Working Papers des Fachhochschul-Studiengangs Informationsberufe – Eisenstadt – www.infomanager.at/KM/

Wissen & Management

2004/2

Berichte aus

Forschung

Entwicklung

und Praxis

Karl Popper

Peter F. Drucker

gefördert durch:

FH plus

Eisenstadt, im Juni 2004

Sehr geehrte Leserin,
sehr geehrter Leser,

mit **Wissen & Management** präsentieren wir Ihnen eine Veröffentlichungsreihe des FH-Studiengangs Informationsberufe / Eisenstadt. Viermal pro Jahr möchten wir Ihnen ausgewählte Arbeitsergebnisse vorstellen. Unser Thema dabei ist der *Umgang mit Wissen in komplexen Organisationen*.

Dieses Mal beschreibt der Beitrag Methoden zur automatischen Inhaltserschließung. Außerdem stellt die Autorin Anforderungen an Softwaretools für „intelligentes Retrieval" auf und vergleicht drei ausgesuchte Produkte.

Wir bedanken uns bei der Lenzing AG für die Zusammenarbeit und wünschen Ihnen eine interessante Lektüre und uns Ihre zahlreichen Rückmeldungen.

Sebastian Eschenbach
Mag. Dr. rer. soc. oec., Dr. rer. nat.
sebastian.eschenbach@fh-burgenland.at

Monika Bargmann
Mag. (FH) für Informationsberufe
monika.bargmann@fh–burgenland.at

Automatische Inhaltserschließung von Textdokumenten

eine Analyse von Softwaretools für maschinelle Indexierung und „intelligentes“ Retrieval am Beispiel der Lenzing AG F&E

*Elisabeth Schwarz**

Aufbauend auf einem theoretischen Einführungsteil soll am Beispiel der Lenzing AG F&E zum einen aufgezeigt werden, welche Methoden der automatischen Inhaltserschließung und -strukturierung in drei ausgewählten Indexierungs- und Retrievalsoftwaretools eingesetzt werden. Zum anderen wird beschrieben, welche Möglichkeiten sie bieten bzw. welche Anforderungen sie abdecken können.

Information Retrieval (IR)

Ziel eines IR-Systems ist es, alle relevanten und möglichst wenig irrelevante Dokumente als Antwort auf eine Suchabfrage zu präsentieren. Der Fokus liegt dabei auf natürlichsprachlichen Texten, die meist keine definierte Struktur besitzen und semantische Mehrdeutigkeiten aufweisen. Zur Interpretation des Inhalts eines Dokuments werden syntaktische und semantische Informationen aus dem Text entnommen und mit dem Informationsbedürfnis eines Nutzers abgeglichen. Wie solche Informationen aus Texten extrahiert und wie sie zur Relevanzbestimmung eines Textes genutzt werden, sind zentrale Fragestellungen im IR, mit denen sich zum Teil auch dieses Working Paper beschäftigt.

Die zwei grundsätzlichen Suchstrategien sind die Suche mit Suchbegriffen – z.B. Boolesche Suche, Fuzzy Search, Natural Language Query – und die Suche mittels Browsing – meist Navigation in einem hierarchischen „Themenkatalog“.

Es besteht also ein sehr enger Zusammenhang zwischen der Art der Inhaltserschließung (manuelle/automatische Indexierung) und den Recherchemöglichkeiten.

„Vocabulary Problem“ und semantisches Retrieval

Wenn in Dokumenten gleiche Sachverhalte mit unterschiedlichen Wörtern (Synonyme) bzw. unterschiedliche Sachverhalte mit denselben Wörtern (Polyseme) beschrieben werden, wird das „vocabulary problem“ am deutlichsten sichtbar. Verfahren,

die die Bedeutungsebene berücksichtigen, beziehen meist statistische Verfahren mit ein, um Häufigkeiten gemeinsam vorkommender Terme zu erfassen. Erweiterungen von Suchabfragen (Query Expansion), die aus einem Thesaurus oder aus der Verwendung eines Relevance Feedback-Mechanismus stammen können, dienen zur Verfeinerung der Suchabfrage. Weitere Ansätze beziehen in der Dokument- und Abfragerepräsentation den Kontext mit ein, in dem sich die Terme befinden.

Konzepte der Wissensstrukturierung und –repräsentation

Die beschriebenen Modelle haben unabhängig vom Gebrauch durch Mensch und/oder Maschine zum Ziel, Inhalte aufgrund ihrer Bedeutung (semantische Ebene) zu erfassen. In der Regel werden sie für ein bestimmtes Fachgebiet oder wenige Fachgebiete erstellt. Dabei produzierte Zusatzinformationen über Dokumente werden als Metadaten bezeichnet.

Hinter den verschiedenen Begriffen zur Wissensstrukturierung und –repräsentation stecken oft dieselben oder ähnliche Konzepte, die sich nur in den unterschiedlich starken Ausprägungen von Merkmalen und in der Art der Strukturierung unterscheiden.

Ein **Thesaurus** setzt sich aus natürlichsprachigen Begriffen zusammen, die aufeinander bezogen sind und deren hierarchische, äquivalente und assoziative Beziehungen dargestellt werden. Mittels terminologischer Kontrolle werden Mehrdeutigkeiten und Unschärfen der natürlichen Sprache aufgelöst. Die Wahl des Zugangsvokabulars soll sowohl fachlich korrekte Benennungen enthalten als auch solche, die Nutzer tatsächlich bei der Informationssuche verwenden.

In einem **semantischen Netz** wird Wissen in Form von Objekten und Relationen in einer netzartigen, graphischen Form dargestellt. Die Knoten können Objekte, Konzepte oder Begriffe repräsentieren, während die Kanten als Verweise zwischen diesen Einheiten dienen.

Ontologien bauen auf dem Konzept semantischer Netze auf. Aufgebaut sind sie in Taxonomien, hierarchischen Strukturen von sich einander ausschließenden Klassen. Das Konzept der Vererbung von Eigenschaften vermeidet Redundanzen und ist Basis für Inferenzmechanismen. Inferenz ist ein Mechanismus zum Ableiten „neuen" Wissens, Wissen, das nicht explizit formuliert und abgelegt sein muss. Wurden z.B. die Relationen ‚arbeitet_in (Mitarbeiter, Projekt)' und ‚behandelt_Thema (Projekt,

Thema)' und die logische Folgerungsregel ‚Arbeitet ein Mitarbeiter X in einem Projekt Y zum Thema Z, dann hat dieser Mitarbeiter X auch Kenntnisse zum Thema Z' definiert, so wird über die Abfrage, wer Kenntnisse zu einem bestimmten Thema hat, Mitarbeiter X gefunden, obwohl dies im Wissensmodell nicht explizit festhalten wurde. Die bekannteste Ontologiesprache, um Dokumente mit Metadaten zu versehen, ist das vom World Wide Web Consortium entwickelte Resource Description Framework (RDF).

Im Zuge der Entwicklung von XML (Extensible Markup Language) hat sich das Konzept semantischer Netze unter dem Namen „**Topic Maps**" sehr verbreitet. Es handelt sich dabei um ein XML-basiertes Austauschformat für semantische Netze. Ihr Einsatz ermöglicht die Verlinkung von Informationsressourcen durch die semantische Modellierung ihrer Inhalte. Gemäß dem Schema semantischer Netze bestehen Topic Maps aus Topics (Konzepte) und Relationen (Associations). Jedes Topic kann beliebig viele interne und externe Ressourcen beinhalten.

Ein weiteres Thema, das oft im Zusammenhang mit Topic Maps fällt, ist die Visualisierung von Wissen. Dokumentsammlungen oder Rechercheergebnisse werden dabei in einer zwei- oder dreidimensionalen Karte dargestellt, wobei Wissenskarten auf eine manuell inhaltserschlossene und Themenkarten auf eine automatisch inhaltserschlossene Dokumentbasis zurückgreifen.

Eine **Klassifikation** besteht aus Klassen, in denen gleichartige Objekte und Themen zusammengefasst werden. Vorwiegend werden hierarchische Relationen eingesetzt. Die hierarchischen Beziehungen dieser systematisierten Zusammenstellung von Begriffen werden durch Notationen repräsentiert.

Das Konzept einer **Taxonomie** ist vergleichbar mit dem einer Klassifikation. Dieser Begriff ist stark vom englischsprachigen Raum geprägt und wird bei Produktanbietern automatischer Klassifizierungssoftware häufig verwendet. Dokumente können nicht nur einer Klasse, sondern mehreren passenden Klassen zugeordnet werden, bei entsprechenden Anwendungen auch mit Hilfe von automatischen Kategorisierungsalgorithmen. Neben der Struktur besteht eine Taxonomie auch aus einer Applikation in Form eines Navigationstools.

Methoden der inhaltlichen Erschließung von Textdokumenten

Neben der manuellen Inhaltserschließung bzw. den computerunterstützten Verfahren, auf die in diesem Beitrag nicht näher eingegangen wird, gewinnen automatische Indexierungsverfahren durch die Zunahme elektronisch vorliegender Textdokumente immer mehr an Bedeutung. Bei umfangreichen Dokumentsammlungen werden über die Texte Indices gelegt, um die Suchgeschwindigkeit zu erhöhen. Als Index-Technik setzt man in den meisten Fällen invertierte Dateien ein, vergleichbar mit einem Buchindex. Die Terme werden automatisch und abhängig von dem jeweils verwendeten Indexierungsverfahren aus den Dokumenten entnommen.

Einfache Stichwortextraktion bzw. **Volltextinvertierung** in Kombination mit booleschem Retrieval ist als Grundlage für die Freitextsuche weit verbreitet, wenngleich sie nicht den Ansprüchen eines automatischen Indexierungsverfahrens entspricht, da keine Termauswahl erfolgt.

Kollektions- und retrievalorientierte Verfahren gehen weg von logischen Verknüpfungsoperationen; bei solchen Systemen werden die gewichteten Indexterme der Suchanfrage und der Dokumentrepräsentation miteinander verglichen und als Ergebnis eine Dokumentreihenfolge geliefert. Es handelt sich dabei um einen statistischen Ansatz, durch den aufgrund der Berechnung von Termgewichten ein Ranking möglich wird. Es werden dabei mit Ausnahme der Stoppwörter nur jene Terme in den Index aufgenommen, die bestimmten Häufigkeitskriterien entsprechen – bezogen auf Dokumentebene, Dokumentsammlung, Ort des Vorkommens im Dokument etc. – und die mit Hilfe verschiedenster Algorithmen ausgewählt werden. Alle analysierten Indexierungs- und Retrievalsoftwareprodukte verwenden statistische Ansätze. Die zwei bekanntesten Gewichtungsmodelle sind das Vektorraummodell und probabilistisches Gewichten. Letzteres eignet sich für Retrieval-Verfahren, die mit Gewichtung erst auf Basis eines bereits gewonnenen Suchergebnisses arbeiten. Dies bezeichnet man als Relevance Feedback.

Informations- bzw. computerlinguistische Verfahren setzen bei der Morphologie, der Syntax und/oder der Semantik an und sind sprachabhängig. Vom Ansatz her lassen sie sich in wörterbuch- und regelbasierte Verfahren unterscheiden. Linguistische Verfahren bewegen sich in der Praxis meist auf morphologischer Ebene (z.B. Wortstammreduktion, Kompositazerlegung) und werden, als vorgeschalteter Prozess, in Kombination mit vorwiegend statistischen oder begriffsorientierten Verfah-

ren eingesetzt. Syntaktische Analysen dienen dazu, aufgrund der Wortnachbarn die Wortart zu bestimmen und korrekte Grundformreduktion zu ermöglichen. Auch Mehrwortbegriffe (Phrasen, Nominalgruppen) können so identifiziert werden. Bei semantischen Analysen wird versucht, kontextuelles Wissen zu verarbeiten, die Unterscheidung von Bedeutungen ist aber mit einfachen informationslinguistischen Verfahren nicht lösbar. Für die Weiterentwicklung der natürlichen Sprachverarbeitung bietet sich beispielsweise die Kombination von statistischen Methoden und semantischer Modellierung (vgl. Ontologie) an.

Im Falle der **Konzept-Indexierung** werden als Basis für den Index nicht Terme, sondern Konzepte verwendet. Da ein Term verschiedene Konzepte unterschiedlich stark repräsentieren kann, muss ein Term in den verschiedenen Dokumenten von mehreren Konzept-Codes mit unterschiedlichen Gewichtungen für das jeweilige Dokument abgebildet werden. Es gibt viele Techniken zum Clustering von Termen, sie stützen sich aber alle auf die Häufigkeit zweier gleichzeitig in einem Dokument vorkommender Terme. Solche Konzept-Indexierungssysteme können auf neuronalen Netzwerkalgorithmen basieren. Damit lässt sich bestimmen, welche Beziehung die Terme zueinander haben und welche Konzepte sie bilden.

Bei **begriffsorientierten Verfahren** und **wissensbasierten Ansätzen** handelt es sich um additive Verfahren, da entsprechende Indexterme aus einem kontrollierten Vokabular automatisch zugeteilt werden. Für die Textanalyse werden wieder statistische und informationslinguistische Verfahren verwendet. Um auch den Kontext, in dem die Wörter auftreten, zu berücksichtigen, reichen begriffsorientierte Verfahren nicht aus, weshalb Wissensbasen aus dem Forschungsfeld der künstlichen Intelligenz miteinbezogen werden (Expertensysteme). Als Grundlage dient das Modell der semantischen Netze.

Um den Einführungsteil abzuschließen, seien noch kurz **verwandte Verfahren** der automatischen Indexierung angeführt: **Hypertext-Indexierung** (Hypertextverlinkungen werden als Erweiterung der Konzepte in den Objekten genutzt, in denen der Link enthalten ist), **künstliche neuronale Netze** (durch Vernetzung von sehr vielen einfachen Schaltungen werden die Signalverarbeitungsprozesse zwischen den menschlichen Nervenzellen nachgeahmt), **Pattern Matching-Verfa**hren (Mustererkennungsverfahren, bei dem anhand erlernter oder implementierter Muster Terme oder Termgruppen erkannt werden), **automatische Textklassifikation** (regelbasiert, Catalog-by-example, statistisches Clustering) und **Text Mining** (angewandte Technologien zur Termextraktion und Aufdecken von Mustern und Beziehungen zwi-

schen Termen und Dokumenten sind maschinelles Lernen, statistische und linguistische Analysetechniken).

Aktuelle Trends und Entwicklungen

Im Folgenden sollen unter anderem durch Miteinbezug von Experteneinschätzung kurz die aktuellen Trends und Entwicklungen in der (vorwiegend) automatischen Inhaltserschließung von Textdokumenten aufgezeigt werden.
Der Trend geht zunächst in Richtung des Aufbaus von speziellen, vorwiegend manuell erstellten Wissensmodellen mit direkter Verknüpfung zu den Dokumenten auf Basis von XML (vgl. Ontologien, Topic Maps), die sich vor allem in der Automobil- und Maschinenbaubranche durchzusetzen beginnen und Taxonomien aufgrund ihrer eingeschränkteren Möglichkeiten verdrängen könnten. Im Bereich der automatischen Indexierung sind derzeit Anwendungen im Einsatz, die über Indexierung (alle Formate), Klassifikation zur Kategorisierung, Text Mining-Technologien und Information Extraction versuchen, digitale Inhalte zu erschließen, es werden aber auch automatische Pattern Recognition-Verfahren zur Erstellung von Wissensnetzen (Ontologien) eingesetzt.
Zur Automatisierung der Indexierungsprozesse machen sich Anbieter vor allem statistische Verfahren und künstliche Intelligenz zunutze. Dabei werden zur Einbeziehung der semantischen Ebene Thesauri, Klassifikationen, Topic Maps, Taxonomien etc. verwendet, um knapp gehaltene Suchstatements effektiv verarbeiten zu können. Der Kombination von statistischen (bzw. neuronalen) und linguistischen Verfahren wird weiterhin eine große Zukunftschance zugeräumt, auch in Verbindung mit wissensbasierten Ansätzen. Generell sind Verfahrenskombinationen im Vormarsch. Weiters liegen in korpusbasierten Verfahren Potentiale (vgl. Konzept-Indexierung), die über Einzelwortebene hinausgehend das Vorkommen gemeinsam auftretender Wörter berücksichtigen. Geforscht wird zudem im Bereich Text Mining, wo versucht wird, aus Inhalten seltene, aber aussagekräftige Terme zu extrahieren (Informationsobjekte wie etwa regionale Begriffe, Aufdecken neuer thematischer Trends etc.). Es zeichnet sich ab, dass z.B. Mining-Methoden bzw. Verfahren auf Ontologiebasis den statistischen Verfahren Konkurrenz machen könnten.
Probleme liegen meist in der Verarbeitung von wenig strukturierten Altdaten; zweiter wichtiger Punkt ist die Frage der Zugänglichkeit zu strukturiertem und unstrukturiertem Wissen mittels Such- und Navigationsmöglichkeiten, um eine Nutzung überhaupt zu ermöglichen. Eine gemeinsame Oberfläche zu allen Informationen ist daher von großer Bedeutung.

Analyse von Indexierungs- und Retrievaltools

Basierend auf einer Situations- und Anforderungserhebung konnten für die Lenzing AG F&E Anforderungen an eine Indexierungs- und Retrievalsoftware abgeleitet werden. Auf die spezifischen Informationsbedürfnisse der chemischen Industrie wird in diesem Beitrag nicht eingegangen. Im Zuge dieser Erhebung wurden neunzehn Forschungsmitarbeiter mit Erfahrung bei Informationsrecherchen und der Nutzung des bestehenden Dokumentenmanagement-Systems befragt. Näher analysiert wurden die Produkte Portal-in-a-Box von Autonomy, RetrievalWare von Convera und SmartDiscovery von Inxight.

Anforderungen

- Die Software sollte die Leistung einer automatischen Indexierung bestehender Informationsquellen und eines Retrievals über eine Suchoberfläche erbringen: Große intern produzierte forschungsrelevante Dokumentbestände werden im DMS abgelegt. Die Software sollte zudem in der Lage sein, die bereits bestehenden Metadaten/Attribute zu übernehmen und suchbar zu machen und die verwendeten Datenformate zu unterstützen. Weiters müssen interne Informationsquellen wie Filesystem, Intranet, Access- und Oracle-Datenbanken etc. erschlossen werden können. Externe Quellen wie relevante Webseiten sollten bei Bedarf ebenfalls miteinbezogen und gespidert werden können. Somit könnten über eine Suchoberfläche zahlreiche interne und externe Quellen abgedeckt werden; dadurch ist es nicht mehr notwendig zu wissen, in welcher Quelle die gesuchte Information gespeichert ist.
- Die Suchmöglichkeiten sollten auch den ungeübten Nutzer beim Formulieren von Suchabfragen unterstützen – zum Beispiel mittels Eingabe von Abfragen in natürlicher Sprache oder Konzeptsuchen – bzw. eine Verfeinerung der Suche ermöglichen, damit das Informationsbedürfnis möglichst klar formuliert („vocabulary problem“) und eine Trefferflut verhindert werden kann. Boolesche Suche wird im DMS nur mäßig genutzt. Eine sprachenübergreifende Suche für Deutsch, Englisch und Französisch sollte grundsätzlich möglich sein, ebenso die Einbindung einer Suche nach Experten für bestimmte Forschungsgebiete. Taxonomien sollten nicht nur als Indexierungswerkzeug eingesetzt werden können, sondern dem Nutzer als „thematischer“ Sucheinstieg zur Verfügung stehen, was sich besonders zum Verschaffen eines thematischen Überblicks eignen würde.

- Alertingfunktionen/Benachrichtigungsdienste sollen den Nutzer je nach Interessensprofil auf das Vorliegen neuer, für ihn relevanter Information hinweisen. Solche Automatismen ersparen eine aktive Suche des Nutzers und verhindern, dass möglicherweise wichtige Informationen übersehen werden.
- Die Art und Weise der Trefferdarstellung soll den Nutzer beim Finden der benötigten Information helfen, indem die wichtigsten Dokumente mittels eines Relevanz-Rankings oben angeführt, Kurzausschnitte der Textpassagen mit den Suchbegriffen und Suchbegriffe in den Trefferdokumenten selbst (Term Highlighting) angezeigt werden. Damit soll gewährleistet werden, dass ein Nutzer möglichst schnell abschätzen kann, ob ein Dokument für ihn von Interesse ist, da die Frage der Relevanz eine subjektive ist und die Entscheidung darüber vom System nur unterstützt werden kann. Eine Visualisierungskomponente, wie zum Beispiel Themenlandkarten, sollte grundsätzlich eingebunden werden können.
- Die verwendete Technologie der Software sollte aufgrund des Vorliegens zigtausender eingescannter Dokumente im DMS fehlertolerant sein.
- Die Software sollte auf einen unternehmensweiten Einsatz hin skalierbar sein.
- Die Software sollte eine zeit- (und kosten-)aufwändige nachträgliche manuelle oder ev. computergestützte Indexierung der Dokumente im DMS überflüssig machen. Die Anschaffungs- und Wartungskosten für die Software sollten die Kosten für eine manuelle und/oder computerunterstützte Indexierung (inkl. Thesauruserstellung, benötigter Software und Fachpersonal) nicht übersteigen.

Der Einsatz eines derartigen Indexierungs-, Klassifizierungs- und Retrievaltools könnten in der Lenzing AG F&E folgende Vorteile bringen:

- Automatische Indexierung und Suche kann in allen elektronisch verfügbaren und ausgewählten Informationsquellen erfolgen. Vor allem die Einbindung aller internen Quellen (DMS, Filesystem, Intranet, Datenbanken etc.) bzw. einiger relevanter Webseiten scheint sinnvoll zu sein.
- Die vorhandenen Suchmöglichkeiten neben Boolescher Suche (natürlichsprachige Abfragen, Konzeptsuche anstatt Stichwortsuche) entlasten den Nutzer bei der Formulierung von Abfragen und ermöglichen Suche auf Bedeutungsebene. Das Produkt RetrievalWare beispielsweise bietet Bedeutungsauswahl, Einbeziehen von Synonymen und verwandten Begriffen an. Dies hat eine höhere Precision zur Folge. Außerdem muss für die Suche in verschiedenen Quellen nur eine Suchlogik angewendet werden. Bei Produkten von Autonomy und Convera sind eine sprachenübergreifende Suche und die Einbindung einer Suche nach Experten

möglich. Der Aufbau und die Integration von Taxonomien zur automatischen Klassifizierung/Indexierung und als Sucheinstieg werden unterstützt. Personalisierungs- und Alertingfunktionen können genutzt werden.

- Der Nutzer wird beim Finden von relevanten Dokumenten durch Relevanz-Ranking, „Kurzzusammenfassungen" und Term Highlighting unterstützt. Das Einbinden von Visualisierungskomponenten (bspw. auf Basis von Taxonomien) ist möglich.
- Viele der forschungsrelevanten Dokumente wurden und werden durch Einscannen erfasst; den Auswirkungen fehlerhafter OCR-Prozesse kann mit fehlertoleranten Technologien (Autonomy oder Convera) entgegengewirkt werden, die auch fehlerhafte oder falsch geschriebene Wörter in Dokumenten wiederauffinden. Dadurch wird ein höherer Recall erreicht.
- Der Einsatz solcher Retrievalsoftwaresysteme kann unternehmensweit ausgebaut werden (Skalierbarkeit). Der hohe Umfang an Funktionalitäten (Expertennetzwerke, sprachenübergreifende Suche, Personalisierung, automatische Klassifizierung etc.) kann nach Bedarf ein- und ausgebaut werden. Die Erschließung und Suche nach Bild, Ton- und Videomaterial mit Hilfe zusätzlicher Module/Tools (möglich durch die Mustererkennungsverfahren von Autonomy und Convera) könnten für die Marketingabteilung von Interesse sein.
- Die nachträgliche manuelle oder computergestützte Beschlagwortung der Dokumente im DMS zur Verbesserung der Suche ist zeit- (und kosten-) aufwändig, da neben Thesaurusaufbau fachkundige Indexierer eingesetzt werden Die Kosten für eine Softwarelizenz mit Basisfunktionalitäten (ohne Taxonomie, sprachenübergreifender Suche, Visualisierung etc.) und Anbindung an alle Quellen belaufen sich auf etwa 100.000 bis 120.000 Euro mit Wartungskosten zwischen 15 und 20 Prozent vom Kaufpreis der Software. Erfahrungswerte für die Administration solcher Systeme sprechen von etwa 10 Prozent eines Personentages pro Woche. Die Kosten für ein computerunterstütztes Indexierungstool und den Aufbau eines hierfür geeigneten Thesaurus würden sich auf ähnliche Beträge belaufen, hingegen mit laufend hohen Personalkosten. Von einer nachträglichen manuellen Indexierung ohne Computerunterstützung ist aus Zeit- und Kostengründen abzusehen.
- Der Einsatz derartiger Tools ist branchenunabhängig und auch für die chemische und pharmazeutische Industrie geeignet. Im deutschsprachigen Raum sind dies beispielsweise Henkel, Novartis (Autonomy), Roche, EMS Chemie (Convera) etc.
- Technologien, die vorwiegend auf statistischen Methoden beruhen (Indexierung, Taxonomieaufbau, Klassifizierung), wie beispielsweise jene von Autonomy, er-

lauben zwar möglicherweise eine schnellere Einführung des Systems; aber gerade dann, wenn eine automatische Indexierung (i.S. einer Klassifizierung) und Suche auf Bedeutungsebene erfolgen soll, sind semantische Technologien (z.B. von Convera) als qualitativ hochwertiger einzustufen.

Der Lenzing AG diente vorliegende Analyse als Übersicht, welche Alternativen zur manuellen Indexierung und reinen Dokumentenmanagementsystemen vorhanden sind und wird als Entscheidungsgrundlage herangezogen, wenn die Lösung in einem Indexierungs- und Retrievaltool gesehen werden. Ein überblicksmäßiger Vergleich der drei Softwareprodukte befindet sich im Anhang.

Zusammenfassend können

- die Integration aller Dokumentenquellen und Formate,
- die nutzerentlastenden Funktionalitäten beim Suchprozess,
- die niedrigeren Kosten im Vergleich zu vorwiegend manuellen Methoden und
- die Skalierbarkeit bis zu einem unternehmensweiten Einsatz

als wesentliche Vorteile bei der Einbindung, der Organisation und dem Finden von elektronisch verfügbaren Dokumenten hervorgehoben werden.

Anmerkungen

* Mag. (FH) Elisabeth Schwarz schloss im Jahr 2003 den Fachhochschul-Studiengang Informationsberufe in Eisenstadt ab. Dieser Bericht basiert auf ihrer Diplomarbeit „Methoden der Inhaltserschließung und Wissensstrukturierung von elektronischen Textdokumenten“. Die Autorin ist per eMail unter e.schwarz@epsilontelecom.com bzw. lisi_schwarz@hotmail.com zu erreichen.

Anhang

Diese Tabelle zeigt einen Vergleich der Funktionalitäten der Softwareprodukte Portal-in-a-box, RetrievalWare und SmartDiscovery. Dieser ist stark vereinfacht und dient als Überblick.

	Portal-in-a-box – Autonomy	**RetrievalWare – Convera**	**SmartDiscovery – Inxight**
1. Indexierung & Retrieval aller elektronischen Infoquellen	✓	✓	✓
2. Übernahme bestehender Metadaten (Altsystem)	✓	✓	✓
3. Suchmöglichkeiten			
3.1. Natürliche Sprache	✓	✓	-
3.2. Konzeptsuche	statistische, Pattern Matching Verf.	semantisches Netz, Pattern Matching Verf.	auf vorwiegend linguistischer Basis
3.3. Boolesche Suche	✓	✓	✓
3.4. Suche nach ähnlichen Dokumenten	✓	✓	✓
3.5. sprachenübergreifende Suche	ja, aber problematisch	ja, aber problematisch	-
3.6. Taxonomien	automatisch/manuell	manuell	automatisch/manuell
4. Alerting - Benachrichtigungsdienste	automatisch	manuell definierbar	-
5. Trefferdarstellung			
5.1. Relevanzranking	statistisch, Relevanz Feedback	statistisch, semantisch	statistisch, linguistisch
5.2. Kurzausschnitte mit Suchbegriffen	✓	✓	✓
5.3. Term Highlighting	✓	✓	✓
5.4. Visualisierungskomponente	✓	von Drittanbietern	✓
6. Fehlertoleranz	Pattern Matching Verf.	Pattern Matching Verf.	-
7. Skalierbarkeit	✓	✓	✓

Quellenverzeichnis

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern Information Retrieval. New York: Addison-Wesley.

Billhardt, H., Borrajo, D., & Maojo, V. (2002). A Context Vector Model for Information Retrieval. Journal of the American Society for Information Science, 53, 236-249.

Burkart, Margarete (1997). Thesaurus. In Buder, M. et al. (Eds.), Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit (pp. 160-179). München: Saur.

Croft, W. Bruce (Ed.) (2000). Advances in Information Retrieval. Recent Research from the Center for Intelligent Information Retrieval. Boston, Dordrecht, London: Kluwer Academic Publishers (KAP).

Delphi Group (2002). Taxonomy & Content Classification. Market Milestone Report. Online unter: http://www.delphigroup.com/coverage/taxonomy.htm (2002-11-19).

Dublin Core Metadata Initiative (DCMI) (2003). DCMI Frequently Asked Questions (FAQ). Online unter: http://www.dublincore.org/resources/faq (2003 -05-17).

Karagiannis, D., & Telesko, R. (2001). Wissensmanagement. Konzepte der Künstlichen Intelligenz und des Softcomputing. München, Wien: Oldenbourg.

Knorz, G. (2001). Visualisierung von Zusammenhängen. Von der Wissenskarte zur interaktiven graphischen Topic Map. Online unter: http://www.iud.fhdarmstadt.de/iud/wwwmeth/publ/paper/Marburg2001/2001_10_15_Visualisierung_Endversion_pdf.pdf (2002-12-18).

Kowalski, G.J, & Marbury, M.T. (2000). Information storage and retrieval systems. Theory and Implementation. 2nd ed. USA: Kluwer Academic Publishers.

Montague Institute (2003). What's an ontology? Online unter: www.montague.com/E1/Society/review/ontologies.shtml (2003-02-06).

Nohr, H. (2001). Automatische Indexierung. Einführung in betriebliche Verfahren, Systeme und Anwendungen. Potsdam: Verlag für Berlin-Brandenburg.

Ontoprise GmbH (2001). Semantisches Knowledge Retrieval. Online unter: http://www.ontoprise.de/documents/Ontoprise_Whitepaper_Knowledge_Retrieval.pdf (2003-02-06).

Persönliche E -Mails an die Verfasserin von Lutz Thielmann, Manfred Hauer, Dieter Merkl, László Domokos

Studiengang Informationsberufe

Information & Knowledge Management

Fachhochschul Studiengänge

Burgenland

Bildung im Herzen Europas.

Der Fachhochschul-Studiengang Informationsberufe ist Österreichs akademische Ausbildung zur Informationsexpertin (Abschluss Mag. (FH) nach 8 Semestern Studium). Hier lernen Sie, Menschen professionell mit Information zu versorgen und dafür Informations- und Kommunikationstechnologie (IKT) einzusetzen. Das heißt:

- Information suchen
  z. B. professionelle Recherche in der Unternehmensberatung

- Information organisieren
  z. B. in der Dokumentation eines Medien-Unternehmens oder in einer wissenschaftlichen Bibliothek arbeiten

- Information vermitteln
  z. B. das Intranet eines internationalen Konzerns gestalten

Kontakt

Homepage: http://www.infomanager.at und http://www.fh-burgenland.at
E-mail: sebastian.eschenbach@fh-burgenland.at oder office.ib@fh-burgenland.at
Adresse: Campus 1, A-7000 Eisenstadt
Telefon: +43-(0)5-9010-6020

Informationsnachmittage im Sommer 2004

12. August, 26. August, 9. September, jeweils von 16 bis 18 Uhr