



# Von textbasiertem zu inhaltsorientiertem Wissensmanagement

## Die Rolle von Terminologien und Ontologien

*Ulrich Reimer*

Business Operation Systems  
Esslenstr. 3  
CH-8280 Kreuzlingen  
ulrich.reimer@acm.org

### **Zusammenfassung**

Das Auffinden relevanter Information in Textdokumenten erfolgt heutzutage mehrheitlich durch Freitext-Retrieval, welches jedoch oft unbefriedigende Ergebnisse liefert. Eine Verbesserung lässt sich erzielen, wenn Dokumentinhalte unter Verwendung einer Terminologie oder gar einer Ontologie explizit beschrieben sind, so dass Suchanfragen auf diesen inhaltsbeschreibenden Strukturen ausgewertet werden, anstatt (nur) auf der Textoberfläche. Der Beitrag stellt zwei unterschiedliche Anwendungen vor, wo die Beschreibung und das Auffinden von Informationen erst durch den Einsatz von Terminologien praktikabel wurden.

### **Abstract**

Searching for relevant information in text documents is often very disappointing. Freetext retrieval as the underlying technology of today's search engines can be improved considerably when a description of the contents of the documents can be searched instead of just the words occurring in the documents. This paper argues that a content-oriented retrieval needs to be based on a terminology or even an ontology. The paper presents two applications to illustrate that such an approach is feasible.

## **1 Sozialisations- vs. Kodifizierungsstrategie**

Eine wesentliche Unterscheidung in zwei Arten von Wissensmanagementstrategien geht zurück auf (Hansen et al. 99). Die Wahl der Strategie ist wesentlich durch das Geschäftsmodell eines Unternehmens (oder einer Organisationseinheit darin) bestimmt. Für Unternehmen, deren



Kernkompetenz eher im innovativen Bereich liegt, ist das Erfahrungswissen der Mitarbeiter von zentraler Bedeutung. Solches Wissen ist vorwiegend impliziter Natur und kann folglich nicht (ohne Weiteres) expliziert werden. Erfahrungswissen lässt sich deshalb nur durch die betreffenden Experten selber einbringen. Die gezielte Förderung des Wissensaustauschs durch Netzwerke steht in Unternehmen mit innovativem Charakter deshalb im Vordergrund ihrer Wissensmanagementstrategie (Sozialisationsstrategie).

Unternehmen, deren Geschäftsmodell primär auf der Massenproduktion von Gütern oder Dienstleistungen basiert, setzen vorwiegend Wissen ein, das sich schon mehrfach bewährt hat, deshalb gut verstanden und leicht in Form von Handlungsanweisungen explizierbar ist. Zur Optimierung ihres Geschäfts sind Unternehmen solcher Art deshalb an der Wiederverwendung des Wissens ihrer Mitarbeiter interessiert. Da dieses Wissen leicht zu explizieren ist, kann es in einer zentralen Wissensablage hinterlegt werden (Kodifizierungsstrategie). Unternehmensweite, zentrale Ablagen dieser Art werden Unternehmensgedächtnisse genannt (Lehner 00, van Heijst et al. 96, Abecker et al. 98).

Eine der größten Schwierigkeiten im Aufbau von Unternehmensgedächtnissen liegt darin, das enthaltene Wissen auf einfache Weise für Mitarbeiter zugreifbar zu machen. Da Wissen in der Regel in Textdokumenten kodiert ist, spielen Verfahren des Text-Retrieval hier eine zentrale Rolle.

## **2 Erschließung von Wissen aus Textdokumenten**

Das Auffinden von Textdokumenten, die relevantes Wissen enthalten, geschieht in der Regel mittels Freitext-Retrieval. Eine Suchanfrage besteht aus einer Menge von Wörtern, die in den zurückgelieferten Dokumenten vorkommen müssen. Freitext-Retrieval funktioniert gut, wenn sich die gewünschten Dokumente durch wenige Wörter identifizieren lassen. Sehr oft ist dies nicht der Fall, so dass entweder gar keine oder zu viele Dokumente zurückgegeben werden, unter denen sich möglicherweise auch relevante befinden, die dann aber erst herauszusuchen sind.

Eine Verbesserung kann durch die Erschließung des Textinhalts und dessen Darstellung in einer formalen Notation erreicht werden. Eine Suchanfrage wird auf diesen formalen Strukturen ausgewertet, wodurch eine (gewisse) Unabhängigkeit von den unterschiedlichen Varianten, dieselben Sachverhalte in natürlicher Sprache zu formulieren, erreicht wird. Relevantes Wissen lässt

sich zielgerichteter auffinden und man „ertrinkt“ nicht mehr in der Informationsflut (Reimer 04).

Die inhaltliche Texterschließung erfordert jedoch einen enormen manuellen Aufwand, der in der Regel kostenmäßig nicht vertretbar ist. Wünschenswert ist deshalb eine automatische Texterschließung. Ansätze hierfür lassen sich grob in drei Hauptrichtungen untergliedern. Wissensbasierte Verfahren unterstützen die linguistische Analyse eines Textes durch Hintergrundwissen über den Weltausschnitt, der dem Text zugrunde liegt (siehe z.B. Hahn/Reimer 99). Dieses Wissen geht über rein terminologisches Wissen hinaus, da es die Bedeutung der Begriffe durch ihre Eigenschaften und ihre Beziehungen untereinander näher definiert. In diesem Fall spricht man auch von Ontologien (Gruber 93, Reimer 04). Für realistische Weltausschnitte hat das benötigte Hintergrundwissen einen erheblichen Umfang, so dass dessen Bereitstellung in der Regel ebenfalls zu kostspielig ist. Lediglich in hinreichend eng eingrenzbar Anwendungsszenarien können wissensbasierte Verfahren realistisch einsetzbar sein.

Rein linguistische Verfahren führen gewisse syntaktische Analysen durch, um für einen Text relevante Begriffe sowie eventuell Beziehungen zwischen ihnen zu identifizieren (Hahn/Romacker 00, Engels/Lech 03). Unter Umständen verwenden sie auch typische Syntaxmuster, um Aussagen aus einem Text zu extrahieren (Berry 03, Hovy et al. 01).

Statistische Verfahren verzichten auf eine weitgehende linguistische Analyse und ziehen stattdessen Häufigkeitsverteilungen von Wörtern in einer Dokumentenkollektion heran, um die für einen Text signifikanten Begriffe zu bestimmen (Reimer 92, Mani/Maybury 99). Linguistische und statistische Verfahren weisen naturgemäß gewisse Fehlerraten auf, weshalb sie nur dort einsetzbar sind, wo dies tolerierbar ist.

Zwischen Freitext-Retrieval auf der einen Seite und der Erschließung von Textinhalten auf der anderen Seite – sei es manuell oder automatisch – gibt es die Möglichkeit, das Retrieval von Textdokumenten durch den Einbezug von Terminologien bzw. Ontologien zu verbessern. Das folgende Kapitel stellt anhand von zwei Anwendungen dar, dass ein solcher Ansatz nicht nur zu signifikanten Verbesserungen führt, sondern auch praktikabel ist.

### 3 Ergänzung von Texten um modelliertes Wissen

Wie oben argumentiert wurde, ist sowohl die manuelle als auch die automatische Inhaltserschließung von Texten nur bedingt einsetzbar. Benötigt werden jedoch Verfahren, die nicht nur eine deutliche Verbesserung von Freitext-Retrieval erzielen, sondern auch breit und kostengünstig einsetzbar sind. Gegenwärtig fokussiert deshalb eine Reihe von Ansätzen darauf, Textdokumente um formale Strukturen zu ergänzen, die lediglich deren Inhalt charakterisieren, ihn aber nicht repräsentieren. Dies ist von der Grundidee her die aus den Bibliotheks- und Informationswissenschaften wohlbekannte Indexierung. Da auch diese eingeschränkte Art der Inhaltserschließung automatisch kaum praktikabel ist (Reimer 92), bleibt die manuelle Indexierung. Die Erfahrung zeigt allerdings, dass Mitarbeiter in einem Unternehmen Textdokumente kaum indexieren (trotz bestehender Möglichkeit in den Applikationen). Das hat mehrere Gründe:

- zusätzlicher Zeitaufwand bei gleichzeitigem Zeitdruck,
- die Indexierung ist nur ein optionaler Arbeitsschritt,
- es steht kein Vokabular zur Verfügung, aus dem ausgewählt werden kann.

Die Einführung von Terminologien greift am letzten Punkt an. Die folgenden beiden Abschnitte illustrieren anhand zwei unterschiedlicher Anwendungen, dass der Einsatz von Terminologien für die Realisierbarkeit von Informationssystemen mit hoher Retrieval-Güte entscheidend sein kann.

#### 3.1 Skills Management: Einsatz einer manuell erstellten Terminologie

In seiner primären Aufgabe stellt Skills Management sicher, dass Mitarbeiter eines Unternehmens mit bestimmten Fähigkeiten leicht aufzufinden sind. Dazu werden Beschreibungen der Fähigkeiten und Erfahrungen der Mitarbeiter benötigt sowie die Suchmöglichkeit in diesen Beschreibungen. Darüber hinaus gehende Aspekte von Skills Management sind der Abgleich der für ein Unternehmen benötigten Qualifikationen mit den Mitarbeiterprofilen sowie Maßnahmen, die Ist-Situation den Anforderungen anzugleichen, z.B. durch gezielte Fortbildungsmaßnahmen sowie das Einwerben neuer Mitarbeiter. Im Folgenden betrachten wir ausschließlich den Aspekt der Suche nach Mitarbeitern mit bestimmten Fähigkeiten.

Es gibt sehr unterschiedliche Ansätze für Skills Management (z.B. Ackermann et al. 99, Becerra-Fernandez 00, Maybury et al. 00, Sure et al. 00). Der hier beschriebene, bei Swiss Life entwickelte Ansatz (Reimer et al. 03) sieht für jeden Mitarbeiter eine Home Page auf dem Unternehmens-

Intranet vor, die sich als ein strukturiertes Textdokument darstellt. Mitarbeiter mit bestimmten Fähigkeiten können dann – zunächst – mittels Freitext-Retrieval über das Intranet aufgefunden werden.

Da es sich bei den Home Pages um sehr kurze Dokumente handelt, machen sich die Schwächen von Freitext-Retrieval besonders bemerkbar, denn aufgrund ihrer Kürze ist die Wahrscheinlichkeit ziemlich klein, dass in einer Anfrage derselbe Begriff für eine Fähigkeit verwendet wird wie in den Seiten aller Mitarbeiter, die diese Fähigkeit aufweisen. Diese Problematik lässt sich durch die Verwendung einer Terminologie lösen: Die Beschreibung von Fähigkeiten erfolgt mit demselben Vokabular wie die Formulierung einer Suchanfrage. Für den Benutzer entsteht dadurch kein Mehraufwand. Durch die Vorgabe eines konkreten Vokabulars wird er sogar unterstützt, da er lediglich Begriffe aus einer vorgegebenen Menge auswählt und sich nicht selber adäquate Begriffe überlegen muss – Letzteres ist in der Regel viel aufwändiger.

Eine Terminologie darf nicht zu groß sein, damit die Auswahl relevanter Begriffe einfach und schnell bleibt. Die Terminologie für die Skills-Management-Anwendung bei Swiss Life enthielt etwa 700 Begriffe und deckte die Bedürfnisse der Einzellebensversicherung, der Informatik und der Personalabteilung ab. Die Erfahrung mit der Anwendung zeigte, dass dies tendenziell schon zu viele Begriffe für diesen Zweck waren.

Neben einer Beschreibung der Fähigkeiten eines Mitarbeiters enthält eine Home Page weitere Angaben, wie beispielsweise zur Ausbildung, zur Mitarbeit in Projekten, zu früheren Tätigkeiten etc. Hierbei handelt es sich um frei formulierten Text, der mittels Freitext-Retrieval suchbar ist. Während des Einsatzes der Skills-Management-Anwendung bei der Swiss Life traten Fälle auf, wo es sehr hilfreich gewesen wäre, eine Suchanfrage auf einen Teil der Home Page einschränken zu können, z.B. auf Informationen zur Mitarbeit in Projekten.

Die Einschränkung einer Suchanfrage auf solche Kontexte lässt sich erreichen, indem man die Untergliederung einer Home Page in einer formalen Struktur explizit macht und sie so hinterlegt, dass sie in einer Anfrage auswertbar ist. Die einzelnen Textabschnitte in dieser Untergliederung sind durch Begriffe aus einer Terminologie indexiert. Die Terminologie entspricht dem Informationsmodell, das der Untergliederung zugrunde liegt. Da eine Home Page durch Instanziierung aus einem vorgegebenen Grundmuster erzeugt wird, ergeben sich die formale Struktur und die dahinter liegende

Indexierung als Nebeneffekt der Instanziierung automatisch, ohne Zutun des Benutzers.

Sobald die Möglichkeit besteht, eine Home Page formal in Kontexte zu untergliedern, lässt sich dieser Vorgang auch rekursiv anwenden. So könnte ein Abschnitt „Mitarbeit in Projekten“ statt aus unstrukturiertem Text aus einer Liste von Projektbeschreibungen bestehen. Jede Projektbeschreibung würde sich wiederum aus verschiedenen Attributen zusammensetzen, z.B. aus Projektname, Auftraggeber, inhaltliche Beschreibung, verwendete Technologie, Methoden, etc. Man sieht, dass auf diese Weise aus einer textuellen Beschreibung eine begriffliche Beschreibung wird, die auf einer zugrunde liegenden Ontologie basiert. Dieser Übergang ist fließend und nichts anderes als eine zunehmend formaler werdende Modellierung. Aspekte hoher Wichtigkeit lassen sich detaillierter modellieren, andere Aspekte entsprechend weniger detailliert, eventuell lediglich als Freitext.

### **3.2 Reformulierung von Anfragen mit automatisch gelernter Terminologie**

Die im vorangehenden Abschnitt beschriebene Anwendung zu Skills Management setzt eine Terminologie schon bei der Erstellung von Dokumenten (Home Pages) ein, um ein besseres Retrieval zu erzielen. Im Falle von bestehenden Dokumentensammlungen ist das nur mit erheblichem Aufwand durch eine nachträgliche manuelle Indexierung möglich. Eine automatische Indexierung wäre eine Alternative, jedoch mit allen Einschränkungen, wie sie in Kapitel 2 genannt wurden, kaum praktikabel. Ein Ansatz, der einen Mittelweg zwischen manueller und automatischer Indexierung darstellt, wurde in einem weiteren Projekt bei der Swiss Life verfolgt (Reimer et al. 03).

Das Projekt hatte die Aufgabe, das Retrieval auf einer umfangreichen, sehr wichtigen Dokumentensammlung zu verbessern. Eine zunächst durchgeführte Feldstudie hat ergeben, dass die Mitarbeitenden, die diese Dokumentensammlung benutzen, nicht bereit sind, Zeit in die Formulierung einer Anfrage zu investieren. Es bestand der ausdrückliche Wunsch, nicht mehr als zwei bis drei Suchbegriffe eingeben zu müssen, um ein befriedigendes Ergebnis zu erhalten. Selbst die Auswahl von Suchbegriffen aus einer vorgegebenen Terminologie war für diesen Benutzerkreis schon deutlich zu aufwändig und nicht akzeptabel.

Es wurde deshalb entschieden, in der Gestaltung der Suchfunktion dem Wunsch der Mitarbeitenden nach einer einfachen und schnellen Eingabe von Suchbegriffen nachzukommen. Erhält ein Benutzer nicht das gewünschte Suchergebnis, ist er eher gewillt – so die Vermutung – ein wenig mehr Zeit in die Formulierung der Suchanfrage zu investieren als zu Beginn. Die Suchfunktion sollte deshalb vorsehen, den Benutzer bei der Reformulierung der Anfrage zu unterstützen, um eine Retrievalverbesserung zu erzielen.

Für die Reformulierung einer Anfrage wird ein Mechanismus benötigt, der es erlaubt, dem Benutzer neue Suchbegriffe vorzuschlagen, von denen ein besseres Ergebnis erwartet wird. Zu diesem Zweck werden Relationierungen zwischen Suchbegriffen benötigt, die es nicht gab. Da die Dokumentensammlung einer strikten Terminologie unterliegt (die jedoch nirgends durch ein Glossar oder ähnliches explizit gemacht war), ließen sich mit Hilfe eines Textanalyse-Systems (Engels/Lech 03) die relevantesten Begriffe aus den Dokumenten automatisch extrahieren und in einem assoziativen Begriffsnetz anordnen. Dort sind Begriffe, die semantisch assoziiert sind, über eine Relation verbunden, die mit einem Gewicht zwischen 0 und 1 für die Assoziationsstärke versehen ist. Unter Verwendung dieses Netzes wird dem Benutzer zusätzlich zum Retrieval-Ergebnis eine Liste von Begriffen angeboten, die mit seinen Suchbegriffen am stärksten assoziiert sind. Aus diesen kann er auswählen, um die ursprüngliche Suchanfrage zu modifizieren.

Prinzipiell könnte das gelernte Begriffsnetz herangezogen werden, um eine Suchanfrage auch automatisch um zusätzliche Begriffe zu erweitern. Da die Anzahl der Assoziationen eines Begriffs im Netz durchschnittlich sehr hoch ist, wäre eine solche automatische Anfrageerweiterung zu unspezifisch, selbst wenn man nur Begriffe mit hoher Assoziationsstärke hinzunimmt. Ein Benutzer wählt dagegen gezielt aus den ihm angebotenen Suchbegriffen entsprechend seines (nur ihm bekannten) Informationsbedarfs aus.

Auf der erwähnten Dokumentensammlung, die insgesamt etwa 1000 Seiten umfasste, wurden ungefähr 15'000 Begriffe extrahiert, denen jeweils als Maß für ihre Relevanz ein Wert zwischen 0 und 1 zugewiesen war. Davon wurden die 1'500 relevantesten Begriffe für das assoziative Netz herangezogen.

Eine nach der Produktivsetzung des Systems vorgenommene Evaluation hat ergeben, dass die Benutzer eine Suchanfrage im Durchschnitt mehr als einmal reformulieren – ein deutlicher Indikator, dass diese Funktionalität einen Mehrwert bietet und angenommen wurde.

## 4 Ausblick

Freitext-Retrieval nimmt alle Dokumente in ein Suchergebnis auf, in denen die in der Suchanfrage verwendeten Wörter vorkommen. Häufig lassen sich auf diese Weise recht gute Suchergebnisse erzielen, oft sind sie jedoch völlig unbefriedigend. Eine größere Unabhängigkeit von den in einem Text verwendeten Wörtern erhält man durch eine Indexierung der Dokumente.

Statistische Verfahren der automatischen Indexierung können eine Verbesserung bringen, weisen jedoch in der Regel ebenfalls nicht akzeptable Fehlerraten auf. Eine automatische Indexierung mittels textverstehender Systeme ist dagegen nur unter ganz speziellen Rahmenbedingungen praktikabel, wie z.B. eine abgegrenzte Domäne, Restriktion auf eine spezielle Textsorte, Einschränkung des Textverstehens auf die Extraktion bestimmter Arten von Fakten. Eine manuelle Indexierung durch den Benutzer kann praktikabel sein, wenn eine Terminologie zugrunde gelegt wird. Der vorliegende Beitrag skizziert zwei solche Anwendungsszenarien, wo Terminologien für die Beschreibung von Informationsinhalten herangezogen werden.

Eine grundlegende Änderung in der Mächtigkeit textverstehender Systeme wird sich in absehbarer Zeit kaum ergeben. In den nächsten Jahren werden deshalb im Zuge der unter dem Begriff „Semantic Web“ (Davies et al. 03) entstehenden Ansätze Terminologien und Ontologien eine zunehmend wichtige Rolle spielen – sowohl im Falle von intranet-basiertem Wissensmanagement, wie auch für die Organisation von Wissen auf dem Internet. Die in diesem Beitrag beschriebenen Ansätze sind zwei Beispiele dafür, wie sich die Organisation und die Wiedergewinnung von Informationen erheblich verbessern lässt, ohne dass die dazu nötigen Aufwände unrealistisch hoch wären.

## Danksagung

Ein Teil der in diesem Beitrag beschriebenen Projektergebnisse entstand im Rahmen des Projekts OnToKnowledge, das von der Europäischen Kommission (Projektnummer IST-1999-10132) und vom Schweizerischen Bundesamt für Bildung und Wissenschaft (Projektnummer BBW 99-0174) gefördert wurde.



## **5 Literaturverzeichnis**

- Abecker, A.; Bernardi, A.; Hinkelmann, K.; Kühn, O.; Sintek, M. (1998). Toward a Technology for Organizational Memories. In: *Intelligent Systems*, Vol.13, No.3, 1998, pp.40–48.
- Ackerman, M.S.; McDonald, D.; Lutters, W.; Marumatsu, J. (1999). Recommenders for Expertise Management. In: *Workshop on Recommender Systems: Algorithms and Evaluation at ACM SIGIR Conference 1999*.
- Becerra-Fernandez, I. (2000). Facilitating the online search of experts at NASA using expert seeker people-finder. In: U. Reimer (ed): *Proc. Third Int. Conf. on Practical Aspects of Knowledge Management (PAKM 2000)*, 2000.
- Berry, M. (2003). *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer-Verlag, 2003.
- Davies, J.; Fensel, D.; van Harmelen, F. (eds) (2003). *Towards the Semantic Web. Ontology-Driven Knowledge Management*. John Wiley, 2003.
- Engels, R.H.P; Lech, T.Ch. (2003). Generating Ontologies for the Semantic Web: *OntoBuilder*. In: J. Davies, D. Fensel, F. van Harmelen (eds): *Towards the Semantic Web. Ontology-Driven Knowledge Management*. John Wiley, 2003, pp.91–115.
- Gruber, T.R. (1993). A Translation Approach to Portable Ontologies. In: *Knowledge Acquisition*, Vol.5, No.2, 1993, pp.199–220.
- Hahn, U.; Reimer, U. (1999). Knowledge-Based Text Summarization – Saliency and Generalization Operators for Knowledge Base Abstraction. In: I. Mani, M. Maybury (eds): *Advances in Scalable Text Summarization*. Cambridge/MA: MIT Press, 1999, pp.215–232.
- Hahn, U.; Romacker, M. (2000). On ‘Deep’ Knowledge Extraction from Documents. In: *RIAO’2000 – Conference Proceedings of the 6th RIAO Conference. Content-Based Multimedia Information Access*. Vol.2. Paris, April 12–14, 2000.
- Hansen, M.T.; Nohria, N.; Tierney, T. (1999). What's Your Strategy for Managing Knowledge? In: *Harvard Business Review*, 1999, pp.106–116.
- Hovy, E.H.; Gerber, L.; Hermjakob, U.; Junk, M.; Lin, C.-Y. (2001). Question Answering in Webclopedia. In: *Proc. of the Ninth Text Retrieval Conference (TREC-9)*, 2001, pp.655–664.
- Lehner, F. (2000). *Organisational Memory*. Hanser-Verlag, 2000.
- Mani, I.; Maybury, M. (eds) (1999). *Advances in Scalable Text Summarization*. Cambridge/MA: MIT Press, 1999, pp.215–232.
- Maybury, M.; D'Amore, R.; House, D. (2000). Automatic expert finding. In: *International Journal of Technology Research Management*, Vol.43, No.6, 2000, pp.12–15.
- Reimer, U. (1992). Verfahren der automatischen Indexierung. Benötigtes Vorwissen und Ansätze zu seiner automatischen Akquisition: Ein Überblick. In: R. Kuhlen (ed): *Experimentelles und praktisches Information Retrieval: Festschrift für Gerhard Lustig*. Konstanz: Universitätsverlag, 1992, 171–194.
- Reimer, U.; Brockhausen, P.; Lau, T.; Reich, J.R. (2003). *Ontology-Based Knowledge Management at Work: The Swiss Life Case Studies*. In: J. Davies, D. Fensel, F. van

- Harmelen (eds): Towards the Semantic Web. Ontology-Driven Knowledge Management. John Wiley, 2003, pp.197–218.
- Reimer, U. (2004). Wissensbasierte Verfahren der Organisation und Vermittlung von Information. In: R. Kuhlen, Th. Seeger, D. Strauch (eds): Grundlagen der praktischen Information und Dokumentation. Band 1. K.G. Saur, 2004.
- Sure, Y.; Maedche, A.; Staab, S. (2000). Leveraging corporate skill knowledge – from ProPer to OntoProper. In: U. Reimer (ed): Proc. Third Int. Conf. on Practical Aspects of Knowledge Management (PAKM 2000), 2000.
- van Heijst, G.; van der Spek, R.; Kruizinga, E. (1996). Organizing Corporate Memories. In: Proc. of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Alberta, Canada, 1996.