homogen strukturiert, statisch, Metadaten meist vollständig und umfangreich	Auf Unternehmensprozesse ausgelegt, heterogen strukturiert, dynamisch, wenig Metainformation
Umfang der Systeme	Sehr gross	Eher klein (Unternehmen) / „unendlich“ (Internet)
Anfragesprache	Komplex, gute Retrieval-Qualität.	Einfach, natürliche Sprache.
Such-Kompetenz der Anwender[10]	Sehr hoch professionelle Informationsvermittler	Gering, untrainierte Enduser
Fachkompetenz der Anwender	Mittel, meist nur als Vermittler tätig	Hoch, professionelle Anwender mit meist sehr spezifischem Fachwissen

Tabelle 2: Merkmale IR - EIR

[10] vgl. Abbildung 6

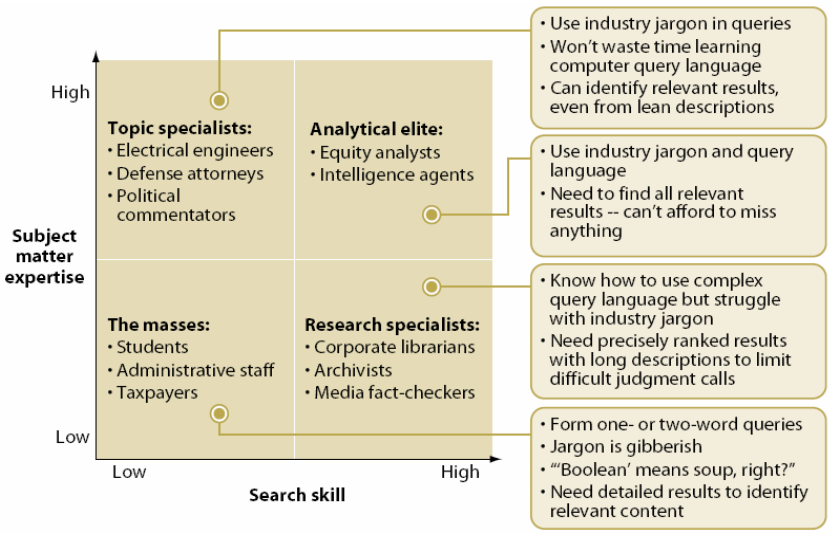

Abbildung 6: Anwender unterscheiden sich im Such- und Domänenwissen; Quelle: Forrester Research Inc.

3.1.1 Besonderheiten des Internet

Im Gegensatz zum klassischen Information Retrieval umfasst die Suche im Web grundsätzlich alle im Internet verfügbaren Datenquellen. Die Besonderheiten werden im Folgenden angegeben:

- riesige Datenmenge mit exponentiellem Wachstum (vgl. Abbildung 7): Grösse „surface web" (direkt adressierbare Dokumente) ca. 167 Terabytes, Grösse „deep web" (indirekt in Datenbanken abgelegte Informationen) ca. 91850 Terabytes [@BERG]
- hohe Volatilitätsrate; Schätzung: 40% des Webs verändert sich monatlich [@KOTA]
- hoher Anteil an unstrukturierten und redundanten Daten; Schätzung: 30% der Daten sind Mirror-Seiten, Kopien oder ähnliche Seiten, hohe semantische Ähnlichkeit [@KOTA]
- aufgrund der Eigenschaften des Webs sehr verteilter heterogene Datenbestand (vgl. Abbildung 8)
- grosse Qualitätsunterschiede z.B. tote Links, veralterte Seiten, Scann- und Tippfehler, grammatikalische Fehler usw.
- Sprache: Aufgrund der Verbreitung des Web hoher Anteil an Mehrsprachigkeit

» die Anwender kamen anfangs aus dem wissenschaftlichen/studentischen Bereich. In den letzen Jahren Verschiebung zu privaten Nutzern (ungefähr 600 Mio. Menschen haben Zugang zum Internet)
» Veränderung von einer rein wissenschaftlichen Ausrichtung zu einer Informations- und Handelsplattform
» die klassischen Metriken Recall und Precision (vgl. Kap.2.1) können wegen der unbekannten Menge potentieller relevanter Dokumente im WWW nicht gemessen und in Zahlen ausgedrückt werden. (Menge der relevanten Dokumente geht gegen unendlich)

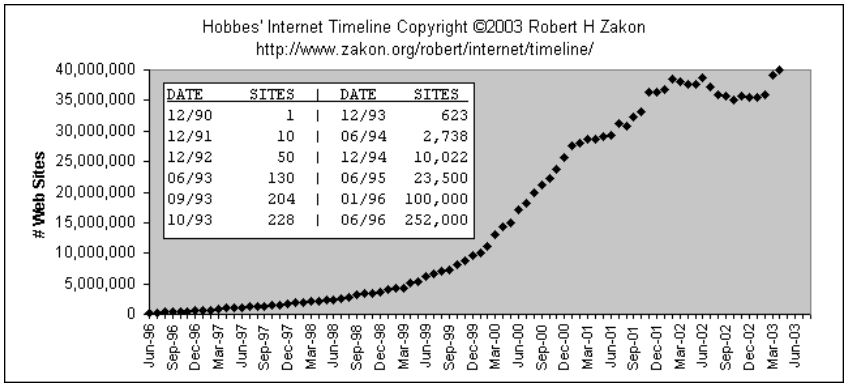

Abbildung 7: Wachstum der Web-Sites nach Zakon; Quelle: http://www.zakon.org/robert/internet/timeline/

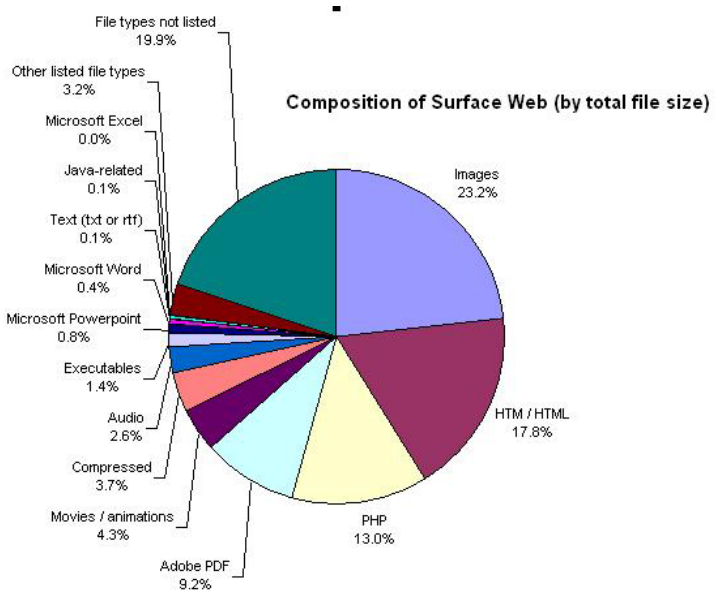

Abbildung 8: Zusammensetzung des Internets; Quelle: [@BERK]

3.1.2 Besonderheiten des Intranet

Information Retrieval im Intranet umfasst grundsätzlich alle durch einen Firewall-Rechner geschützten Datenquellen. Im Unterschied zum IR im Internet bestehen vor allem folgende Unterschiede:

- die Datenquellen befinden sich alle in einer abgeschlossenen Einheit und deshalb ist die Datenmenge viel kleiner als im Internet
- der Datenbestand ist über die Jahre hinweg gewachsen
- es besteht eine Vorstellung des Informationsbedarfs der Benutzer, da in der Regel die relevanten Geschäftsprozesse abgebildet sind
- die Qualität der Inhalte ist grösser
- Strukturierung und Formatierung der Inhalte ist einheitlicher
- der Anteil an PDF-Dokumenten und anderen Nicht-HTML-Formaten ist grösser
- es sind aktuell oder potentiell Metadaten-Standards vorhanden
- die technische Infrastruktur ist viel einheitlicher (v.a. Einsatz von CMS-Systemen)
- die Anzahl der Benutzer ist beschränkt (zum Vergleich ca. 600 Mio. im WWW)

3.2 „Big Picture“ eines Enterprise Information Retrieval Systems

Wie aber muss nun ein solches System aufgebaut sein, um diesen Ansprüchen gerecht zu werden?

Auf Grund der Tatsache dass Information kein statisches Gut ist (also nichts, was man auf Vorrat halten kann) muss man Information Retrieval Systeme sowohl strukturell (Architektur) betrachten als auch bzgl. Ihrer Laufzeitaspekte (dynamische Aspekte).

3.2.1 Architektur

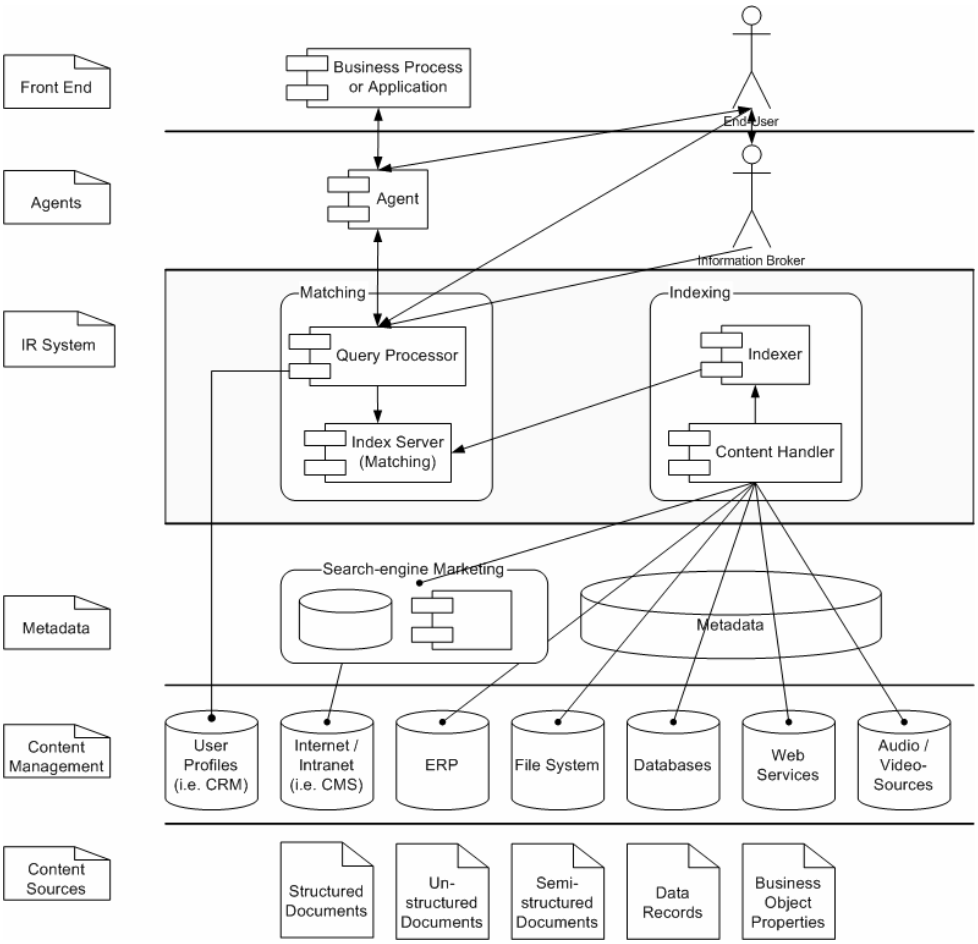

Abbildung 9: Aufbau eines EIRS

Praktisch alle kommerziellen IR-Systeme haben zwei Haupt-Komponenten, die Indexing- und die Matching-Komponente. Während die Indexing-Komponente für die interne Repräsentation der Daten zuständig ist (in der Regel ist dies neben den Basis-Daten vor allem ein normalisierter Index) ist die Matching-Komponente dafür zuständig, eine Suchanfrage gegen diesen Index abzugleichen und somit den auf die Anfrage passenden Teilbereich an Inhalten zu ermitteln. Normalerweise wird die Anfrage zu diesem Zweck ebenfalls normalisiert. Für die Verarbeitung und Normalisierung von Anfragen ist ein Query-Prozessor zuständig. Durch den Einsatz mehrerer Query-Prozessoren können unterschiedliche Anfrageszenarien oder Sprachen unterstützt werden.

Das Gegenstück zum Query Processor ist der Content Handler. Dieser ist ebenfalls dazu da, dem Indexer Content in einer normalisierten Form zu liefern. Normalisiert heisst beispielsweise die Entfernung von Stoppwörtern, Reduktion auf Grund- bzw. Stammform etc. In einigen Fällen ist der Content Handler auch nur für die Konvertierung verschiedener Datenformate (Office, Web etc.) in ein einheitliches internes Format zuständig während alle weiteren Normalisierungsschritte dann vom Indexer ausgeführt werden. So lassen sich dann sehr leicht zusätzliche Formate in den Indexierungsprozess einbinden, indem einfach zusätzliche Content Handler integriert werden. Für Indexer und Matching-Komponente sind diese Schnittstellen zur Aussenwelt in der Regel transparent.

Dennoch gibt es zwei Ebenen in der Architektur eines IRS die nicht direkt Bestandteil eines Produkts sind aber dennoch im Gesamtsystem eine entscheidende Bedeutung einnehmen: An der Schnittstelle zu den Inhalten ist dies die Ebene des (Enterprise) Content Management sowie des Managements der Metainformationen. Wie später noch erläutert wird, werden in klassischen IRS Inhalte immer (und ausschliesslich) zum Zwecke der späteren Suche aufbereitet. Dieser Sachverhalt ist in realen Umgebungen (namentlich in Unternehmen und im Internet) in der Regel nicht gegeben. Diese Rolle wird hier durch den Einsatz von Content Management Systemen und dem gezielten Einsatz von Metadaten bzw. Methoden des Suchmaschinenmarketings übernommen. Nur die ganzheitliche Betrachtung dieser Aspekte wird dem Gesamtsystem den gewünschten Erfolg bringen.

An der Schnittstelle zum Anwender (Information Worker) gibt es ebenfalls Varianten, die alternativ zum direkten (synchronen) Zugriff auf ein IRS angewandt werden können. Hier ist insbesondere der Einsatz von Infor-

mation Brokern (z.B. Informationsvermittlungsstellen) zu nennen, aber auch die Nutzung von Softwaretechnologien wie Agenten. Die Indikation für diese alternativen Dialogformen ist normalerweise dann gegeben, wenn der Anspruch an die Komplexität der abzufragenden Informationen sehr hoch ist (z.B. medizinische oder juristische Fachinformationen), auf der anderen Seite aber das Wissen über die zu verwendende Anfragesprache sehr gering ist. In diesen Fällen braucht es Mediäre, entweder als Person oder als Systemkomponente.

3.2.2 Laufzeitverhalten

Insbesondere im Falle von sehr gutem und anspruchsvollem Fachwissen (Domain Knowledge) und geringer Kompetenz in der Benutzung und Anwendung von IRS bzw. Anfragesprachen ist das Laufzeitverhalten eines IRS von grosser Bedeutung.

Mögliche Laufzeitkomponenten sind beispielsweise Dialogschnittstellen, mit denen ein System in der Lage ist, eine Anfrage eines Anwenders schrittweise und systematisch zu verbessern.

Dieser Aspekt des Dialogs mit dem System ist enorm wichtig, da nur so Abfragen entwickelt werden können die auch tatsächlich ein Maximum an Informationen zurückliefern.

Dies hat mehrere Gründe: Information Worker sind selten in der Lage, eine ad hoc Anfrage zu formulieren, die Ihrem tatsächlichen Informationsbedarf entspricht. Dies liegt zum einen daran, dass man oft gar nicht weiss, was man eigentlich sucht, sondern erst dann wenn man ein bestimmtes Dokument liest bemerkt, dass dieses relevant ist. Zum anderen liegt das aber auch schlicht daran, dass man in der Regel mit den Möglichkeiten der verfügbaren Anfragesprachen nicht vollständig vertraut ist.

Dieses Problem wird verstärkt dadurch, dass Information Worker bislang nicht für systematische Informationsarbeit ausgebildet werden, ganz im Gegensatz zu professionellen Information Brokern (also den Nutzern klassischer IR-Systeme).

Ein weiterer Aspekt des Laufzeitverhaltens ist die Fähigkeit eines Systems, auf der Grundlage des Dialogverhaltens von Anwendern Wissen zu

extrahieren, welches dann in anderen (vergleichbaren) Dialogen wieder verwendet werden kann. Dieses Konzept bezeichnet man oft als Prediktion (Vorhersage) des Informationsbedarfs auf der Grundlage von Mustern im Dialogverlauf.

3.3 Relevante Aspekte von Enterprise IRS

3.3.1 Einsatz von Suchmaschinen

Mit Suchmaschinen (roboterbasierte Suche) wird traditionell im Internet nach Information gesucht. Basis der Suchmaschine sind die Such-Roboter (Web-Roboter), auch Crawler oder Spider genannt. Dieses spezielle Werkzeug der Suchmaschine durchforstet rekursiv laufend das gesamte Web nach angebotenen Webseiten. Aus den Texten werden Nicht-Stichwörter entfernt, sinntragende Wörter extrahiert und als Stichwörter (Inhaltsbeschreibung) an die Suchmaschine gemeldet. In der Suchmaschine werden invertierte Files, so genannte Indizes gepflegt, die von jedem Stichwort auf die entsprechende Seitenadresse verweisen. Als Beispiele für Suchmaschinen können Google[11] oder AltaVista[12] genannt werden.

Aber das Indexieren von Webseiten ist ein viel komplexerer und anspruchsvollerer Vorgang als von den klassischen Datenbanken her bekannt. Die enorme Anzahl existierender Webseiten, ihre rasante Zunahme und die Frequenz der Änderung machen eine laufende Indexierung schlicht unmöglich. Schätzungen zufolge sind 30% bis 40% der im Netz vorhandenen Webseiten durch Suchmaschinen indexiert [@KOTA].

3.3.2 Metasuchmaschinen

Metasuchmaschinen sind Systeme, die meistens nicht selbst mit der ganzen Architektur und Funktionalität einer Suchmaschine ausgestattet sind.

[11] http://www.google.com
[12] http://www.altavista.com

Für den Benutzer bieten sie aber gleichermassen die Möglichkeit, eine selbst formulierte Anfrage über ein Interface einzugeben. Die Anfrage wird nun an mehrere Suchmaschinen weitergeleitet. Die zurückgelieferten Ergebnisse werden dann zusammengeführt, Duplikate rausgefiltert und eine neue Rangfolge gebildet. Mit diesem Vergehen bieten Metasuchmaschinen dem Benutzer eine höhere Retrieval-Qualität an. Als Beispiel für Metasuchmaschinen sei hier MetaCrawler[13] angegeben.

3.3.3 Einsatz von Katalogen

Kataloge oder Verzeichnisse sind weit verbreitet im World Wide Web. Das Katalogsystem (sog. Ordnungsassistent) wird manuell von Redaktoren aufgebaut und ist strukturiert nach Rubriken. Diese sind nach den eigenen Klassifikationen der Anbieter geordnet. Durch die systematische Anordnung findet der Benutzer sehr gut einen Überblick gebenden Einstieg in ein Sachgebiet. Danach findet er durch die Navigation in den hierarchisch aufgebauten Sachgebieten schnell Information mit hoher Relevanz. Weil die meisten Web-Kataloge sehr umfangreich geworden sind bieten sie ausserdem eine Stichwortsuche (boolesche) an. Der berühmteste Vertreter von Web-Katalogen ist sicherlich Yahoo![14].

3.3.4 Informationsassistenten

Intelligente Agenten (sog. technische Informationsassistenten[15]) sind eine Reaktion auf die zunehmende Verbreitung des Internets und der damit verbundenen steigenden Komplexität. Agenten führen spezifische Aufgaben autonom und teilweise asynchron aus und unterstützen somit die Nutzer beim effizienten und zielgerichteten Arbeiten im Internet. Das Spektrum an intelligenten Agenten ist hierbei sehr breit: Such-, Orientierungs-, Zertifizierungs-, Transaktions- und Kommunikations-Agenten[16]. Wesentlich für selbständig agierende Agenten ist ihre Fähigkeit, sich in ihren Aufgabengebieten intelligent zu bewegen, dass sie auf ihre Umwelt reagieren und ihre Lernfähigkeit. Dazu nutzen Agentensysteme Mittel aus

[13] http://www.metacrawler.com
[14] http://www.yahoo.com
[15] nach Kuhlen: Die Konsequenz von Informationsassistenten, S.411
[16] nach Kuhlen: Die Konsequenz von Informationsassistenten, S.224 ff.

dem Gebiet der künstlichen Intelligenz aus. Als Beispiel für einen Agenten kann ResearchIndex[17] angegeben werden: Dieser ist spezialisiert auf wissenschaftliche Publikation und automatisiert viele Arbeiten wie beispielsweise die Textanalyse von PDF-Dokumenten auf Zitate. Ein weiteres Beispiel ist „LEO“, eine intelligenter Assistent, die einem Benutzer des Portals des Kanton Zürich dabei hilft, sich in dem Universum aller Websites des Kantons, seiner Gemeinden und Verbände, zurechtzufinden.

Abbildung 10: Suchmaske Kanton Zürich; Quelle: http://www.ktzh.ch/

3.3.5 Suchmaschinenmarketing

Suchmaschinenmarketing beschäftigt sich mit der Suche oder besser gesagt dem Gefundenwerden im Internet via Suchmaschinen, Metasuchmaschinen und Kataloge. Unter dem Begriff Suchmaschinenmarketing werden daher alle Massnahmen zusammengefasst, die dazu dienen, dass nach Eingabe einer Suchanfrage ein (relevantes) Dokument in der Rangierung der Trefferliste möglichst weit oben erscheint UND darüber hinaus die Trefferdarstellung den User zum Anklicken des Treffers bewegt.

[17] http://www.researchindex.com

Suchmaschinenmarketing wird vor allem von Unternehmen betrieben, die bestrebt sind, dass die Informationen über ihre Produkte, Dienstleistungen und dergleichen im Internet gefunden werden resp. besser gefunden werden als die der Konkurrenz. Im Vordergrund der Massnahmen steht dabei die verbesserte Auffindbarkeit der eigenen Website nach Eingabe eines relevanten Suchbegriffs. Um eine optimale Positionierung zu erreichen wäre eine genaue Kenntnis der Rankingalgorithmen der jeweiligen Internetsuchmaschine von Nöten. Diese gelten jedoch als eines der am besten gehüteten Geheimnisse im WWW. Da diese Algorithmen andererseits auf ähnlichen grundlegenden Prinzipien wie z.B. die Häufigkeiten des/der Suchbegriffe an verschiedenen Stellen eines Dokuments oder die Anzahl und Qualität der Verlinkungen von externen Seiten beruhen, können Internetseiten in der Regel durch relativ einfache Massnahmen bis zu einem gewissen Punkt für Suchmaschinen optimiert werden.

Neben der Suchmaschinen-Optimierung der eigenen Website besteht die Möglichkeit spezielle, auf einen bestimmten Suchbegriff optimierte HTML-Seiten zu erstellen, um so die Positionierung innerhalb der Trefferliste zu verbessern. Solche Dienstleistungen werden von entsprechenden Search Engine Optimization (SEO) Firmen angeboten. Darüber hinaus besteht bei einigen Internetsuchdiensten seit Einführung des sog. pay-per-click-Modells durch das Unternehmen Overture[18] im Jahre 1998 die Möglichkeit, sich Trefferpositionen zu Suchbegriffen zu erkaufen bzw. zu ersteigern. Die Darstellung dieser „sponsored links“ oder „AdWords“ (Google) erfolgt gewöhnlich über, neben oder innerhalb der aus dem Suchmaschinen-Index generierten Trefferliste.

3.3.6 Einsatz und Bedeutung von Metadaten

Bedingt durch die Tatsache, dass das im Internet (oder auch Intranet) repräsentierte Wissen fast ausschliesslich nicht vor dem Hintergrund der späteren Suche organisiert wurde, kommt dem Wissen über dieses Wissen – also den so genannten Metainformationen – entscheidende Bedeutung zu. Folglich dienen Metadaten zur Wissensrepräsentation bzw. zur Inhaltserschliessung im WWW (vgl. Kap.2.4). Sie bieten also eine Möglichkeit zur inhaltlichen, strukturierten Beschreibung (Struktur, Datentyp, Wertebereich, Semantik u.a.) von Webseiten. Metadaten sind zudem auch für

[18] http://www.overture.com

die Metabeschreibung von nicht- textbasierten Objekten nützlich (z.B. Bilder). Suchmaschinen wiederum verwenden (berücksichtigen) die Metadaten, um die Suche zu verbessern.

Für die Speicherung und Übertragung von Metadaten gibt es eine Reihe von Datenformaten und Datenmodelle. Diese Standards werden im Folgenden beschrieben:

» Dublin-Core-Metadaten[19]
Dieses Format dient zur bibliographischen Beschreibung von Dokumenten und anderen Objekten. Es vereinigt 15 Kernelemente zur einfachen Beschreibung einer Vielzahl von Dokumenten.

» Learning-Object-Metadata-Specification (LOM)[20]
Die Spezifikation definiert ein wesentlich komplexeres Beschreibungsformat als das Dublin Core Element Set. Die LOM-Spezifikation setzt sich aus neun Top-Level-Elementen zusammen, die sich wiederum aus Unterelementen zusammensetzt. LOM dient zur Beschreibung digitaler Lehrinhalte.

» Platform for Internet Content Selection (PICS)[21]
Die PICS-Spezifikation erlaubt die Kennzeichnung von Webseiten um z.B. Altersfreigabe von Webseiten zu bestimmen. PICS ist ein Framework zur qualitativen Beschreibung und Auswahl digitaler Inhalte.

» Resource Description Framework (RDF) / Semantic Web[22]
Das RDF ist eine Spezifikation für ein Modell zur Beschreibung der Semantik von Metadaten. RDF setzt sich aus drei Komponenten zusammen und wird mit Hilfe von XML definiert.

» W3C-Standard[23]
Definition der Syntax und Verwendung der HTML-Metatags, wie Sie in den meisten Webseiten Verwendung finden.

---

[19] http://dublincore.org/
[20] http://ltsc.ieee.org/wg12/index.html
[21] http://www.w3.org/PICS/
[22] http://www.w3.org/RDF/
[23] http://www.w3c.org

4 Marktüberblick: Anbieter von IR-Systemen

Auf dem IR-Markt existiert eine Vielzahl von Vendors, die Lösungen für die unterschiedlichsten Anwendungsbereiche anbieten. Abbildung 11 gibt eine Marktübersicht im Bereich der Anbieter von Information Retrieval Software. Die Anbieter können in drei grössere Kategorien eingeteilt werden, die im folgenden erläutert werden.

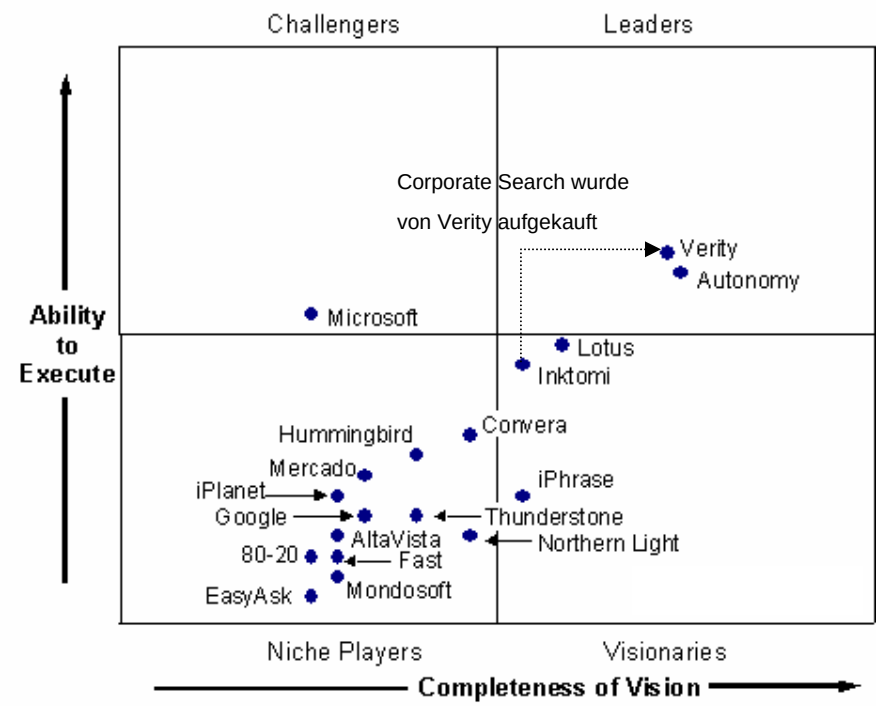

Abbildung 11: Search Engine Magic Quadrant; Quelle: Gartner Group, 2002

4.1 Search Applikationen

Search Applikationen sind umfassende Software-Plattformen bestehend aus mehreren Teilprodukten mit vielen APIs[24], die es einem Kunden ermöglichen, eine massgeschneiderte Lösung für seinen Bedarf der Informationssuche im weitesten Sinne zu erstellen. Search Applikationssoftware wird hauptsächlich für die Suche in Datenbanken und anderen Informationsrepositories eingesetzt. Die Informationssuche in Websites (Internet und Intranet) stellt meistens nur einen Sonderfall dar. Konsequenterweise werden Search Applikationen verhältnismässig selten für die reine Websuche verwendet. Die Anschaffung einer Search Applikation liegt meistens in der Grössenordnung von Hunderttausenden Dollars nur schon für Lizenzen.

4.1.1 Europsider – relevancy 6.0

Produkt:

relevancy 6.0 unterstützt die Integration und den schnellen inhaltsbasierten Zugriff auf alle unternehmensrelevanten Informationen und ist für die Segmente Wissensportale, spezielle Dokumentenkollektionen, Compliance Management, digitale Bibliotheken und Issues Management optimiert. Je nach Marktsegment und Applikation können die benötigten Dienste modular zusammengestellt werden.

Webseite: http://www.eurospider.com

[24] Application Programming Interface (API)

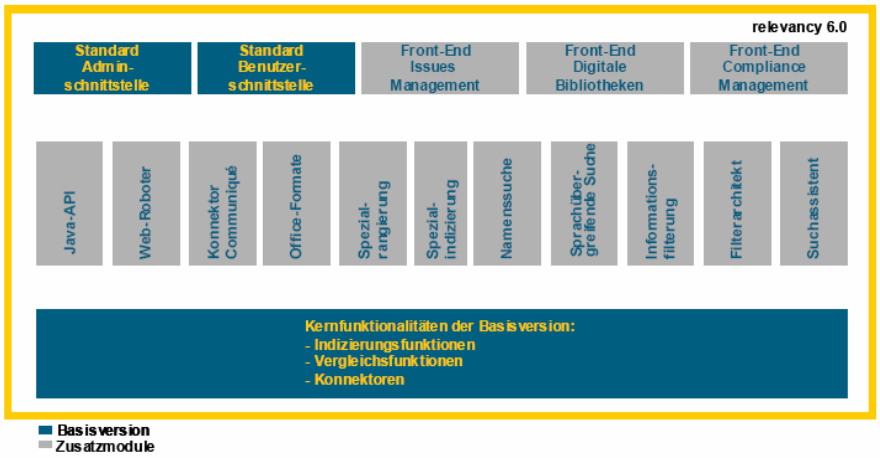

Abbildung 12: Durch die Dienste angebotenen verschiedenen Funktionalitäten von relevancy, die in Kern- und Zusatzfunktionalität unterschieden werden; Quelle: Eurospider

Technische Anforderungen:

» Unix, Linux, Windows NT und Windows XP (auf Anfrage)
» mind. 256 MB RAM, 10 GB Harddisk
» Webserver: Apache 1.3, IIS 5
» Datenbank: PostgreSQL 7.1, Oracle 8/9.

Kunden (Auswahl):

» Credit Suisse Group, ETH Zürich, UBS AG, Roche

Pricing:

» auf Anfrage

4.1.2 Autonomy – Content Infrastructure

Produkt:

Die Architektur von Autonomy ist modular und skalierbar aufgebaut und basiert im Wesentlichen auf der Bayes'schen Inferenztheorie und der Informationstheorie von Claude Shannon. Das Kernelement der Autonomy Content Infrastructure (ACI) bildet die Dynamic Reasoning Engine (DRE).

Webseite: http://www.autonomy.com

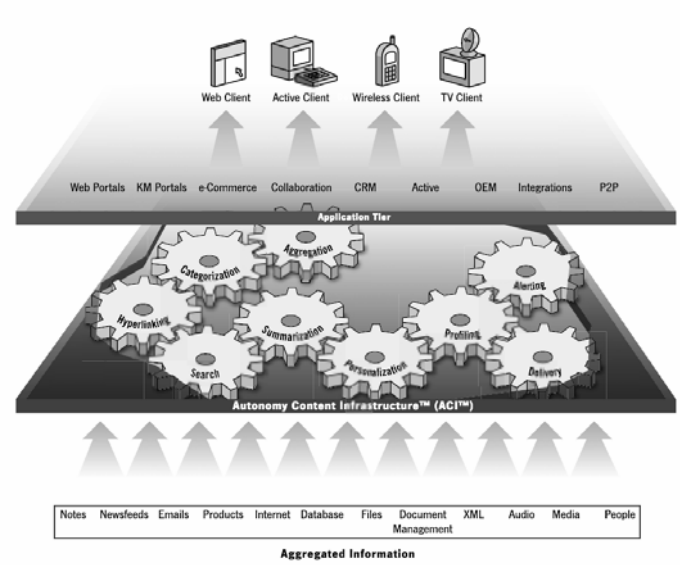

Abbildung 13: Autonomy Content Infrastructure; Quelle: Autonomy

Technische Anforderungen:

» Microsoft Windows NT, 2000 oder XP, HP-UX, Linux, Solaris,

Kunden (Auswahl):

» BAE Systems, Deutsche Bank, Ericsson, Henkel, Infineon

Pricing:

» auf Anfrage (beginnt bei ca. 100’000.- $), Klassenbibliothek erhältlich

4.1.3 Vivisimo – Clustering Engine

Produkt:

Durch die Verwendung mehrerer Clustering-Methoden, die auf dem Gebiet der künstlichen Intelligenz basieren, bietet Vivísimo eine sehr hilfreiche Visualisierung der Suchergebnisse.

Webseite: http://vivisimo.com

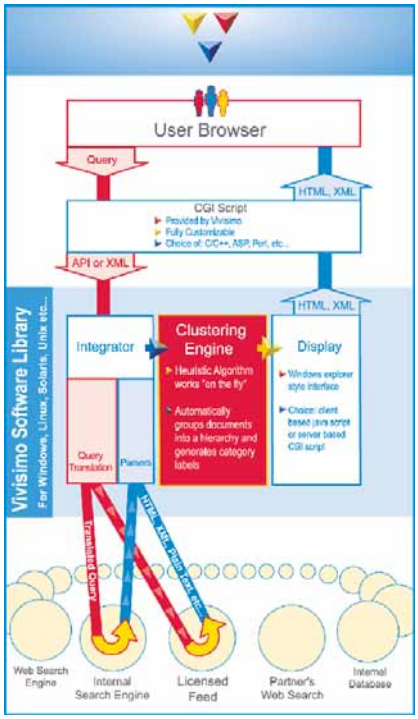

Abbildung 14: Aufbau der Vivisimo Software Library inkl. der Clustering Engine; Quelle: Vivisimo

Technische Anforderungen:

» Windows, Linux, Solaris und andere Unix-Systeme
» HTML-Version: Standard-Browser
» JavaScript-Version: IE 4+ oder Netscape 4+

Kunden:

» keine Angaben

Pricing:

» auf Anfrage

4.1.4 Verity – UltraSeek

Produkt:

Veritys Ultraseek unterstützt Anfragen in natürlicher Sprache, und beinhaltet erweiterte Relevanz-Algorithmen sowie eine Technologie zur linguistischen Analyse von Begriffen, so dass ganze Themenfelder je Begriff erschlossen werden können.

Webseite: http://www.verity.com

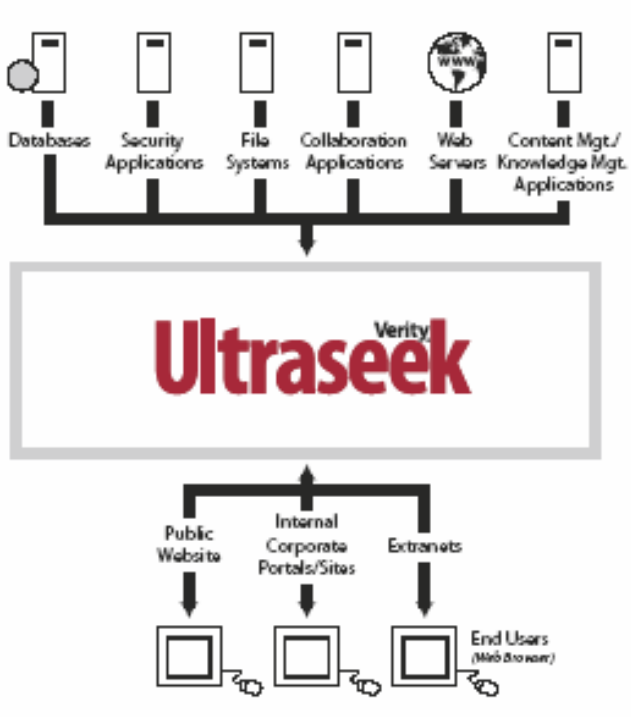

Abbildung 15: Einsatz von UltraSeek in verschiedenen Informationumgebungen; Quelle: Verity

Technische Voraussetzungen:

» Solaris 2.6, Solaris 7 und Solaris 8 auf einer Sun Sparc, Microsoft Windows NT 4.0 und höher (intelbasiert), Microsoft Windows 2000 auf einem PC, Red Hat Linux 6.0 oder höher auf einem PC
» 128MB RAM Minimum. Zusätzlicher Speicher dringend angeraten (schnellere Antwortzeiten)
» 100MB Festplatte für die Anwendung

Kunden (Auswahl):

» Bluewin, Credit-Suisse Group, SBB, UBS AG, Kanton Luzern

Pricing:

» Betrieb auf einem Server: 1'995.- $ für 1-3000 Seiten; 4'995.- $ für bis zu 10000 Seiten

» Enterprise = Multi-Server-Betrieb: 2'995.- $ für 1-3000 Seiten; 7'495.- $ für bis zu 10000 Seiten, Preise für umfangreicheren Betrieb auf Anfrage

4.2 Kompakte Searchprodukte

Kompakte Searchprodukte für Corporate Search haben den Fokus auf der Ermöglichung der Suche auf Websites (Internet und/oder Intranet).

4.2.1 Google – Google Search Appliance

Produkt:

Die Google Search Appliance besteht aus einer integrierte Hardware- und Softwarelösung. Dieser Kleincomputer wird innerhalb des Unternehmensnetzwerkes installiert und soll Kunden unabhängig vom Google-Online-Dienst machen. Das System durchsucht die Firmenrechner nach Dokumenten in beliebigem Format.

Webseite: http://www.google.com/appliance/

Abbildung 16: Modell GB-5005; Quelle: Google

Technische Voraussetzungen:

» Google-spezifisches Linux-System auf gelieferter Hardware

» IE 5+ oder Netscape 4+

Kunden (Auswahl):

» Cisco Systems, Sun Microsystems, Royal Bank of Canada

Pricing:

» Modell GB-1001: 28'000.- $ für 150'000 Dokumente, 50'000.- $ für 300'000 Dokumente, für sicheres System-Crawling zusätzlich 10'000.- $

» Modell GB-5005: 230'000.- $, einschliesslich 1 oder 2 Kollektionen mit bis zu 1.5 Millionen Dokumente pro Kollektion, sicheres System-Crawling

» Modell GB-8008: 450'000.- $ für ein Server-Rack mit sicherem System-Crawling, zusätzliche Load-Balancing Funktionen, und Kapazität für bis zu 5 Kollektionen mit 4 Millionen Dokumenten pro Kollektion

4.2.2 Atomz – Atomz Search

Produkt:

Atomz Search ist eine web-native Suchmaschine, welche die Webseiten durchforstet und die relevanten Informationen in einer „remote" Datenbank speichert.

Webseite: http://www.atomz.com/

Abbildung 17: Atomz Search Architektur; Quelle: Atomz

Technische Voraussetzung:

» alle übers Internet und Intranet zugänglichen Webserver

Kunden (Auswahl):

» Palm, GoreTex, CBSNews

Pricing:

» Free Trial-Version für bis zu 500 Seiten, auf Wunsch Updates, Berichte, keine Werbung – nur ein Atomz Logo

» Gekaufte Version: je nach Anzahl der Domains und Seiten 15'000.- $ und mehr pro Jahr

4.2.3 e-serve – e-serve®Business

Produkt:

e-Serve basiert auf Java und Internet-Technologie wie XML, sowie wissensbasierten Problemlösungs- und Such-Technologien.

Web: http://www.e-serve.ch

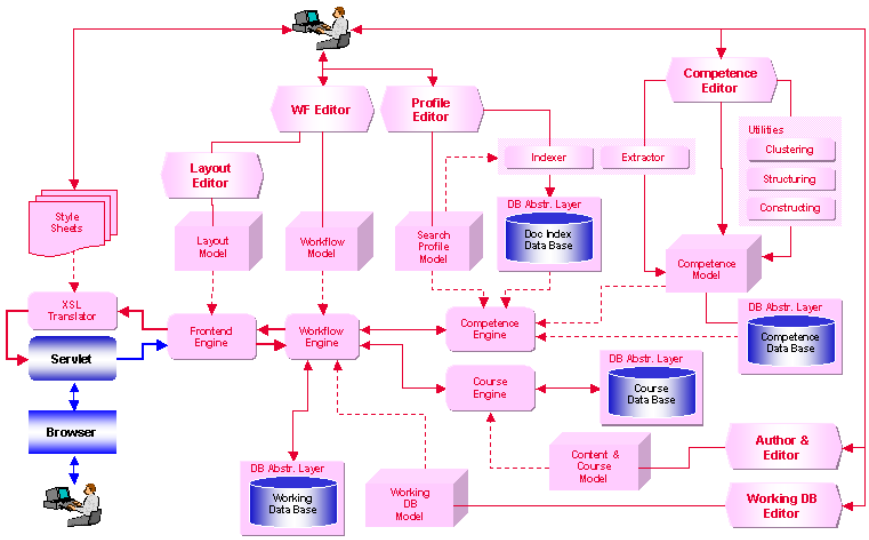

Abbildung 18: e-serve Architektur; Quelle: e-serve AG

Technische Voraussetzungen:

» Aufgrund der Technologie ist e-Serve unabhängig von der Plattform und läuft auf allen kommerziellen Rechnern

Kunden (Auswahl):

» Tarmed, UBS, Schindler, Coop

Pricing:

» auf Anfrage

4.3 Integrierte Searchprodukte

Integrierte Searchprodukte umfassen Searchprodukte, die mit einer bestimmten Softwareplattform integriert verbunden sind und sich deshalb vorzüglich für die Suche nach Informationen eignen, die auf dieser Plattform selbst verwaltet werden. Beispiele für solche Plattformen sind vor allem Content Management Systeme (CMS), Document Management Systeme und Portale, aber auch z.B. Microsofts Windows Internet Plattform. Integrierte Searchprodukte sind meistens schlecht(er) für Informationssuche ausserhalb ihrer angestammten Plattform ausgerichtet.

4.3.1 Microsoft – Sharepoint Portal Server

Produkt:

Der Sharepoint Portal Server 2.0 ist eine flexible Portallösung, die das Suchen, Freigeben und Veröffentlichen von Dokumenten in Unternehmen erleichtert. Microsoft Sharepoint Portal Server 2003 ist Teil der .NET-Server-Familie und basiert somit weitgehend auf .NET-Technologie. Wie alle .NET-Server-Produkte integriert der Sharepoint Portal Server sehr gut mit anderen Microsoft-Produkten, kann aber seine volle Funktionalität auch nur in Zusammenarbeit mit anderen Server-Produkten von Microsoft (SQL Server, Exchange Server und evtl. CM Server) entfalten.

Im Grunde genommen ist SPS keine "reine" Portallösung, sondern er beinhaltet ein Portalsystem (Webpart-Technologie), kann aber je nach Anforderungen auch als einfaches Dokumenten-Management-System oder als Search-Portal eingesetzt werden.

Wichtigste Stärke ist sicherlich die Tatsache, dass Sharepoint zu einem günstigen Preis ein umfangreiches Produkt bietet. Es erübrigt sich in der Regel die Notwendigkeit, weitere Software-Produkte für die genannten Funktionsbereiche einzusetzen.

Webseite: http://www.microsoft.com/sharepoint/

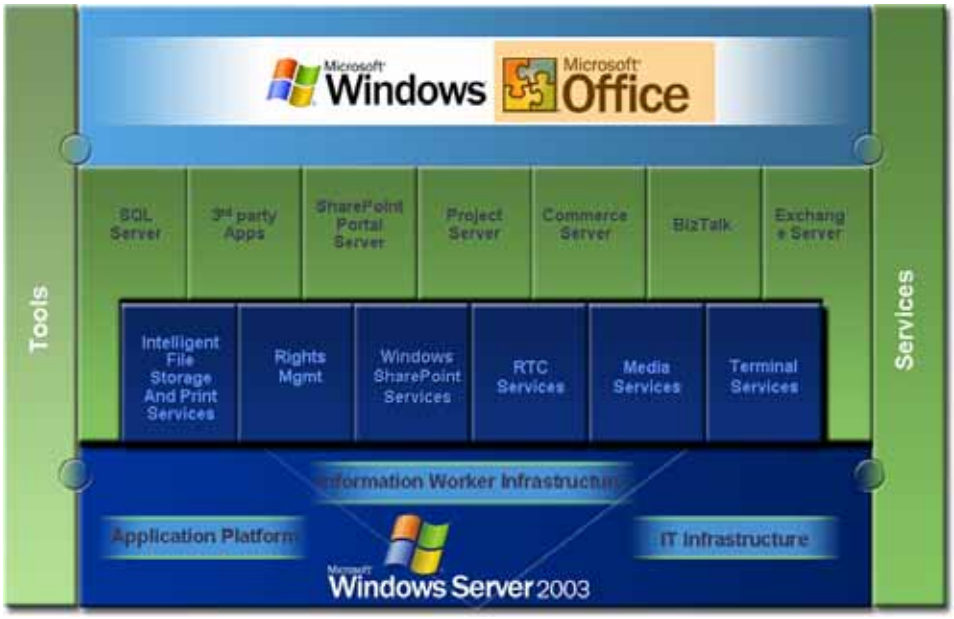

Abbildung 19: Sharepoint als Teil der Information Worker Architektur; Quelle: Microsoft

Technische Voraussetzung:

» Windows 2003 Server mit IIS 6.0
» SQL Server 2000

Kunden:

» Microsoft weltweites Intranet

Pricing:
Unterschiedlich je nach Lizenzvertrag, Richtwerte:

» Server-Software: Sharepoint PortalServer 2003 mit ca. 3999.- $ pro Server
» Client-Lizenzen: pro Anwender ca. 71.- $ oder Internetpauschallizenz

4.3.2 IBM – WebFountain

Produkt:

WebFountain ist eine web-weite Data Mining- und -Entdeckungsplattform, die Trends, Muster und Zusammenhänge aus riesigen Mengen unstrukturierter oder halb-strukturierter Textdaten zieht. Die riesige Suchmaschine wird bereits von Grosskonzernen, Regierungen und Marktforschungsun-

ternehmen genutzt, um die Marktfähigkeit von Produkten oder die Effizienz von Werbekampagnen zu testen.

Webseite: http://www.almaden.ibm.com/webfountain/

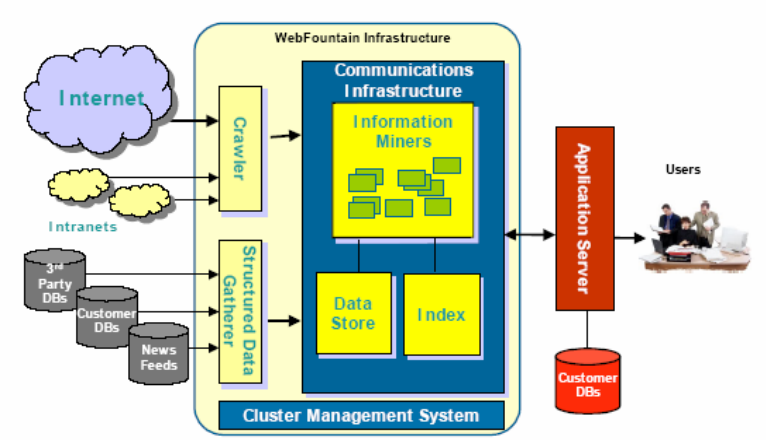

Abbildung 20: WebFountain Architektur; Quelle: IBM

Technische Voraussetzung:

» Linux

Pricing:

» Nutzung ab ca. 100'000.- $

4.4 OpenSource Systeme

Im OpenSource Bereich gibt es keine Lösung, die sich stark aus der Menge hervorhebt und eine bedeutende Präsenz erlangt hat. Einige Systeme sind unten aufgeführt:

» Lucene
Webseite: http://jakarta.apache.org/lucene/docs/index.html
» OpenFTS
Webseite: http://openfts.sourceforge.net/
» mnoGoSearch
Webseite: http://mnogosearch.org/

4.5 Auswahl eines EIR-Systems

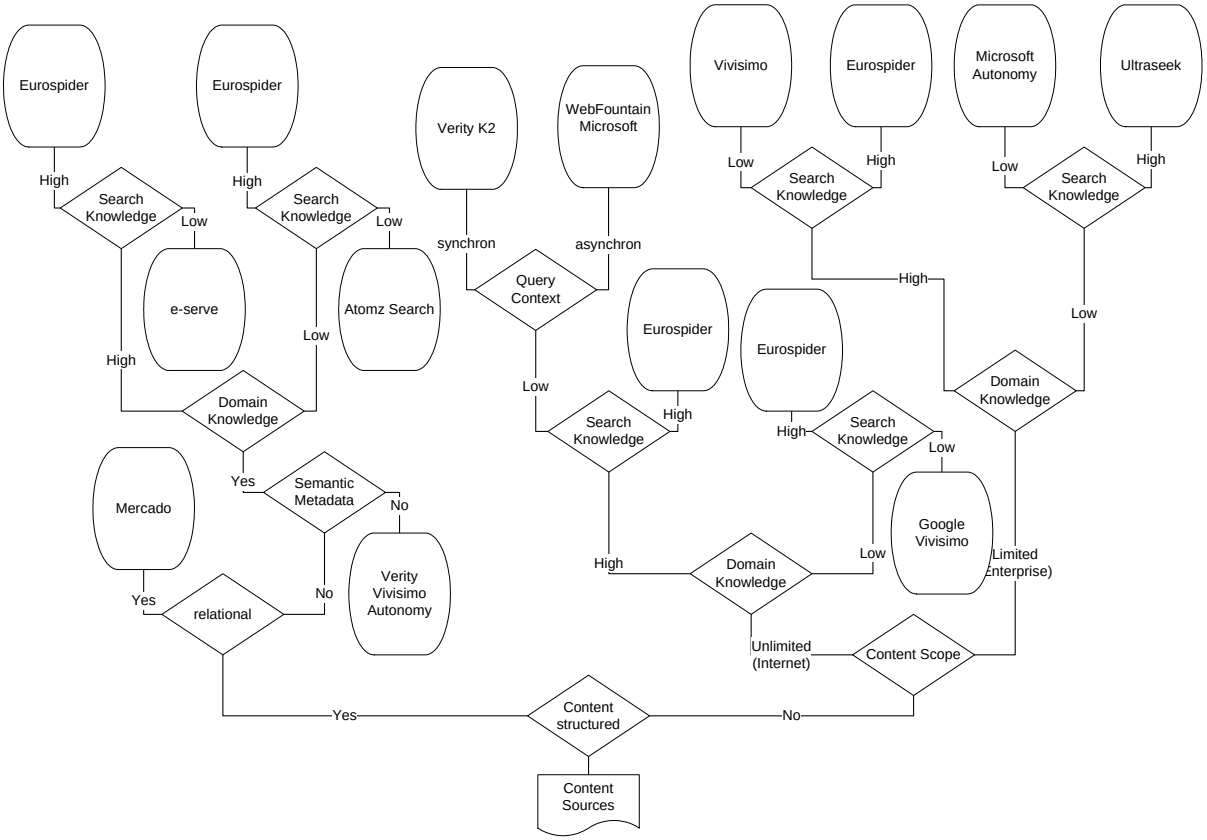

Abbildung 21: Entscheidungsbaum mit Angaben von Anbietern

Die Auswahl eines geeigneten IR-Systems ist ein sehr komplexer Prozess. Der abgebildete Entscheidungsbaum (vgl. Abbildung 21) ist deshalb als grobe Orientierungshilfe zu sehen und muss in jedem Fall individuell verifiziert werden.

5 Leistungsangebot

namics kann auf umfangreiche Erfahrung bei Konzeption und Implementierung von Enterprise Information Retrieval Systemen zurückgreifen. Basierend auf zahlreichen Projekten in verschiedenen Branchen haben sich die folgenden Leistungspakete entwickelt.

Die einzelnen Leistungen können einzeln bezogen werden, bauen aber aufeinander auf und ermöglichen dem Kunden so eine individuelle und transparente Unterstützung seiner Projektvorhaben und sind somit der Garant für einen erfolgreichen Projektverlauf.

5.1 IR-Assessment

Welches Potential kann in Ihrem Unternehmen durch IR-Technologien ausgeschöpft werden?

Dieses Paket besteht aus einer **Analyse des Status Quo**, basierend auf einem standardisierten Interview mit den relevanten Stellen aus IT, Marketing und Organisation. Diese Ergebnisse werden im Hinblick auf das spezifische Verbesserungspotential bewertet und im Rahmen eines Workshops daraus **Szenarien** erarbeitet welche beschreiben wie dieses Verbesserungspotential ausgeschöpft werden könnte und welche Kostenfolgen dies hätte. Ergebnis ist ein **Big Picture** des aktuellen Status Ihres Unternehmens in Bezug auf die informationelle Grundversorgung und eine **Roadmap** wie dieser Status im Rahmen Ihrer geschäftlichen Ziele weiterentwickelt werden kann.

Der Preis für dieses Pakt beträgt pauschal 9'800 CHF.

5.2 Produkte-Evaluation

Mit unserer strukturierten, sowohl wissenschaftlichen wie auch empirisch abgesicherten Evaluations-Methode, stellen wir sicher, dass Ihre IR-Potenziale individuell und umfassend ausgeschöpft werden können.

Die Evaluation eines geeigneten Produktes ist die wichtigste Entscheidung in Bezug auf Erfolg oder Misserfolg Ihrer IR-Vorhaben!

Es ist deshalb von entscheidender Bedeutung alle relevanten Aspekte in Bezug auf Technologie (Funktionalität), Firmenorganisation und Informationsbedarf aller Geschäftsprozesse (insbesondere Marketingprozesse) zu kennen und im Sinne Ihrer Unternehmensziele zu bewerten.

Durch die Kombination von wissenschaftlich abgestützter Beratungskompetenz, Anbieterneutralität und Erfahrungswissen ist namics der ideale Partner für Ihr Evaluationsvorhaben.

Der Aufwand für ein Evaluationsvorhaben richtet sich nach dem Umfang der Arbeiten (Anzahl Anbieter, Umfang des Analyseprozesses, Prototyping). Eine sehr gute Basis für eine Produktevaluation ist das IR-Assessment.

Typische Projektvolumen für Evaluationsprojekte bewegen sich im Bereich 10’000-50'000 CHF.

5.3 Beratung und Implementierung

Für bestimmte Anbieter bietet namics auch die geeigneten Implementierungsleistungen an. Insbesondere wenn IR-Technologie in Verbindung mit CMS eingesetzt werden soll ist hier von grossen Vorteilen auszugehen (Kosten und Nutzen) wenn diese Leistungen „aus einer Hand“ bezogen werden.

Derzeit bieten wir Implementierungsleistungen für die Produkte folgender Hersteller an:

- Microsoft Searcj (als Teil von Sharepoint Portal Server)
- Eurospider Relevancy
- Google Search Appliance
- Autonomy
- Lucence
- Verity Ultraseek und Verity K2

Die Kosten richten sich je nach Umfang der Arbeiten und setzen sich zusammen aus den Lizenzkosten der Produkte und den Dienstleistungen für das Customizing. Typische Projektvolumen bewegen sich im Bereich 50’000-500'000 CHF.

5.4 Kontakt und weiterführende Informationen

www.namics.com/IR

info@namics.com

6 Ressourcen

Literatur

[@RIJS] van Rijsbergen, C.J.: *Information Retrieval*
http://www.dcs.gla.ac.uk/Keith/Preface.html

[@FUHR] Fuhr, Norbert: *Skriptum Information Retrieval*
http://www.is.informatik.uni-duisburg.de/
teaching/lectures/ir_ss03/index.html

[@KOTA] Kobayashi, Mei; Takeda, Koichi: *Information Retrieval on the Web*, IBM Research, Tokyo, 25.April 2000
http://citeseer.nj.nec.com/kobayashi00information.html

[@BERG] Bergmann, M.K.: *The Deep Web: Surfacing Hidden Value*
http://www.brightplanet.com/technology/deepweb.asp

[@BERK] Studie der "School of Information Management and Systems", Univ. of California, Berkeley
http://www.sims.berkeley.edu/research/projects/how-much-info-2003

Webseiten

namics Information Retrieval Webseite
http://www.namics.com/ir

Cross-Language IR Evaluation Forum
http://clef.iei.pi.cnr.it:2002

Gesellschaft für Informatik, Fachgruppe Information Retrieval
http://www.uni-hildesheim.de/~fgir/