SemanticMiner^{TM}: Ein integratives Ontologie-basiertes Knowledge Retrieval System *

Eddie Mönch
Ontoprise GmbH, Haid-und-Neu-Str. 7, D-76131 Karlsruhe
moench@ontoprise.de

Abstract: Oft stellt sich bei der Analyse von Wissensprozessen in Unternehmen heraus, dass der einfache Zugriff auf das vorhandene Unternehmenswissen in Dokumenten nicht möglich ist. Für den Zugriff auf Dokumenten- und Datenbestände des Unternehmens nehmen die Technologien des *Information Retrieval (IR)* eine zentrale Rolle ein. Im Folgenden beschreiben wir die Theorie des SemnaticMiner^{TM}-Systems, das heißt Methoden und Technologien sowie weiterführende Ansätze, um mithilfe semantischer Technologien aus dem Information Retrieval ein *Knowledge Retrieval (KR)* zu erreichen.

1 Einführung in Ontologie-basiertes Wissensmanagement

Bereits Aristoteles versuchte in seiner Kategorienlehre, die Dinge der Welt nach bestimmten Kriterien zu untersuchen und zu ordnen. Daraus entstand über Jahrhunderte eine philosophische Subwissenschaft namens Ontologie. Diese vergleichsweise neue Bezeichnung, die sich aus dem Griechischen zusammengesetzt — „ontos“ für Sein und „logos“ für Wort [Sow00] — wird benutzt, um die Lehre vom Sein zu unterscheiden von der Lehre des Seienden in den Naturwissenschaften.

Die Informatik entlehnte den Begriff der Ontologie zum Zwecke der Repräsentation und Nutzung von Wissen. Seit Anfang der neunziger Jahre wurden Ontologien zu einem beliebten Forschungsthema in Teilgebieten der Künstlichen-Instelligenz-Forschung. In letzter Zeit breitet sich die Idee der Ontologie auf immer mehr Bereiche aus, wie Intelligent Information Integration, Cooperative Information Systems, *Information Retrieval*, Electronic Commerce und Knowledge Management (für weitere Beispiele sei auf [Sow00] verwiesen). Der Grund für die stetig wachsende Popularität von Ontologien, liegt größtenteils an dem, was sie versprechen: Ein geteiltes und gemeinsames Verstehen einer Domäne, das zwischen Personen und Anwendungssystemen kommuniziert werden kann (vgl. [Fen01]).

Bedeutung: Ontologien werden entwickelt um eine maschinen-verarbeitbare Semantik an Informationsressourcen, die zwischen verschiedenen Agenten (Software und Menschen) kommuniziert werden kann, bereitzustellen.

* Ausführliche Version unter http://www.ontoprise.de/documents/SemanticMinerKR.pdf

Definition und Eigenschaften. Die am häufigsten zitierte Definition für Ontologie ist die von Gruber: „Eine Ontologie ist eine formale, explizite Spezifikation einer gemeinsamen Konzeptualisierung“ [Gru93].

Durch eine explizite Spezifikation der Entitäten (Konzepte), die mit anderen Entitäten über Axiome (Relationen) verknüpft, oder mit Attributen detaillierter beschrieben werden entsteht daraus eine Ontologie. Üblicherweise sind Ontologien in Taxonomien mit mehrfacher Vererbung und disjunkten Unterkategorien organisiert. Neben dieser Kategorisierung beschreiben sie für einen Wissensbereich ebenfalls Regeln, die die Konzepte durch Constraints oder Inferenzregeln in Beziehung setzten. Diese werden typischerweise in logischen Formalismen repräsentiert, die auf der Prädikatenlogik basieren.

F-Logic. Für das SemanticMiner-System verwenden wir die Sprache *Frame-Logic* (F-Logic). F-Logic entspricht syntaktisch gesehen einer Obermenge der Prädikatenlogik erster Stufe (*FOL, first order logic*), wobei die Ausdrucksmächtigkeit beider Sprachen allerdings äquivalent ist. F-Logic ist eine logik- und objektorientierte Sprache, die 1995 von Kifer et al. [KLW95] entwickelt wurde. Sie verbindet die Ausdrucksstärke von Normallogik (Horn-Logik mit Negationen) mit den Datenmodellierungsmöglichkeiten des objektorientierten Ansatzes. Da die grundlegenden Prinzipien der Vererbung, Kapselung, Klassenbildung, Polymorphie und Typüberprüfung durch die Ausdrucksstärke und die Inferenzmöglichkeiten von Logik ergänzt werden, ist sie besonders für die Modellierung von Ontologieen geeignet.

Das Allwissenden-Paradigma. Mit der Verwendung einer Ontologie akzeptiert man automatisch das „Allwissenden“-Paradigma, das einem traditionellen Ansatz der Kognition in sozialen Systemen entstammt. Wissen wird dabei in einer einzigen, von allen geteilten kohärenten Struktur repräsentiert und organisiert, völlig unabhängig von wem, wie, wo und warum dieses Wissen ursprünglich geschaffen wurde. Der heute aufstrebende Ansatz der „Verteilten Intelligenz“ basiert hingegen auf der Annahme, dass Wissen immer und unteilbar mit verschiedenen sogenannten Kontexten verknüpft ist, wie beispielsweise Individuen, Gruppen, Zeiträumen und Orten und daher nicht generell zentral organisiert werden kann: Wissen ist demnach immer kontextspezifisch [NSB00]. Zu erwähnen ist auch, dass sich die spätere Nutzergruppe des angestrebten wissensbasierten Systems auf die Ontologie geeinigt haben muss [Gru95]. Durch diese Formalisierung wird jedoch Mehrdeutigkeit vermieden.

Weitere Ansätze existieren um Wissensmodelle aufzubauen. Eine ebenfalls verbreitete Methode ist die Verwendung von TopicMaps für die Einordnung und Kategorisierung von Begriffen. Hierbei werden vorhandene Themen (Topics) miteinander verbunden, ein semantisches Netz entsteht. TopicMaps eignen sich insbesondere zur Navigation vorhandener Begrifflichkeiten. Ontologien stellen zusätzlich zur Navigationsunterstützung mächtigere Modellierungsmöglichkeiten zur Verfügung, welche zusätzliche Funktionen des Wissensmodells ermöglichen [SM01].

Im Gegensatz zu allen anderen Technologien bestehen weitere Zusatznutzen von Onto-

logien darin, dass sie Ableitungen erlauben und Auswertungen der oben beschriebenen regelbasierten Zusammenhänge mittels einer Inferenzmaschine (z.B. OntoBroker^{TM}) erlauben. Implizites Wissen wird dadurch ebenfalls abgefragt und dargestellt — explizit gemacht.

2 Information Retrieval

Für den Begriff bzw. das Gebiet des Information Retrieval (IR) gibt es keine allgemein akzeptierte Definition oder Abgrenzung. Historisch gesehen wurde IR zum besseren (Wieder)auffinden von wissenschaftlicher Literatur entwickelt. Auch wenn dieses Gebiet nach wie vor einer der Schwerpunkte des IR ist, haben sich sowohl der Bereich der Objekte, mit denen IR umgeht, als auch die Aufgabenstellung erweitert. Eine Beschreibung gibt die Fachgruppe Information Retrieval der Gesellschaft für Informatik [Fuh96]:

„Im Information Retrieval werden Informationssysteme in Bezug auf ihre Rolle im Prozess des Wissenstransfers vom menschlichen Wissensproduzenten zum Informationsnachfragenden betrachtet.“ Ziel des IR ist es also, gespeicherte Daten (Texte, strukturierte Daten, Bilder, Fakten u.a.) so aufzubereiten und anzubieten, dass sie bei einem konkreten Informationsbedarf mit problemgerechten Suchstrategien möglichst präzise und vollständig herausgesucht werden können.

2.1 Qualitätsbewertung von IR-Systemen: Recall und Precision

Die am häufigsten verwendeten Maße zur Beurteilung der Güte eines IR-Systems sind *Recall* und *Precision*. Durch diese beiden Maße wird die Suche mit einem IR-System aufgrund des gelieferten Retrievalergebnisses bewertet. Grundlage bildet der Begriff der Relevanz eines Dokuments.

Eine Reihe von verschiedenen Definitionen des Begriffs Relevanz sind beispilsweise in [Kai93] zu finden. Wir verwenden die Definition *Relevanz* nach [CLvRC98]:

Definition 2.1 (Relevanz) *Wenn der Benutzer ein Dokument zu einer haben will, dann ist dieses relevant zu dieser Anfrage.*

Nun können die beiden Maße Recall und Precision definiert werden [BYRN99]:

Definition 2.2 (Recall) *Recall stellt das Maß für die Vollständigkeit des Retrievalergebnisses dar und ist definiert als das Verhältnis zwischen gefundenen, relevanten Dokumenten und der Gesamtzahl der im Dokumentenbestand vorhandenen relevanten Dokumente.*

Genauer gilt: Gegeben sei ein Informationsbedarf I und eine Anfrage q des Benutzers. Dann berechnet sich der Recall durch

$$req(q, I) = \frac{|\mathcal{R}(q, I)|}{|\mathcal{R}(I)|}, \tag{1}$$

wobei $|\mathcal{R}(I)|$ die Anzahl aller relevanten Dokumente zum Informationsbedarf I und $|\mathcal{R}(q, I)|$ die Anzahl der mit der Anfrage q gefundenen, zum Informationsbedarf I relevanten Dokumente bezeichnet (vgl. Abbildung 1).

Der Wertebereich des Recalls geht von 0 bis 1. Ein Recall von 0 wird für das schlechteste Ergebnis, 1 für das bestmögliche vergeben.

Definition 2.3 (Precision) *Precision dient zum Messen der Genauigkeit des Retrievalergebnisses und als Indikator für die Fähigkeit eines IR-Systems, nicht relevante Dokumente nicht auszugeben. Precision ist definiert als das Verhältnis der gefundenen relevanten Dokumente zur Zahl aller Dokumente.*

Genauer gilt: Gegeben sei ein Informationsbedarf I und eine Anfrage q des Benutzers. Dann berechnet sich Precision durch

$$pres(q, I) = \frac{|\mathcal{R}(q, I)|}{|\mathcal{E}(q)|}, \tag{2}$$

wobei $|\mathcal{R}(q, I)|$ die Anzahl der mit der Anfrage q gefundenen, zum Informationsbedarf I relevanten Dokumente und $|\mathcal{E}(q)|$ die Anzahl aller mit Anfrage q gefundenen Dokumente bezeichnet (vgl. Abbildung 1).

Der Wertebereich von Precision geht ebenfalls von 0 bis 1. Auch bei Precision wird versucht, den Wert zu maximieren.

Sinnvoll ist nur die Betrachtung beider Maße, da Recall die Zahl der irrelevanten, ausgegebenen Dokumente unberücksichtigt lässt und leicht auf das Maximum von 1 gesetzt werden kann, indem alle im Dokumentenbestand vorhandenen Dokumente ausgegeben werden. In diesem Fall wäre dann allerdings der Precisionwert sehr niedrig. Die alleinige Betrachtung von Precision wiederum würde nichts über die Vollständigkeit des Retrievalergebnisses aussagen. Precision allein könnte dadurch maximiert werden, dass nur sehr wenig Dokumente ausgegeben werden.

Bei Suchen mit einem hohen Anspruch auf Vollständigkeit des Suchergebnisses wir ein hoher Recall angestrebt, so dass wir innerhalb des SemanticMiner Systems ein größeres Augenmerk auf dieses Maß legen (siehe hierzu Kapitel 3.1).

3 Knowledge Retrieval — Semantisches Information Retrieval

Ein Indikator für die Retrievalqualität der derzeitigen Ad-hoc-IR-Systeme[1] stellen die Ergebnisse der jährlich stattfindenden TREC-Veranstaltungen dar. TREC bezeichnet eine Veranstaltung, bei der Softwareimplementierungen von derzeitigen Algorithemn im IR auf ihre Qualität getestet werden. In [Har00] werden die Ergebnisse der an TREC teilgenommenen Ad-hoc-IR-Systeme über die letzten Jahre verglichen. Es zeigt sich, dass

[1] Unter Ad-hoc-Suche versteht sich die vollautomatische Suche.

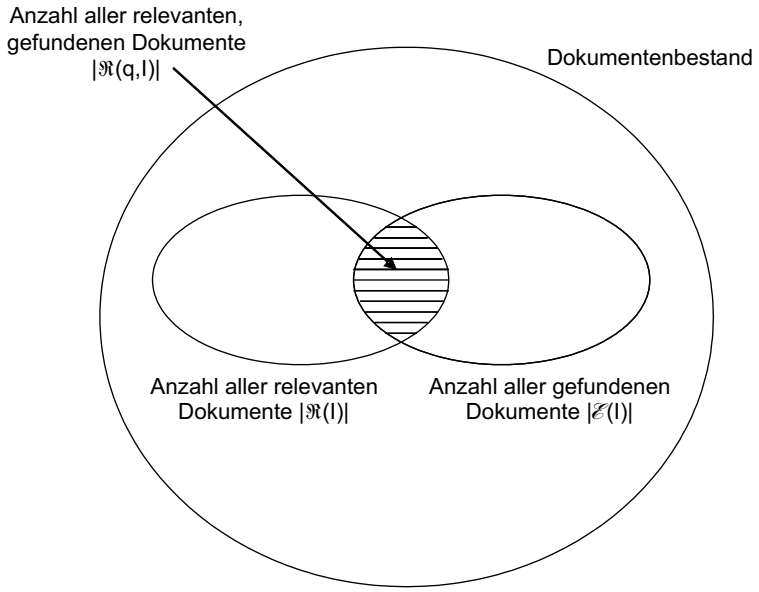

Abbildung 1: Recall und Precision für einen gegebenen Beispiel-Informationsbedarf

seit 1996 bei den Ad-hoc-IR-Systemen eine Stagnation hinsichtlich der Retrievalqualität (Recall/Precision) zu verzeichnen ist. Daraus läßt sich folgern, dass nach dem derzeitigen Kenntnisstand der Wissensschaft die Entwicklung von Retrievalalgorithmen und Indexierungsalgorithmen ausgereizt ist.

3.1 Abhängigkeit der Retrievalqualität von der Anfrage

Die Qualität eines Ad-hocSuchdienstes im Sinne von Recall und Precision ist sehr abhängig von der aktuellen Anfrage. Diese Eigenschaft wurde bei Ad-hoc-IR-Systemen in [Har00] praktisch bestätigt. Es wurden unterschiedliche Ad-hoc-IR-Systeme bei TREC pro Anfrage verglichen und es zeigte sich, dass ein Ad-hoc-IR-System bei der einen Anfrage eine sehr hohe Qualität im Vergleich zu den anderen Ad-hoc-IR-Systemen und bei anderen Anfragen nur eine sehr schlechte Qualität im Vergleich zu den anderen Ad-hoc-IR-Systemen besitzen kann.

3.2 Änderung der Anfrage

Das Ziel der Änderung der Anfrage besteht in der Adaptivität des Wortschatzes des Benutzers an das IR-System. Es ist eine sehr verbreitete Methode. Es existieren sehr viele

automatische Anfragemodifikationsalgorithmen in der Literatur, z.B. [BMS98].

Definition 3.1 (Anfragemodifikation) *Anfragemodifikation entspricht der automatischen Änderung einer Anfrage aufgrund von Zusatzwissen (Thesaurus, Relevanz Feedback, Statistiken, usw.) mit dem Ziel, bessere Retrievalergebnisse zu erzielen. Es besteht dabei die Gefahr des* Anfrageabtriebs *(engl.* Query Drift*), also die Gefahr, dass die erweiterte Anfrage nicht mehr den ursprünglichen Informationsbedarf widerspiegelt.*

Teilweise werden für den Begriff Anfragemodifikation auch die Begriffe *Anfrageerweiterung* oder *Anfragereformulation* verwendet.

Unser Verfahren unterscheidet sich von bekannten Anfragemodifikationsalgorithmen dahingehend, dass die Anfrage von der Dokumentenmenge völlig abgekoppelt und die Erweiterung wie in Kapitel 1 beschrieben allgemein gültigen Status besitzt. Ebenfalls wird dadurch der Gefahr des Anfrageabtriebs entgegengesteuert.

3.3 Query-Expansion

Die Abhängigkeit der Retrievalqualität von der Anfrage unterstützt unsere Motivation im SemanticMiner^{TM}-System den Fokus auf die Anfrage an ein Ad-hoc-Suchdienst zu legen. Der dem System zugrunde liegende *Query-Expansion*-Ansatz kann den Anfragen an das Ad-hoc-IR-System automatisch ontologisches Wissen hinzufügen und so die Qualität der Antworten verbessern. Das führt zu einer Verbesserung des Recall-Wertes, da mehr relevante Dokumente durch die qualitative Erhöhung der Suchterme gefunden werden. Über den Precision-Wert kann keine allgemeine Aussage getroffen werden, da die Anzahl der mit der Anfrage q gefundenen, zum Informationsbedarf I relevanten Dokumente — also $|\mathcal{R}(q, I)|$ — als auch die Anzahl aller mit Anfrage q gefundenen Dokumente, also $|\mathcal{E}(q)|$ steigt.

Jedoch schaut sich typischerweise der Suchende nur die ersten 10 bis 20 Dokumente eines Suchergebnisses an. Durch die Ranking-Funktion im SemanticMiner^{TM}-System führt dies, in Kombination mit der Query-Expansion durch ontologisches Wissen, zu einer wesentlichen Steigerung des „subjektiven“, also für den Suchenden relevanten (die ersten 10-20 Dokumente), Precision-Wertes, da die Dokumente mit hoher Termübereinstimmung *aller* Anfrageterme das höchste Ranking erfahren.

Aus [Har00] kann man ebenfalls folgern, dass der Wechsel eines Ad-hoc-Suchdienstes während einer Suche bei gleicher Anfrage durchaus Sinn macht. Die zugrunde liegenden Ad-hoc-IR-Systeme des SemanticMiner^{TM}-Systems sind für das System transparent und können beliebig ausgetauscht oder ergänzt werden.

3.4 Semantik der Anfrage

Ein weiterer Nachteil allgemeiner IR-Ansätze ist, dass eine reine syntaktische Suche nach Begriffen stattfindet, ohne dabei die Bedeutung der Wörter innerhalb der einzelnen Dokumente zu berücksichtigen. Dies führt bei der Suche zu einer hohen Zahl von Treffern, bei denen auch Dokumente gefunden werden, in denen der Begriff in anderer Bedeutung verwendet wird. Eine Suche nach ähnlichen Begriffen, bzw. Eingrenzungen oder Verallgemeinerungen kann von diesen statistischen Ansätzen nicht unterstützt werden.

Während der letzten 30 Jahre fand eine anhaltende Diskussion darüber statt, ob zur Unterstützung natürlicher Sprachverarbeitung (Natural Language Processing, NLP) auf syntaktische oder semantische Technologien fokussiert werden solle. In beiden Lagern wurden Ansätze diskutiert und vorangetrieben. Immer deutlicher hat sich gezeigt, dass beide Technologien und insbesondere das Zusammenspiel zwischen statistischen Verfahren und semantischer Modellierung die wichtigsten Ansatzpunkte für die Weiterentwicklung der natürlichen Sprachverarbeitung darstellen.

3.5 Integration und Auswertung strukturierter Daten

Durch die Kombination einer Suchanfrage mit (semi)strukturierten Daten (Listen, Datenbanken, Metadaten) und logischen Regelzusammenhängen wird die Mächtigkeit der vorgestellten Ansätze in Abschnitt 3.4 weiter erhöht. Ziel ist es dabei, als Ergebnisliste keine Verweise auf Dokumente, die wiederum den gesuchten Inhalt enthalten, zu liefern, sondern tatsächliche Informationen aus Dokumenten zu lösen und als konkrete Antworten zu präsentieren.

Dies geschieht durch die Bildung von *Kollokationen*, wobei eine Kollokation eine Wortgruppe darstellt, mit der üblicherweise ein Grundbegriff, ein Gegenstand oder eine Handlung bezeichnet wird. Kollokationen wie „maschinelle Übersetzung“ oder „Anwendung schließen“ werden in der Terminologie als eigenständige Termini angesehen. Durch die Bildung von Korrelationslisten aus einer Datenbank oder mittels des Excel2F-Logic-Exports, werden die Kollokationen dem SMS bekannt gemacht.

Dadurch kann man auch aus völlig unstrukturierten Informationsquellen und Textdokumenten zum Beispiel geeignete Mitarbeiter in Unternehmen identifizieren, indem man zu dem gesuchten Begriff eine Kollokationsabfrage über die Mitarbeiter-Datenbank stellt. Ebenfalls können somit Wissenslücken in Unternehmen aufgedeckt oder Wettbewerberübersichten generiert werden.

3.6 Deduktion

Wie in Abschnitt 1 bestehen weitere Zusatznutzen von Ontologien darin, dass sie Ableitungen erlauben und Auswertungen der beschriebenen regelbasierten Zusammenhänge

mittels der Inferenzmaschine des OntoBrokers erlauben. Implizites Wissen wird dadurch ebenfalls abgefragt und dargestellt - explizit gemacht. Das heißt, dass alle Informationen, die durch Regeln ausgewertet wurden, also nur implizit vorlagen, im SemanticMiner™-System als explizite Informationen dargestellt werden. Der End-Nutzer des Systems ist somit nicht in der Lage zu unterscheiden, ob die Information, die ihm präsentiert wird, explizit vorhanden war, oder durch *Deduktion* anhand von Ableitungsregeln (*„inferencing rules“*) ermittelt wurde.

Literatur

[BMS98] C. Buckley, M.Mandra, and A. Singhal. Improving Automatic Query Expansion. In *21st ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 206–214, 1998.

[BYRN99] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, Addison-Wesley, 1999.

[CLvRC98] F. Crestani, M. Lalmas, C.J. van Rijsbergen, and I. Campbell. Is this Document Relevant? ... Probably - A Survey of Probablistic Models in Onformation Retrieval. *ACM Computing Surveys*, 30:528–552, December 1998.

[Fen01] D. Fensel. *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, Berlin, 2001.

[Fuh96] N. Fuhr. Ziele und Aufgaben der Fachgruppe Information Retrieval, January 1996. http://ls6-www.informatik.uni-dortmund.de/ir/fgir/mitgliedschaft/brochure2.html.

[Gru93] T.R. Gruber. A translation approach to portable ontology specifications. *Knowledge Aquisition*, (5):199–220, 1993.

[Gru95] T.R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, (43):907–928, 1995.

[Har00] D. Harman. What We Have Learned, and not learned, from TREC. In *BCS-IRSG: 22nd Annual Colloquium on IR Research*, pages 2–20, April 2000. http://irsg.eu.org/irsg2000online/papers/harman.htm.

[Kai93] A. Kaiser. *Computer-unterstütztes Indexieren in Intelligenten Information-Retrieval Systemen. Ein Relevanz Feedback orientierter Ansatz zur Informationserschliessung in unformatierten Datenbanken*. PhD thesis, Wirtschaftsuniversität Wien, 1993.

[KLW95] M. Kifer, G. Lausen, and J. Wu. Logical Foundations of Object-Oriented and Frame-Based Languages. *Journal of the ACM*, 42:741–843, 1995.

[NSB00] S. Neumann, L. Schuurmans, and M. Bonifacio. Verteilte Systeme im Wissensmanagement. *Information Management und Consulting*, (15):75–82, 2000.

[SM01] S. Staab and A. Mädche. Knowledge Portals: Ontologies at Work. *AI Magazine*, 2(21), 2001.

[Sow00] J. Sowa. Knowledge Representation: Logical, Philosophical, and Computational Foundations, 2000.