Ontologiebasierte Wissensextraktion

Klaus Unterstein

13. November 2001

Inhaltsverzeichnis

1 Einleitung **2**

2 Begriffsklärung **2**
2.1 Ontologie(n) . . . 2
2.1.1 Definition . . . 2
2.1.2 Motivation, Zweck und Einsatzmöglichkeiten . . . 2
2.1.3 Beschreibung . . . 3
2.1.4 Bewertung . . . 4
2.2 (Wissens-) Extraktion . . . 4
2.3 Ontologiebasierte Wissensextraktion . . . 5

3 Ontologiebasierte Wissensextraktion **5**
3.1 Klassifizierungsmöglichkeiten für ontologiebasierteWissensextraktion . . . 5
3.1.1 Autonomiegrad . . . 6
3.1.2 Strukturiertheit der Eingabedaten . . . 6
3.1.3 Verfahren allgemein . . . 6
3.1.4 Verfahren im Detail . . . 7
3.1.5 Extraktion on-demand vs. Vorab-Extraktion . . . 8
3.1.6 Methoden . . . 8
3.2 Methoden . . . 8
3.3 Vor- und Nachteile im Vergleich . . . 9
3.4 Bewertung der Ansätze . . . 10
3.5 Praxis/Trends . . . 10

4 Zusammenhänge zu den anderen Beiträgen **10**

5 Schlußwort **11**

6 Literaturangaben **12**

1 Einleitung

Ausarbeitung zum Referat "Ontolologiebasierte Wissensextraktion", welches von Klaus Unterstein am 23.10.2001 im Rahmen der "PG-402: Wissensmanagement" gehalten wurde. Es wurde versucht, einen allgemeinen Überblick über das Prinzip der ontologiebasierten Wissensextraktion zu geben. Nach der einführenden Begriffsklärung wird anschliessend auf den ontologiebasierten Wissensextraktionsprozeß eingegangen. Es folgt eine Beschreibung der verwendbaren Methoden, welche unterschiedliche Vor- und Nachteile besitzen. Diese werden bewertet. Nachfolgend gibt es noch einen kurzen Ausblick auf aktuelle und zukünftige Entwicklungen. Anschliessend folgt eine Integration der Thematik zu anderen Bereichen, wonach das Schlußwort folgt.

2 Begriffsklärung

Um über die Thematik problemlos sprechen zu können und Mißverständnisse vorzubeugen, werden die verwendeten Begriffe vorab erklärt/geklärt.

2.1 Ontologie(n)

2.1.1 Definition

Es gibt viele verschiedene Definitionsmöglichkeiten für den Begriff "Ontologie". Gewählt wurde die Definition von Gruber, die kurz und prägnant die wichtigsten Elemente beschreibt. Nach Gruber wurde im Jahre 1993 der Begriff Ontologie definiert als: "An ontology is a formal, explicit specification of a shared conceptualization." Frei übersetzt beschreibt eine Ontologie explizit eine formales, verteiltes semantisches Modell eines bestimmten, uns interessierenden Bereichs. Ein "semantisches Modell" (conceptualization) bezieht sich hier auf ein abstraktes Modell eines Phänomens in der Welt, welche die relevanten Konzepte dieses Phänomens identifiziert. "Explizit" (explicit) heißt, daß der Typ dieser Konzepte, die benutzt werden, und die Zwänge bei ihrer Benutzung explizit definiert sind. "Verteilt" (shared) soll heißen, daß die Ontologie allgemein anerkanntes Wissen beinhaltet, das nicht auf ein Individuum eingeschränkt ist, sondern von Gruppen akzeptiert wird. "Formal" (formal) heißt, daß die Ontologie durch Maschinen verwendbar sein soll. Verschiedene Stufen der Formalität sind möglich.

2.1.2 Motivation, Zweck und Einsatzmöglichkeiten

Die Anzahl der gespeicherten Informationsquellen und die Anzahl der verschiedenen Formate dieser Informationen wächst immer weiter an. Dies erschwert das Finden, den Zugriff und auch die Zusammenfassung von Informationen zu einem bestimmten Themenbereich. Weiterhin besteht eine grosse Lücke zwischen dem semantischen Modell von Informationen und der gespeicherten Form. Die Verwendung von Ontologien soll das Finden und den Zugriff auf relevante Informationen innerhalb der unzähligen Informationsquellen erleichtern. Dies

setzt eine wohldefinierte Semantik der semantischen Modelle und die Festlegung bestimmter Begriffe zwingend voraus, welches wiederum die Eindeutigkeit der Begriffe zur Folge hat. Somit dienen Ontologien auch als Kommunikationshilfe zwischen Mensch und Maschine, die den Austausch von Semantik **und** Syntax unterstützen kann. Hintergrund-Wissen (oder implizit vorhandenes Wissen) kann mittels Ontologien zur Verfügung gestellt werden, um zum Beispiel die Leistung von Informations-Extraktions-Systemen zu erhöhen. Besonders wichtig ist der Punkt der Eindeutigkeit. Nach Ogden Richards (1923) gibt es für jedes Sache ein oder mehrere Symbole. Dieses Symbol repräsentiert eine Sache. Das Symbol erweckt den Begriff, welches sich dann auf die Sache bezieht. Dies wird durch das **semiotische Dreieck** beschrieben. Das Problem ist, daß sich ein Symbol auf verschiedene Sachen beziehen kann. Als Beispiel wäre das Symbol der Schriftzug "Jaguar", der sich sowohl auf ein Auto (die Automarke Jaguar), als auch auf das besagte Tier beziehen kann. Der Kontext kann diese Mehrdeutigkeit auflösen, wodurch der Begriff, den man darunter versteht, eindeutig festgelegt wird. Dies ist ein Vorteil von Ontologien. Durch die Vorgabe der Ontologie (z.B. eine Ontologie, die den Bereich "Tierwelt" abdeckt), wird diese Mehrdeutigkeit vermieden. Außer der Eliminierung von Mehrdeutigkeiten soll eine Ontologie domain-relevante Konzepte (und nach Möglichkeit auch nur domain-relevante Konzepte), bestehende Beziehungen zwischen den Konzepten und Axiome sowohl für die Konzepte und den Beziehungen beschreiben. Somit eignen sich Ontologien zum Einsatz in Informations-Extraktions-Systemen zur Integration von Informationen aus heterogenen Quellen und somit auch zur Generierung verschiedener Ziel-Strukturen, die die Informationen speichern soll (Informationsspeicherung) und zur Extraktion weiterer Fakten durch "Schliessen" (Inferenz), indem Hintergrundwissen herangezogen wird.

2.1.3 Beschreibung

Nachdem jetzt allgemein über Ontologien und ihre angestrebte Funktion berichtet wurde, wird nun eine allgemeine Beschreibungsform von Ontologien angeschnitten.

Eine Ontologie wird beschrieben durch:

- Eine Menge von Zeichenketten, die die lexikalen Einträge L für Konzepte und Relationen beschreiben
- Eine Menge von Konzepten C
- Eine Taxonomie von Konzepten (bei einigen Definitionen Heterarchie[1]) H_C
- Ein Satz an nicht-taxonomischen Relationen R (beschrieben durch ihre Domain)
- Relationen F und G, die Konzepte und Relationen verknüpfen

[1] Unter Heterarchie versteht man eine Hierarchie mit der Möglichkeit der multiplen Vererbung innerhalb dieser Hierarchie.

- Die Taxonomie der Relationen (bzw. Heterarchie H_R) (optional)
- Axiome A, die weitere Einschränkungen (Constraints) der Ontologie beschreiben und es erlauben, implizite Fakten explizit zu machen (optional)

Die letzten Punkte sind optional. So gibt es Fälle, in denen Axiome nicht verwendet wurden bzw. es keine Taxonomie der Relationen gab.

2.1.4 Bewertung

Der Einsatz von Ontologien hat Vor- und Nachteile. Vorteilhaft ist, daß es ein einfaches Prinzip ist, welches nur den relevanten Bereich betrachtet (Fokussierung) und irrelevante Elemente weggelassen werden, ähnlich dem Klassenprinzip bei objekt-orientierten Programmiersprachen. Weiterhin bringt die Nutzung von Semantik und Hintergrundwissen, die in die Ontologie integriert werden, weitere Vorteile[2]. Ein weiterer Punkt wäre auch die Möglichkeit der dynamischen Entwicklung einer Ontologie, welche sich mit der Zeit 'automatisch' auf den relevanten Zielbereich fokussiert[3]. Ein grosser Vorteil sind semi-automatische Ansätze, die die Nachteile der manuellen Erstellung/Modellierung und den daraus resultierenden Zeitaufwand und Kostenfaktor reduzieren. Ein anderer Nachteil ist noch das Problem "Vollständigkeit vs. Minimalität". Zum einen möchte man eine möglichst kleine Ontologie haben, die alle relevanten Aspekte des Bereichs abdecken soll, ohne sich zu stark einzuschränken. Zum anderen möchte man eine möglichst umfassende Ontologie für den Bereich entwickeln, welches wiederum zu einer Ausartung dieser Ontologie führen kann. Hier ist es schwierig, eine optimale Mitte zu finden, da man nicht weiß, ob man alle und nur die relevanten Punkte abgedeckt hat. Nachträgliche Änderungen sind möglich, führen aber zu einem Mehraufwand und reduziert die Wahrscheinlichkeit zur Wiederverwendung.

2.2 (Wissens-) Extraktion

Eine mögliche Definition ist:

> Der Prozeß, in dem Information automatisch aus textuellen Dokumenten in eine zur Speicherung in Datenbanken geeignete Form generiert wird.
> *[J. M. Lawler, 1998]*[4]

Es geht um die Extraktion von Informationen. Dies kann man noch weiter klassifizieren mittels verschiedener Kriterien. Mögliche Kriterien wären Quelle, Datenformate, Extraktionsmethoden usw. So gibt es unterschiedliche Quellen, aus denen Informationen extrahiert werden können. Es stehen verschiedene Medien

[2] Beispiel der Mehrdeutigkeit bzgl. dem Wort: Jaguar

[3] Beispiel einer dynamischen Entwicklung während des Betriebs ist die Suchmaschine, die in [DLOE] beschrieben wird

[4] "The process by which information in a form suitable for entry into a database is generated automatically from textual documents." [LAWL]

zur Verfügung, wie z.B. eine Datenbank, das WorldWideWeb, eMails, usw. Weiterhin spielen noch die unterschiedlichen Datenformate eine Rolle. So kann innerhalb einer Quelle eine Anzahl verschiedener Datenformate verwendet werden. So findet man im WWW verschiedene Formate wie HTML, XML, Textdateien, und viele weitere. Außerdem kommen verschiedene Extraktionsmethoden zur Anwendung, deren Anwendung teilweise vom Datenformat und der gewählten Quelle abhängig ist.

2.3 Ontologiebasierte Wissensextraktion

Was ist ontologiebasierte Wissensextraktion? Unter ontologiebasierter Wissensextraktion versteht man die Verwendung von Ontologien innerhalb des Informationsextraktionsprozesses. Allgemein wird mittels Ontologien Informationen aus einer Quelle extrahiert und anschliessend in ein bestehendes System integriert. Die Verwendung von Ontologien kann an unterschiedlichen Stellen innerhalb des Gesamtprozesses stattfinden[5]. Auch wenn beide viele gleiche/ähnliche Methoden verwenden, bringt die Verwendung einer Ontologie gewisse Vorteile gegenüber den allgemeinen Wissensextraktionsverfahren. Die Vorteile, die die Verwendung von Ontologien bringt, wurden bereits erwähnt.

3 Ontologiebasierte Wissensextraktion

Die Verwendung von Ontologien im Extraktionsprozeß nennt man "ontologiebasierte Wissensextraktion". Abhängig vom Anwendungsbereich wird eine Ontologie gewählt. Es können auch unterschiedliche Ontologien für den gleichen Andwendungsbereich verwendet werden, was zu unterschiedlichen Resultaten führen kann. Somit ist die Extraktion flexibel, da man durch den Austausch der Ontologie auch aus anderen Bereichen extrahieren kann. Der Extraktionsprozeß liefert Informationen für die semantische Annotation der Texte, die als Nebenprodukt die Klassifikation der Daten liefert. Durch diese Klassifikation können die Daten direkt integriert werden (z.B. Integration der Daten in eine Datenbank). Weiterhin helfen Ontologien bei der Portierung von Daten in und aus verschiedenen Formaten (aufgrund der expliziten Festlegung der Ontologie). So ist es möglich, aus verschiedenen Quellen, die unterschiedlich strukturiert sein können, die Informationen in eine Datenbank zu integrieren.

3.1 Klassifizierungsmöglichkeiten für ontologiebasierteWissensextraktion

Das Verfahren der ontologiebasierten Wissensextraktion ist recht allgemein, daher ist eine weitere Klassifizierung nützlich.

[5] Auf diesen Punkt wird später genauer eingegangen.

3.1.1 Autonomiegrad

Eine Möglichkeit wäre der Autonomiegrad. Anfangs wurde innerhalb des Prozesses vieles manuell vollzogen, was zwar durchführbar aber zu zeit- und kostenintensiv ist. Optimal wäre ein vollkommen automatisches Verfahren, was aber derzeit nicht existiert, da die resultierenden Ergebnisse noch starke Mängel haben. Daher konzentrieren sich die aktuellen Entwicklungen auf semi-automatische Verfahren, in denen Vieles automatisiert wurde, es aber noch mehrere Eingriffsmöglichkeiten und Zwänge gibt. So sind bestimmte Ergebnisse zu bewerten, ob ein automatisch erstelltes Ergebnis gut ist oder wenn bei einer automatischen Konzept-Heterarchie-Erstellung Konzepte auftreten, die nicht integriert werden konnten, so muß manuell das Konzept integriert bzw. korrigiert werden.

3.1.2 Strukturiertheit der Eingabedaten

Als Kriterium kann die Strukturiertheit von Eingabedaten gewählt werden. Man unterscheidet zwischen strukturierten Eingabedaten (z.B. Tabellen aus einer Datenbank), semi-strukturierte Daten wie HTML- oder XML-Dateien. Unstrukturierte Daten sind z.B. Textdokumente.

3.1.3 Verfahren allgemein

Der allgemeine Ablauf der ontologiebasierten Wissensextraktion gliedert Alexander Mädche in folgende Abschnitte:

1. Import/Wiederverwendung/Konvertierung von Ontologien und anderen Daten (optional)
2. Extraktion von Daten (bottom-up[6]; top-down[7])
3. Beschneidung (Pruning)
4. Veredelung (Refining)
5. Verifikation/Evaluation[8]

Ich schlage eine weitere Klassifizierung der Verfahren nach ihrer Vorgehensweise vor. So kann man unter “bottom-up” und “top-down” unterscheiden. Der Import-Schritt ist optional. Falls man eine oder mehrere Ontologien importiert, handelt es sich bereits schon um ein “Top-down”-Verfahren. Zu der Extraktion von Daten wird im Anschluß mehr gesagt. Nach der Extraktion der Daten folgt die Beschneidung der Ontologie. Unwichtige (domain-unspezifische) Konzepte und Relationen werden entfernt. Die anschliessende Veredelung beinhaltet eine weitere Fokussierung auf den Anwendungsbereich. Nach diesen Schritten sollte eine Verifikation und Evaluation erfolgen. Danach kann man entscheiden, ob die

[6] Man beginnt mit den Daten und schafft daraus die ‘darüberliegende’ Ontologie.
[7] Man beginnt mit der Ontologie und schaut auf die Daten ‘herunter’.
[8] wobei er den Schritt der Verifikation mittels eines Werkzeugs manuell vornahm.

entstandene Ontologie optimal für den Anwendungsbereich ist. Die fünf Schritte können beliebig wiederholt werden, um eine weitere Verbesserung zu erzielen. Alternativ zu den Verfahren gibt es noch Merging und Mapping. Darauf wird später noch kurz eingegangen.

3.1.4 Verfahren im Detail

Hier werden die einzelnen Verfahren grob eingeteilt.

"Bottom-up":
Angefangen wird mit einem Datensatz, aus dem eine Ontologie erstellt wird, die die Daten strukturiert. Unter Verwendung von zwei Text-Sammlungen[9] wird eine statistische Erfassung der Wörter, ihre Häufigkeit, etc. vorgenommen. Die anschliessende Dimensionsreduktion[10] ist notwendig, um das Text-Clustering zu beschleunigen. Anschliessend wird ein domain-spezifisches Lexikon erstellt, welches die domain-relevanten Konzepte enthält. Durch die Anwendung heuristischer Verfahren werden die Relationen erstellt (semantische Analyse). Darauf folgt die Pruning- und Refining-Phase.

"Top-down":
Anfangs hat man bereits eine allgemeine Ontologie, die dann im Verlauf durch bereichsbezogene Daten an den interessierenden Bereich angepaßt wird (Domain-Fokussierung). Man wählt eine (allgemeine) Ontologie[11] und domain-spezifische Quellen, die importiert werden. Anschliessend werden heuristische Verfahren zur Konzept- und Relationsextraktion angewendet. Die bestehende Ontologie wird durch gefundene Konzepte und Relationen erweitert. Danach folgt die Pruning- und Refining-Phase.

"Merging":
Mittels Merging werden zwei oder mehrere bestehende Ontologien zusammengeführt, um eine neue Ontologie zu erstellen, die die vorigen Ontologien abdeckt. Dies kann notwendig sein, wenn man zwei oder mehr Ontologien über den selben Anwendungsbereich hat, und eine allgemeinere benötigt, die die vorigen Ontologien ersetzen kann.

"Mapping":
Bei Mapping handelt es sich um die Erstellung von Regeln, die Entsprechungen aus den Ontologien zuordnet. Somit werden zwei oder mehrere Ontologien beibehalten und Regeln erstellt, die die entsprechende Konzepte zuordnet. Dies ist hilfreich, wenn die bestehenden Ontologien noch weiter verwendet werden sollen. So könnte man Teile einer Ontologie über 'Person' auf eine Ontologie 'Patient' 'mappen', da es nicht unbedingt passend wäre, eine einzige Ontologie zu erstellen, die beide Bereiche abdeckt.

[9] domain-spezifische und allgemeine Textsammlung
[10] durch NLP, Stammbildung, ...
[11] Die Wahl der Kern-Ontologie hat starke Auswirkungen auf die folgenden Schritte.

3.1.5 Extraktion on-demand vs. Vorab-Extraktion

Weiterhin kann zwischen Extraktion on-demand und Vorab-Extraktion unterschieden werden. Unter Extraktion on-demand versteht man, daß die Suche und Extraktion erst auf Anfrage vorgenommen wird. Als Beispiel sei hier eine Suchmaschine erwähnt, die in [DLOE] detailliert beschrieben wird. Bei einer Suchanfrage, in der unbekannte Konzepte enthalten sind, werden diese in die bestehende Ontologie integriert. Auf der anderen Seite steht die Vorab-Extraktion, die derzeit häufiger vertreten ist. Es wird einmalig im Voraus die Ontologie erstellt, beschnitten und veredelt und später optional der komplette Vorgang wiederholt, um die Ontologie zu aktualisieren, was aber nicht dynamisch während der Anwendung abläuft.

3.1.6 Methoden

Weiterhin kann man nach den verwendeten Methoden eine Klassifikation vornehmen. Dies wird später noch detailliert beschrieben.

3.2 Methoden

Eine Einteilung nach Methoden ist ratsam, da sie sowohl bei "Top-down"- und "Bottom-up"-Verfahren verwendet werden und die Einteilung nach "bottom-up" und "top-down" sehr grob ist.

NLP[12] (z.B. SMES[13]:)

- morphologische Analyse (Stamm)
- Semantik-Analyse
- Erkennung benannter Entitäten
- Nutzung domain-spezifischer Informationen

Text-Clustering:

- Reduktion der Text-Dimension durch NLP
- Clusterbildung (iterativ)
- Klassifikation anhand der Cluster Muster-Abgleich

Induktive Verfahren:

- Erkennung und Klassifikation unbekannter Konzepte
- Erkennung von Relationen zwischen Konzepten

Inferenz (mit Description Logic)

Statistik

[12] Natural Language Processing
[13] Saarbrücken Message Extraction System

3.3 Vor- und Nachteile im Vergleich

NLP:
Vorteilhaft ist an NLP, daß es sich an natürlicher Sprache orientiert und es auf Lexika zugreifen kann. Der Nachteil ist, daß sehr viele Heuristiken angewandt werden müssen, die durch manuell erstellte Regeln definiert werden. Diese Regeln gelten meist auch nur für eine Sprache (oder nur einen Teil dieser Sprache!).

Text-Clustering:
Die automatisierte Variante des Text-Clustering hat den Vorteil, daß durch mehrere Iterationen ein Großteil der Begriffe automatisch in Cluster zusammengefaßt werden. Nachteilig ist, daß das Text-Clustering auf eine Domain eingeschränkt ist und daher nicht flexibel für mehrfache Verwendung in unterschiedlichen Bereichen anwendbar ist[14]. Weiterhin ist die Erklärbarkeit der Einteilung meist schwer verständlich. Was für die Maschine ein optimales Clustering-Ergebnis darstellt, kann für den Benutzer nicht nachvollziehbar sein.

Muster-Abgleich:
Muster-Abgleich ist allgemein anwendbar, doch die Nachteile überwiegen. Es müssen viele Heuristiken mittels manueller Regelerstellung konstruiert werden, was eine starke Einschränkung des Systems ist.

Induktive Verfahren:
Induktive Verfahren sind aufgrund ihrer Automatisierung gut. Problematisch wird es, wenn es Widersprche in der Beispielmenge gibt.

Inferenz: (z.B. Description Logic)
Vorteilhaft ist die Möglichkeit der weiteren Regelableitung durch Inferenz. Weiterhin können einige Verfahren mit unvollständigen und fehlerhaften Daten arbeiten[15], im Gegensatz zu anderen Verfahren, die diese Datensätze ignorieren oder anpassen, was zu Verfälschungen führen kann. Hier wurden viele verschiedene Verfahren bzw. Standards entwickelt.

Statistik:
Statistische Verfahren sind verhältnismässig schnell, zuverlässig und bereits bekannt. Nur können diese Verfahren absurde Ergebnisse liefern. Auch hier gibt es das Problem der Verständlichkeit. So sind Ergebnisse meistens schwer interpretierbar.

Bei dieser Klassifikation muß man beachten, daß es Überschneidungen gibt. Eine klare Trennung ist manchmal nicht möglich. So gehört das Clustering-Verfahren zu Statistik.

[14] Das liegt daran, daß man für das Clustering ein domain-spezifisches Lexikon verwendet.
[15] Ein Beispiel für ein fehlertolerantes System, welches Description Logic verwendet, findet man in [DLOE].

3.4 Bewertung der Ansätze

Die einzelne Anwendung einer Methode ist nicht optimal, da jede Methode Vor- und Nachteile hat. Daher empfiehlt sich die Kombination mehrer Methoden, um die Stärken zu kombinieren und Nachteile einzelner Verfahren zu mildern. Häufig wurde eine Kombination von Text-Clustering mit NLP verwendet, was meist zu recht guten Ergebnissen geführt hat.

3.5 Praxis/Trends

Zukünftig wird die Vereinfachung der Entwicklung von Ontologien die Verbreitung und Verwendung von Ontologien erhöhen. Prinzipiell ist das Verfahren allgemein genug, um in vielen Anwendungsbereichen nutzbar zu sein. So gibt es schon für viele Anwendungsbereiche allgemeine Ontologien, die verwendet werden können, die dann anschliessend für die spezifische Aufgabe optimiert werden sollten. Derzeit ist eine Integration und Verwendung von Ontologien in vielen (Forschungs-)Bereichen vorhanden. Weiterhin arbeiten viele Forschungsprojekte an der Verbesserung der Extraktionsfähigkeiten der einzelnen Methoden. Somit fliessen Erfolge aus diesen Bereichen automatisch auch in die ontologiebasierte Wissensextraktion ein. Da es auch verschiedene Methodenarten gibt, sind auch gute Alternativen vorhanden, falls es in einigen Bereichen nicht zu Verbesserungen kommt. Somit hängt das System nicht an einer Methodenart fest. In Kombination mit fortschreitender Automatisierung können ontologiebasierte Informationsextraktionssysteme den Arbeitsaufwand der Wissensextrakrion erheblich reduzieren. Als Beispiel für einen Einsatzbereich von Informationsextraktionssystemen, die Ontologien verwenden, wäre das Semantic Web und Knowledge-Portale zu nennen, in denen Informationen anhand von Ontologien strukturiert und auch visualisiert werden.

4 Zusammenhänge zu den anderen Beiträgen

Im Bereich Wissensmanagement, in der die Strukturierung, Speicherung, Findung und Extraktion von Daten eine wichtige Rolle spielt, können Ontologien alternativ verwendet werden. So kann der Extraktionsprozeß mittels Ontologien vollzogen werden. Sie bieten eine gute Alternative zu den bisherigen Informations-Extraktions-Verfahren, da durch die Verwendung des integrierten Bereichswissens absurde (z.B. durch Mehrdeutigkeit verursachte) Ergebnisse vermieden werden. Beide Verfahren müssen sich nicht unbedingt ausschliessen, da viele Methoden verwendet werden, die sowohl ontologiebasierte, als auch 'normale' Informationsextraktion verwenden. Im Bereich des Semantic Web treten diese Vorteile der ontologiebasierten Wissensextraktion noch stärker heraus, da innerhalb des Semantic Web Ontologien zur Strukturierung des Inhalts verwendet werden. Bei der Benutzung der gleichen oder ähnlichen domain-spezifischen Ontologie wird die ontologiebasierte Wissensextraktion eindeutig bessere Ergebnisse liefern. Weiterhin könnten Verbesserungen innerhalb des Informationsextraktionsprozesses entstehen, wenn die für den Informationsextraktionsprozeß

verwendete Ontologie mittels optimal gewählten Aggregationen (Data Cubes) als auch der Verwendung des Apriori-Algorithmus von Agrawal zur Findung von gemeinsam auftretenden Wortpaaren, die dann als verwandte Konzepte eingestuft werden könnten, optimiert werden könnte. Ausserdem könnte RDT/DB die durch Ontologien strukturierten Daten und das durch Ontologien mitgelieferte Hintergrund-Wissen im Wissensentdeckungsvorgang integrieren, was zu verbesserten Ergebnissen führen kann. Eine weitere Kopplungsmöglichkeit wäre MIDOS, welches zur Findung von interessanten Subgruppen verwendet wird. Diese Subgruppen können 'wichtige Schlüssel'[16] enthalten (bzw. in diesen Subgruppen könnte man wichtige domain-relevante Konzepte finden), die in die verwendete Ontologie einfliessen sollten (falls sie bisher nicht beachtet wurden). Ales letzten Punkt wäre die Kombination von ontologiebasierter Wissensextraktion und Methoden der Zeitreihenuntersuchung. Man könnte die Entwicklungsgeschichte des Datenbestandes oder die Ergebnisse, die durch den ontologiebasierten Wissensextraktionsprozess geliefert werden, auf bestimmte Muster untersuchen, was Rückschlüsse auf die Güte der verwendeten Ontologie liefern könnte. Ob dies wirklich machbar und nützlich ist, ist schwer abschätzbar.

5 Schlußwort

Abschliessend ist das Verwenden von Ontologien eine hilfreiche Technik, die auf spezielle Bereiche zugeschnitten werden kann (durch die Wahl einer geeigneten Ontologie), aber trotzdem durch ein festes System realisiert wird, da die Methoden ontologie-unabhängig sind. Durch die Änderung der Ontologie kann das System somit an die Aufgabenstellung angepaßt werden. Weiterhin profitiert das Verfahren aufgrund ihrer vielen unterschiedlichen Methoden aus vielen Bereichen der Informatik aus Erfolgen in jedem dieser Bereiche. Nachteilig ist, daß es eine Verkettung von vielen Verfahren ist, die nicht immer optimal verbunden werden können. Daraus entsteht ein sehr komplexes System, welches einen hohen Aufwand birgt. Weiterhin resultieren aus einer fehlerhaften bzw. unvollständigen Modellierung der Ontologie grosse Probleme[17]. Außerdem birgt das Verwenden von Ontologien wieder Mißbrauchsmöglichkeiten. Als Beispiel sei hier das Semantic Web erwähnt, welches auf die Verwendung von Metadaten zur Kennzeichnung der Dokumente beruht. Als Vorstufe kann man die Verwendung der Meta-Tags in HTML-Dokumenten sehen, die von Suchmaschinen ausgelesen werden, um Seiten nach ihrem Inhalt zu klassifizieren. Der Benutzer hat die Aufgabe, passende Begriffe auszuwählen und einzutragen. Dies wurde bisher reichlich mißbraucht, um die Anzahl der Treffer auf einer Seite zu erhöhen. Somit ist eine einfache Mißbrauchsmöglichkeit von Anfang an vorhanden, die dann den Wissensextraktionsprozeß in eine nicht-beabsichtigte domain-unspezifische Richtung beeinflussen kann.

[16] domain-relevante Konzepte

[17] Die Problematik von: Vollständigkeit vs. Minimalität

6 Literaturangaben

Bei bestehendem Interesse kann man folgende hier verwendete Quellen für weitere Details konsultieren.

[OBE98] D. W. Embley, D. M. Campbell, S. W. Liddle, R. D. Smith. *Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents in CIKM'98.*

[OBI'99] A. Mädche, S. Staab, R. Studer. *Ontology-based Information Extraction and Integration in DGfS/CL'99.*

[SOAC] J.-U. Kietz, A. Mädche, R. Volz. *A Method for semi-automatic Ontology Acquisition from a corporate Intranet in EKAW2000.*

[STDS] H. Graubitz, K. Winkler, M. Spiliopoulou. *Semantic Tagging of Domain- Specific Text Documents with DIAsDEM in DBFusion 2001.*

[OBTC] A. Hotho, S. Staab, A. Mädche. *Ontology-based Text-Clustering in IJCAI2000.*

[LOSW] A. Mädche, S. Staab. *Learning Ontologies for the Semantic Web in ECML/PKDD2001.*

[DLOE] A. Todirascu. *Using Description Logics for Ontology Extraction in Ontology Learning 2000 at ECAI2000.*

[LAWL] http://www-personal.umich.edu/ jlawler/routledge/glossary.html

Letzte Anderung ist am 15. November 2001 vorgenommen worden. Erstellt wurde dieses Dokument mit LaTeX.