ter Informationen wie z.B. Zeitschriftenname, Band (Vol.), Ausgabe (No.), Seitenzahl, ISSN, ISBN, Verlag, Herausgeber. Auch das Datum muss aus vielen verschiedenen Schreibarten in „Jahr-Monat-Tag“ umgewandelt werden. Der Typ der Publikation kann manchmal nur aus der URL erraten werden. Man kann annehmen, dass die extrahierten Daten vollständig semantisch korrekt und in dem Extrahierungsprozess bereits bereinigt worden sind. Für die übrigen Felder, wie Autoren, Titel und Zusammenfassung muss dies erst durchgeführt werden. Das bedeutet Entfernung aller überflüssigen Textauszeichnungen (HTML und LaTeX Strukturen), Korrektur und Umschreibung von Sonderzeichen (UTF-8, HTML Entities, TeX Darstellungsart, falsch genutzter Mathematikmodus von TeX). Die Autorennamen, die als zusammengesetztes Textelement vorliegen, müssen getrennt werden und entschieden, welcher Teil der Teilkette der Vorname ist, welcher der Nachname und was nicht zum Namen gehört und entfernt werden soll.

Die obengenannten Probleme sind nur ein Teil von in der CCSB aufgetretenen Aufgaben, die im Projekt größtenteils gelöst wurden. Mehrere zusammengesetzte heuristische Regeln, mit regulären Ausdrücken als Werkzeug, haben die Datenextrahierung und -bereinigung möglich gemacht.

2 Projekt Semantische Methoden und Werkzeuge für Informationsportale (SemIPort)

2.1 Ontologie-basiertes Web Mining zum Aufbau großer Informationsportale

Die Erkennung und Extraktion relevanter Daten im Internet wird zunehmend durch den rapiden Zuwachs an Dokumenten erschwert. Bestehende Ansätze, denen aktuelle Suchmaschinen in der Regel folgen, entgegnen den anfallenden Datenmengen mit immer neuer Rechenleistung. Diese Vorgehensweise wird sich jedoch nicht beliebig fortsetzen lassen. In dem SemIPort Projekt wurde der fokussierte Web-Crawler METIS (http://ontoware.org/projects/metis) zur Identifikation und Extraktion kontextrelevanter Informationen aus dem Internet entwickelt, welcher Hintergrundwissen in Form von Ontologien verwendet.

Grundsätzlich wird zwischen mehreren Arten von Ontologien unterschieden. Zum einen wird eine **Web-Ontologie** modelliert. Diese beschreibt die Struktur und Eigenschaften von Dokumenten im Internet, sowie deren Verknüpfungen mittels sog. *Hyperlinks*. Sie repräsentiert außerdem *Hosts,* auf denen Internet-Dokumente gespeichert werden. In der **Domänen-Ontologie** wird die eigentliche Domäne beschrieben. Das dort gespeicherte Wissen stellt letztendlich das Ziel der fokussierten Suche dar. Zum Aufbau eines Informationsportals für wissenschaftliche Publikationen aus der Informatik beschreibt die Domänen-Ontologie z. B. Fachrichtungen, Eigenschaften von Publikationen und beschreibt u.a. Personen und Forscher.

Im Gegensatz zu Informationsextraktionsmechanismen, die eine Bewertung von Res-

sourcen erst nach der eigentlichen Extraktion zulassen[1], zielt der von uns entwickelte Ansatz auf eine Bewertung von Ressourcen während der eigentlichen Suche ab. Dies ermöglicht eine effektive und effiziente Nutzung vorhandener Kapazitäten[2]. Die fokussierte Suche nach Ressourcen basiert auf der Bestimmung der Relevanz einer Ressource zu einer bestimmten Domäne oder Teildomäne. Generell wird als Suchstrategie die Verfolgung von Relationen zu Ressourcen mit möglichst hoher Relevanz verwendet (fokussiertes Crawlen). Die Berechnung der Relevanz setzt sich zusammen aus der **inhaltlichen Analyse** des Dokumenteninhalts und einer **Bewertung der Verlinkungsstruktur** der Ressource. Das Ergebnis ist ein numerischer Wert, welcher zur Erzeugung einer sortierten Menge an Ressourcen verwendet wird. Diese Menge dient zur weiteren Suche, wobei Ressourcen mit höherer Relevanz vorrangig verarbeitet werden.

Zusammenfassend werden Ontologien zur Modellierung der Umgebungswelt und zur Modellierung der eigentlichen Anwendung (Domäne) eingesetzt. Die Domänen-Ontologie beschreibt dabei Konzepte zu denen weitere Instanzen identifiziert werden sollen. Die Suche nach diesen Instanzen wird mittels einer semantischen Bewertung durchgeführt.

2.2 Personalisierte Benutzerinteraktion mit wissenschaftlichen Informationsportalen

Die Interaktion mit Information besteht nicht nur aus Information Retrieval, sondern auch aus dem Organisieren und Verstehen der gesammelten Daten. Gerade bei wissenschaftlichen Informationssystemen ist eine entsprechende Werkzeugunterstützung wertvoll. Wird solch ein Werkzeug in ein Informationsportal integriert, bietet es zudem die Möglichkeit, dem Benutzer bei der Informationssuche gezielter zu unterstützen, da mehr Indizien über dessen Informationsbedarf im System vorhanden sind, beispielsweise der Teil der persönlichen Dokumentsammlung, mit der sich der Benutzer gerade beschäftigt. Diese zusätzlichen Daten können auch zur Unterstützung anderer Benutzer verwendet werden; hier bieten sich kollaborative Empfehlungssysteme oder sogar einfach nur das Verfügbarmachen der von anderen Benutzern definierten Annotationen, wie eine Themenzuordnung zu Dokumenten, an.

Im Rahmen von SemIPort wurde ein persönliches Dokumentmanagementwerkzeug und ein zugehöriges Empfehlungssystem entwickelt [SJ04,Sc04], welche in Informationsportale integriert werden können. Die Kernidee dabei ist, dass die Benutzer persönliche Wissensbasen auf Grundlage einer gemeinsamen Ontologie aufbauen und diese Wissensbasen vom Empfehlungssystem gesammelt und zur Generierung von Empfehlungen genutzt werden. Die gemeinsame Ontologie ermöglicht hier eine einfache Integration der Daten aus den verschiedenen Quellen, sowie die Herstellung des Bezugs zwischen den Benutzerdaten und dem Informationsraum des Portals auf einer abstrakteren als der reinen Inhaltsebene. Für Empfehlungen kann so z.B. das Wissen genutzt werden, dass

---

[1] An dieser Stelle sei auf Methoden wie HITS-Algorithmus und PageRank-Algorithmus verwiesen.
[2] Vorwiegend Reduktion des Bedarfs an Speicherplatz, Bandbreite und Rechenzeit.