Integration von Datenbanken in Suchmaschinen bei unterschiedlichen Kooperationsgraden

Gunnar Weber
Lehrstuhl Datenbank- und Informationssysteme
Fachbereich Informatik, Universität Rostock
weber@informatik.uni-rostock.de

Zusammenfassung

In Datenbanken enthaltene Informationen machen einen Großteil des im WWW vorhandenen Wissens aus. In den meisten Fällen sind diese Informationen in dynamisch generierte WWW-Seiten eingebunden, so daß sie von den aktuellen Suchmaschinen nicht indexiert werden können. Im Projekt SWING wurden verschiedene Lösungen zur Integration von Datenbanken entwickelt. Ein bereits realisierter Ansatz nutzt lokal beim Datenbank-Anbieter vorhandene Summarizer, um Index-Fragmente, die aus den Datenbankinhalten generiert werden, an den Gatherer zu liefern. Eine weitere Möglichkeit ist die Beschreibung der Funktionalität des Anfrageformulars und des Aufbaus der Ergebnisseiten in Metadaten, die vom Gatherer zur automatischen Generierung von Anfragen und zur Extraktion der Informationen aus den Ergebnisseiten genutzt werden können.

1 Einleitung

Die Gatherer[1] der aktuell verfügbaren Suchmaschinen sammeln nur Daten ein, die im sogenannten öffentlich indexierbaren Web [6] vorhandenen sind. Informationen, die hinter Suchmasken verborgen sind, werden ignoriert. Dadurch wird bei der Recherche im Internet eine gewaltige Menge hochqualitativer Informationen aus großen, durchsuchbaren Datenbanken nicht berücksichtigt. Solche Datenbanken haben typischerweise ein Web-Anfrageinterface, das ein HTML-Formular enthält. Der Zugriff auf die Datenbank ist dann nur über dynamisch generierte Seiten möglich, die als Antwort auf die Nutzeranfrage geliefert werden.

Es gibt verschiedene Varianten für die Integration von Datenbanken in Suchmaschinen. Diese unterscheiden sich in der Möglichkeit des Zugriffs auf die lokale Datenbank:

- *Kooperative Datenbankanbieter* ermöglichen den strukturierten Zugriff auf den Datenbankinhalt (oder spezielle Sichten) über JDBC oder ähnliche Mechanismen. In diesem Fall kann die volle Funktionalität einer Anfragesprache auch in der Datenbankanbindung ausgenutzt werden.

- Bei *nicht-kooperativen Datenbankanbietern* können Anfragen nur über ein Anfrageformular, das im WWW über Parameter versorgt wird, gestellt werden. Das Schema der Datenbank und Anfragefunktionen, die darüberhinaus nutzbar wären, sind nicht bekannt bzw. können nicht an die darunterliegende Datenbank übergeben werden.

In diesem Artikel wird unter anderem ein Extraktionsmechanismus vorgestellt, bei dem vom Anbieter bereitgestellte Metadaten zur Indexierung herangezogen werden. Da der Anbieter im Gegensatz zum nicht-kooperativen Fall aktiv werden muß, um die Indexierung zu ermöglichen, wird die Klasse der *eingeschränkt kooperativen Anbieter* eingeführt: Hier erfolgt der Zugriff auf die Datenbank zwar wieder über das Web-Formular (wie beim nicht-kooperativen Ansatz), das Schema der Datenbank und die Anfragefunktionalität des Formulars werden aber über spezielle Metadaten bekannt gegeben (wie im kooperativen Fall).

Bei der Indexierung von Inhalten, die hinter Anfrageformularen in Datenbanken “versteckt” sind (oft auch als *Hidden Web* bezeichnet), treten drei grundlegende Probleme auf.

Das erste ist das Problem der Skalierung. Nach einer Studie [1] ist das Hidden Web bis zu 500 mal größer als das statische Web. Aus diesem Grund ist eine umfassende Indexierung unmöglich und häufig auch nicht sinnvoll. So ist es z.B. unsinnig, eine Artikelnummer zu indexieren. Die aktuell verfügbaren Suchmaschinen berücksichtigen keine Zusammenhänge zwischen den Informationen eines Doku-

[1] Gatherer sammeln Informationen über Dokumente, die im Internet verfügbar sind.

mentes[2]. Somit reicht es aus, nur unterschiedliche Werte für ein Attribut und keine kompletten Tupel in den Index aufzunehmen. Dies führt zu einer deutlichen Reduzierung der Größe des Indexfragmentes für eine Datenbank.

Das zweite Problem liegt in der automatischen Erfassung der Funktionalität des Anfrageformulars, das für den menschlichen Nutzer konzipiert ist. In diesem Beitrag werden deshalb Metadaten vorgeschlagen, mit denen Anbieter diese Funktionalität in maschinenverständlicher Art und Weise beschreiben können.

Die Aufgabe des Anfrageformulars besteht in der Filterung der für einen Nutzer relevanten Datenbankinhalte. Diese Filterung führt zum dritten Problem, da ein Gatherer alle wichtigen Inhalte einsammeln soll. Somit muß der Gatherer beim Zugang über eine Web-Schnittstelle mehrere Anfragen nacheinander an die Datenbank stellen. Mit zunehmender Komplexität des Anfrageformulars und zunehmender Anzahl möglicher Werte für die Elemente steigt der Zeit- und Ressourcenaufwand für das Abfragen der Datenquelle sehr stark an.

Im kooperativen Fall spielen die letzten beiden Probleme keine Rolle, da direkt auf die Datenbank zugegriffen wird.

Für den nicht-kooperativen Fall sei auf zwei interessante Ansätze verwiesen. An der Stanford Universität wurde ein aufgabenspezifischer Web-Crawler mit der Bezeichnung *Hidden Web Exposer* [7] entwickelt, der beim Bestücken der Formulardaten auf eine endliche Menge von Konzepten bzw. Kategorien zugreift. Die Menge der möglichen Werte pro Konzept wird dynamisch um neu extrahierte Informationen erweitert. Die Zuordnung eines Formularelements zu einem Konzept wird über Ähnlichkeiten zwischen der textuellen Beschreibung des Elements und der Konzeptbeschreibung vorgenommen. In [2] wird das Ausnutzen von kombinierten Analysetechniken des Text Mining vorgeschlagen, um Formularelemente zu Konzepten einer Ontologie zuzuordnen. Obwohl die Verfahren der linguistischen Analyse immer weiter verbessert werden, sind sie immer noch sehr zeitaufwendig und können durch die Mehrdeutigkeit der natürlichen Sprache zu falschen Schlußfolgerungen führen. Durch die Verwendung von Ontologien für die Analyse bzw. das Bestücken der Formularfelder sind beide Ansätze domänenspezifisch und erfordern einen hohen Aufwand bei der Erschließung neuer Domänen.

In Abschnitt 2 wird zunächst ein in der Suchmaschine SWING realisierter Ansatz vorgestellt [5], der die Einbindung von Datenbanken kooperativer Anbieter erlaubt. Der Hauptteil des Artikels (Abschnitt 3) beschäftigt sich mit der Integration von Inhalten teilweise kooperativer Anbieter. Die Grundidee dieses Ansatzes ist es, Metadaten so zu definieren, daß der Gatherer die Datenbankinhalte über die Web-Schnittstelle ermitteln kann.

[2] außer der Entfernung zwischen Wörtern

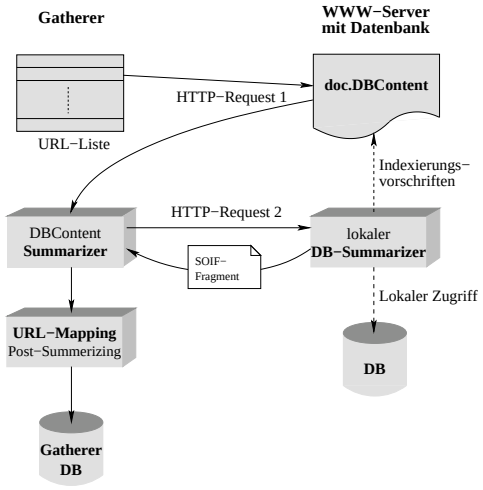

Abbildung 1: Integration von Datenbanken

2 Kooperativer Ansatz - Einbindung von Datenbanken in SWING

Basisbestandteil der Suchmaschine SWING ist das *Harvest-System* [4]. Zur Einbindung von Datenbanken wurde die Gatherer-Komponente des Harvest-Systems modifiziert. Die Grundidee des Ansatzes besteht in der Verwendung lokaler Summarizer, die leicht konfigurierbar sind und vorhandene Standards für den Datenbankzugriff nutzen (DBPerl, JDBC). Ein lokaler Summarizer ist ein Programm, das auf dem WWW-Server des Datenbankanbieters läuft und entsprechend der Konfigurationsangaben generierte Anfragen an die zu indexierende Datenbank stellt.

Damit der Gatherer erkennen kann, daß es sich um eine einzubindende Datenbank handelt, wurde der neue Dokumenttyp `DBContent` eingeführt. Solch ein Dokument muß auf dem WWW-Server, der den Zugriff auf die einzubindende Datenbank ermöglicht, vorhanden sein und enthält folgende Informationen: die Adresse des lokalen Summarizers, die URL der Web-Anfrageschnittstelle, Metadaten wie *Description* und *Keywords* sowie Angaben darüber, welche Attribute aus welcher Relation indexiert werden sollen.

In der Abbildung 1 sind die Schritte graphisch dargestellt, die vom Gatherer durchgeführt werden, um Datenbankinhalte in das Retrievalsystem der SWING-Suchmaschine einzuspeisen.

Die URLs für die Konfigurationsdateien müssen dem Gatherer explizit bekannt gemacht werden. Stößt der Ga-

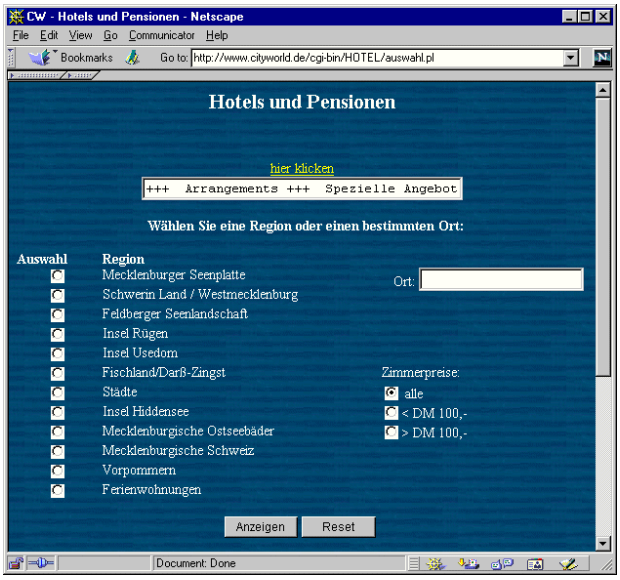

Abbildung 2: DB-Formular

therer bei der Abarbeitung der *URL-Liste* auf solch eine Datei, wird der *DBContent-Summarizer* aufgerufen. Dieser parst die Konfigurationsdatei und ruft dann die URL auf, die für den *lokalen Summarizer* angegeben wurde. Als Parameter wird der Pfad der Konfigurationsdatei relativ zum lokalen Summarizer übergeben. Über diesen Pfad kann der lokale Summarizer auf die Datei zugreifen, um die datenbankspezifischen Informationen zu lesen (Name der Datenbank, zu indexierende Tabellen und Attribute). Anschließend werden SQL-Anfragen zum Ermitteln der Werte für die zu indexierenden Attribute an die Datenbank gestellt. Im letzten Schritt faßt der lokale Summarizer alle Ergebnisse in einem *SOIF-Fragment*[3] zusammen und schickt es an den Gatherer zurück. Dort wird es dann mit der URL für das Anfrageformular und weiteren Metainformationen (Größe des Fragmentes usw.) in der *Gatherer-Datenbank* gespeichert.

In die SWING-Suchmaschine wurde beispielsweise der CityWorld-Hotelführer integriert, der auf einer Oracle-Datenbank basiert. Dieser Führer enthält eine Liste von Hotels und Pensionen in Mecklenburg-Vorpommern mit deren Adressen. Das Anfrageformular (siehe Abbildung 2) gestattet die Suche nach einer Region oder einem Ort, wobei das Ergebnis noch auf bestimmte Preiskategorien eingeschränkt werden kann. Als Ergebnis wird eine Tabelle (Abbildung 4) mit allen Hotels oder Pensionen geliefert, die den Suchkriterien entsprechen. Auf dem WWW-Server, über den der Zugriff auf diese Datenbank erfolgt, ist das Dokument `cityworld.DBContent` vorhanden, das folgende Informationen bereitstellt:

```
Summarizer:   http://www.cityworld.de/cgi-bin/Ora-DB.pl
URL:          http://www.cityworld.de/cgi-bin/auswahl.pl
Title:        CityWorld - Hotels und Pensionen
Description:  Hotelführer M-V
Keywords:     Hotel, Zimmer, Unterkunft, Pension, Preis
Database:     Cityworld
Table:        hotel
Attribute:    name
Attribute:    ort
Table:        region
Attribute:    region
```

Der Index für den Hotelführer enthält somit 3 Attribute, denen alle vorhandenen Hotels, Orte bzw. Regionen zugeordnet sind.

3 Datenbankeinbindung bei eingeschränkter Kooperation

In vielen Fällen haben Anbieter von Web-Datenbanken Bedenken, fremde Programme auf ihren Rechnern zu installieren, die Inhalte nach außen geben. Deshalb soll im folgenden ein Mechanismus vorgestellt werden, der die normale Web-Schnittstelle für die Indexierung nutzt. Dieser Ansatz geht davon aus, daß der Anbieter mit Hilfe von Metadaten beschreibt, welche Funktionalität das Anfrageformular bietet und wie die Ergebnisseiten aufgebaut sind.

Metadaten dienen zur Definition bestimmter Eigenschaften eines HTML-Dokuments, die vom WWW-Browser bei der Darstellung nicht anzeigt werden. Suchmaschinen können diese Informationen interpretieren und sie beispielsweise bei der Bewertung des Dokuments bezüglich der Suchanfrage oder bei der Ausgabe des Suchergebnisses verwenden. Metadaten werden mit Hilfe von *Meta-Tags* angegeben, die im Kopf eines HTML-Dokuments stehen müssen und folgende Syntax haben:

```
<META NAME="Meta-Name" CONTENT="Meta-Wert">
```

Im folgenden Unterabschnitt wird zunächst auf den Aufbau von Web-Formularen und die Probleme bei der automatischen Verarbeitung dieser Formulare eingegangen. Anschließend erfolgt dann eine detaillierte Beschreibung des Gatherers, der die Integration von Datenbanken über Metadaten gestattet. Zur Bestimmung der relevanten Inhalte auf den Ergebnisseiten sind Extraktionsregeln notwendig, die in Abschnitt 3.3 vorgestellt werden. Der letzte Unterabschnitt gibt einen Überblick über alle Metadaten, die zur Integration notwendig sind.

3.1 Aufbau von Web-Formularen

Ein Formular $F = (E_1, D_1), (E_2, D_2), ..., (E_n, D_n)$ ist eine Menge von *(Element, Domäne)*-Paaren, wobei das Element E_i eines der folgenden Standardeingabeobjekte sein

[3] Das *Summary Object Interchange Format* (SOIF) besteht aus einer Liste von Attribut/Wert-Paaren.

kann: Auswahlliste, ein- oder mehrzeiliges Textfeld, Radiobutton oder Checkbox. Die Domäne D_i bezeichnet die Menge aller Werte, die das zugehörge Element E_i annehmen kann. Einige Elemente haben *bestimmte Domänen*, d.h. die gültigen Werte sind in der HTML-Seite eingebettet (Auswahllisten, Radiobuttons, Checkboxen). Andere Elemente wie Textfelder haben *unbestimmte Domänen*, d.h. sie können beliebige Werte annehmen. Zusätzlich werden die meisten Elemente mit einem beschreibenden Text versehen, der dem Nutzer die Semantik des Elements verständlich machen soll. Dieser Text wird im folgenden als *Label* bezeichnet. Jedes Element hat einen internen Bezeichner, der bei der Übertragung der Formulardaten verwendet wird. Bei Radiobuttons bzw. Checkboxen korrespondiert zusätzlich jeder nach außen sichtbare Wert mit einem internen Bezeichner, bei Auswahllisten sind diese optional.

Für die Indexierung der Datenbankinhalte sind die Labels und Domänen relevant, bei der automatischen Ausführung der Formulare werden dagegen die internen Bezeichner benötigt.

Das automatische Extrahieren der Labels und Domänen (bei Radio- und Checkboxen) ist ein schwieriges Problem, da ihre Beziehung zum Formularelement nicht fest vorgegeben ist. Für den menschlichen Nutzer ist diese Zuordnung einfach durch die Entfernung des Labels bzw. des Wertes vom Formularelement zu erkennen, wenn das Anfrageinterface durch den Browser dargestellt wird. Im HTML-Text ist dieser Abstand nicht bestimmbar, da Formularelemente, Domänenwerte, Label und Layoutinformationen willkürlich ineinander geschachtelt werden können. Das Formular müßte somit visualisiert werden, damit ein Gatherer die richtigen Beziehungen über die Abstände ermitteln kann.

Für die Beantwortung von Anfragen sind einige Elemente wesentlich, d.h. die Angabe eines Wertes ist unbedingt erforderlich. Es gibt aber auch Elemente, die nicht spezifiziert werden müssen und nur eine Reduzierung der Ergebnismenge bewirken. Die automatische Bestimmung der wesentlichen Attribute ist kein triviales Problem, auf das hier auch nicht näher eingegangen werden soll.

3.2 Extrahieren der Datenbankinhalte

Die grundlegenden Aktionen des Gatherers, der das Einsammeln von Datenbanken eingeschränkt kooperativer Anbieter erlaubt, ähneln denen eines traditionellen Gatherers. Unterschiede in der Abarbeitung treten erst auf, wenn der Gatherer auf ein Dokument stößt, das datenbankspezifische Metadaten enthält.

Die Abbildung 3 zeigt die grundlegende Arbeitsweise des Gatherers. Der *Gathering-Manager* steuert den gesamten Prozeß der Datensammlung. Er entscheidet, welcher Link als nächstes bearbeitet werden soll und fordert dann das zugehörige Dokument an. Dieses Dokument wird an den *Parser* weitergegeben, der inhaltliche Informationen ermittelt und diese zusammen mit Informationen über das Dokument selbst (URL, Größe, usw.) in strukturierter Form (z.B. SOIF) in der *Gatherer-Datenbank* ablegt. Die extrahierten Links werden in die *URL-Liste* eingefügt und stehen somit dem Gathering-Manager zur Verfügung. Enthält das Dokument datenbankspezifische Metadaten, handelt es sich um ein Anfrageformular für eine Web-Datenbank. Der Gatherer unternimmt folgende Schritte, um die Inhalte der Datenbank zu bestimmen:

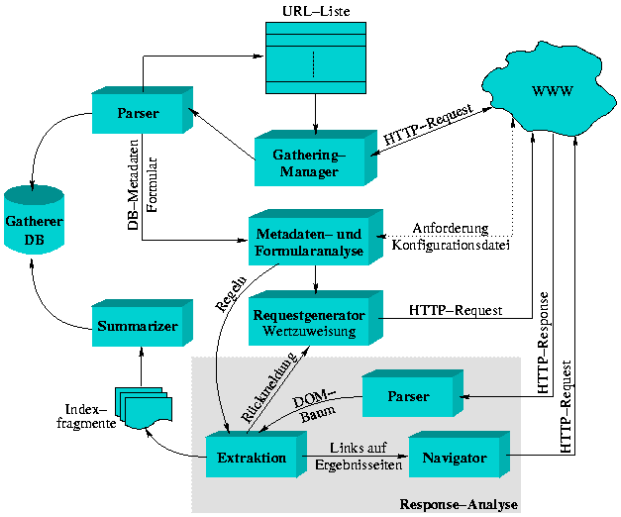

Abbildung 3: Arbeitsweise des Gatherers

- *Metadaten- und Formularanalyse*: Die vom Parser übergebenen DB-Metadaten und das Anfrageformular werden zunächst zerlegt und analysiert. Wird in den Metadaten auf eine Konfigurationsdatei verwiesen, dann muß diese über einen zusätzlichen HTTP-Request angefordert werden. Für das Formular sind folgende Information zu bestimmen: die URL des aufzurufenden Skripts und die Methode für die Parameterübergabe, alle Elemente, deren Typen und interne Bezeichner sowie alle bestimmten Domänen.

- *Request-Generation*: Bei der Erzeugung der Anfragen werden nicht alle Formularelemente berücksichtigt, sondern nur diejenigen, die vom Datenbankanbieter als wesentlich spezifiziert wurden. Ist nur das Element E_i mit der Domäne D_i wesentlich, dann werden $|D_i|$ Anfragen erzeugt. Bei $n > 1$ Elementen wird das kartesische Produkt der Domänen $D_1 \times D_2 \times ... \times D_n$ gebildet, so daß insgesamt $|D_1| * |D_2| * ... * |D_n|$

Anfragen an die Datenbank gestellt werden müssen. Bei unbestimmten Domänen, die mit Hilfe regulärer Ausdrücke definiert wurden (siehe Abschnitt 3.4), ist zusätzlich noch eine Erzeugung der möglichen Werte notwendig.

- *Response-Analyse*: Die Bestimmung der Inhalte auf der Ergebnisseite erfolgt mit Hilfe der in den DB-Metadaten enthaltenen Extraktionsregeln. Einzelheiten zu diesen Regeln können in Abschnitt 3.3 nachgelesen werden. Die Ergebnisseite dient als Eingabe für einen *HTML-Parser*, der daraus einen Analysebaum erzeugt. Innerhalb dieses Baums kann jedes Element im Dokument anhand von Pfadausdrücken eindeutig identifiziert und dessen Inhalt bestimmt werden. Für jede Antwortseite entsteht so ein Indexfragment, das die extrahierten Inhalte zusammenfaßt. Ist das Ergebnis über mehrere HTML-Seiten verteilt, müssen diese ebenfalls angefordert werden. Dafür ist der *Navigator* zuständig, der alle Links auf Ergebnisseiten sammelt und nacheinander abarbeitet.

- *Zusammenfassung*: Der *Summarizer* faßt die Indexfragmente, die pro Anfrage erzeugt wurden, zusammen. Doppelte Werte für ein Attribut werden dabei eleminiert.

Die extrahierten Datenbankinhalte werden zum Schluß gemeinsam mit dem Anfrageformular in der Gatherer-Datenbank abgelegt.

Im folgenden Abschnitt soll näher auf die Extraktionsregeln eingegangen werden, die eigentlich für das Wrapper-Toolkit *W4F* entwickelt wurden. Es existiert eine ganze Reihe solcher Werkzeuge wie *WebL* oder *JEDI*, die ebenfalls Mechanismen zur Extraktion von Daten anbieten [3]. Im Gegensatz zu W4F werden hier aber Skriptsprachen eingesetzt, die das Beschreiben der Ergebnisseiten erschweren.

3.3 W4F-Extraktionsregeln

W4F[4] [8, 9] ist ein Werkzeug zur Generation von Wrappern für Web-Datenquellen. Zur Erstellung robuster Extraktionregeln stellt W4F mit *HEL* (***H**TML **E**xtraction **L**anguage*) eine deklarative Sprache zur Verfügung.

Als Grundlage für die Definition der Regeln dient der Analysebaum, der aus dem HTML-Dokument erstellt wird. Die Zuordnung von HTML-Dokument und Analysebaum ist eineindeutig, d.h. dasselbe HTML-Dokument ist wieder aus dem Analysebaum ermittelbar und umgekehrt.

Der Baum besteht aus einer Wurzel (Bezeichner `html`), internen Knoten und Blättern. Jeder Knoten korrespondiert mit einem HTML-Tag oder Textstück (Bezeichner `PCDATA`). Interne Knoten repräsentieren die geschlossenen HTML-Tags[5] und haben Kinder, auf die über den Bezeichner (Name des zugehörigen HTML-Tags) und den Index[6] (die Reihenfolge des Auftretens) zugegriffen werden kann. Blätter sind entweder offene HTML-Tags (z.B. `img` oder `br`) oder `PCDATA`-Knoten.

Die Navigation entlang des abstrakten Baumes geschieht über Pfadausdrücke. Zur Bildung dieser Ausdrücke stehen zwei Navigationsansätze zur Verfügung:

- Der *Punkt-Operator* nutzt die Baumhierarchie aus, d.h. hinter jedem Element wird getrennt durch den Punkt-Operator ein unmittelbarer Kindknoten im Baum angegeben. Damit läßt sich zu jedem Knoten genau ein Pfadausdruck angeben. Der Pfad `html.head.title` führt z.B. zu dem Knoten, der mit dem `<Title>`-Tag im Kopf des Dokuments korrespondiert.

- Der *Pfeil-Operator* orientiert sich am Dokumentfluß. Das Durchlaufen des Baumes geschieht hier nach dem *depth-first-Prinzip*. Beispielsweise führt der Pfad `html->table[0]` zur ersten Tabelle im Dokument. Die Navigation erfolgt hier über die Hierarchie des Baumes hinweg, so daß z.B. von einem Blatt zu einem anderen Knoten innerhalb des Baumes gesprungen werden kann.

Als Index für ein Pfadelement können auch Intervalle oder Wildcards verwendet werden, so daß mehrere Knoten das Ergebnis bilden. Beispielsweise liefert `html.body->a[*]` alle Anker des HTML-Dokuments.

Die Extraktionsregeln sind nicht nur auf die Knoten selbst beschränkt, sondern sie können auch auf die Informationen zugreifen, die diese tragen. Jeder Knoten hat ein zugeordnetes Textattribut `.txt`. Bei Blättern liefert dieses Attribut im Falle von `PCDATA` das zugehörige Textstück. Bei internen Knoten wird der Wert des Textattributs aus der rekursiven Verkettung aller Unterknoten gebildet. Der Ausdruck `html.body.table[0].tr[0].th[0].txt` ermittelt z.B. die Überschrift der ersten Tabellenspalte. Auf die Eigenschaften der Knoten wie die Werte der Attribute oder die Anzahl der Kinder kann über die Funktionen `getAttr` oder `numberOf` zugegriffen werden. So liefert der Pfadausdruck `html.body->a[0].getAttr(href)` beispielsweise die URL des ersten Links im HTML-Dokument.

Zur Extraktion von Informationen reicht die HTML-Struktur, die durch den Analysebaum bereitgestellt wird,

[4] WysiWyg Web Wrapper Factory

[5] Geschlossene HTML-Tags sind alle HTML-Sprachelemente, die durch eine Start- und Endmarke gebildet werden, z. B. `<table>...</table>`.

[6] Die Vergabe der Indizes geschieht nach dem *left-depth-first-Prinzip*, d.h. die Durchnummerierung gleichnamiger Knoten beginnt zuerst links absteigend im Baum.

nicht immer aus. Dieser Fall tritt z.B. auf, wenn auf einzelne Elemente zugegriffen werden soll, die innerhalb einer Tabellenspalte aufgezählt sind. HEL stellt für solche Fälle die Operatoren `match` und `split` bereit, die mit regulären Ausdrücken arbeiten.

Zur Definition robuster Extraktionsregeln können Bedingungen formuliert werden. Statt fester Indizes stehen dann im Pfadausdruck Variablen, deren Wert jeweils durch eine Bedingung bestimmt wird. Für jede eingeführte Variable muß genau eine Bedingung in der `WHERE`-Klausel angegeben sein, die Verknüpfung erfolgt konjunktiv über das Schlüsselwort `AND`. Die Bedingungen können nicht die Knoten selber einbeziehen, sondern nur deren Werte und verschiedene Vergleichsoperatoren, die in HEL erlaubt sind.

Durch den deklarativen Charakter der Sprache und die Unterstützung bei der Erstellung der Regeln durch ein Werkzeug eignet sich HEL sehr gut für die Beschreibung der Ergebnisseiten.

3.4 Notwendige Metadaten

In der Tabelle 1 sind alle Metadaten aufgelistet, die zur Einbindung von Datenbanken über Web-Formulare benötigt werden. **Fett** hervorgehobene Anteile stehen für Terminale, Nichtterminale sind durch die Zeichen < und > geklammert. Geschweifte Klammern dienen zur Kennzeichnung von Bestandteilen, die null- oder mehrmalig auftreten können und Alternativen werden durch einen vertikalen Strich getrennt.

Bezeichner	*Wert*
ConfigFile	<URL_Konfigurationsdatei>
QueryScript	<URL_Script>**, GET** | **POST**
QueryElement **QueryElements**	<Element> <Element>**,** <Element> {**,** <Element> }
QueryElement.<Element>	/<Einfacher_regulaerer_Ausdruck>/ oder <Wert> {**,** <Wert> }
Result	<gemeinsamer_Teil_einer_Extraktionsregel>
Result.<Attributname>	<W4F_Extraktionsregel> { <Attributwert> {**,** <Attributwert> } }
NavigationURL	<URL_mit_Platzhalter_fuer_Parameter>

Tabelle 1: Notwendige Metadaten für die Datenbankeinbindung

Zur Unterscheidung von anderen Metadaten müssen datenbankspezifischen Informationen immer mit den Initialen `DB` beginnen. Im folgenden werden die einzelnen Metadaten detailliert beschrieben:

- *ConfigFile*: Alle datenbankspezifischen Metainformationen können in einer Konfigurationsdatei zusammengefaßt werden, die unter der angegebenen URL abgelegt ist. Dies hat folgenden Vorteil: das Verzeichnis, in dem diese Datei liegt, kann nur für spezielle Suchmaschinen frei gegeben werden, so daß ein kontrollierter Zugriff auf diese Informationen gewährleistet ist.

- *QueryScript*: Hier kann der Anbieter ein Skript für die Abarbeitung der Anfragen angeben, das anstelle des im Formular spezifizierten Programms verwendet werden soll. Dieses Skript könnte beispielsweise einfachere Anfragen zulassen oder Ergebnisseiten liefern, die leichter zu analysieren sind.

- *QueryElement(s)*: Ein *QueryElement* ist ein für die Anfrage wesentliches Formularelement. Sind mehrere Elemente für eine Anfrage notwendig, dann können diese durch Kommata getrennt in *QueryElements* angegeben werden. Für die Elemente ist der interne Bezeichner zu verwenden. Bei mehreren Elementen ergeben sich die möglichen Anfragebelegungen aus dem kartesischen Produkt der zugeordneten Domänen. Deshalb sollten die notwendigen Anfrageelemente so gewählt werden, daß der Inhalt der Datenbank mit einer minimalen Anzahl von Anfragen bestimmbar ist. Sind z.B. in einem Formular 2 Elemente A und B mit bestimmter Domäne enthalten und B schränkt das Anfrageergebnis von A ein, dann ist nur A anzugeben[7]. Ansonsten müßten alle Kombinationen von A und B angefragt werden, um den gleichen Inhalt zu ermitteln. Weiterhin sollten möglichst Elemente mit bestimmter Domäne verwendet werden, da unbestimmte Domänen nur sehr eingeschränkt beschreibbar sind.

- *QueryElement.<Element>*: Innerhalb einer Domänen können allgemeinere und speziellere Attributwerte auftreten. Die Verwendung eines spezielleren Werts macht in diesem Fall keinen Sinn, da die Ergebnismenge eine Untermenge des Ergebnisses ist, das auf eine Anfrage mit dem zugehörigen allgemeineren Wert geliefert wird[8]. Über diesen Bezeichner können die allgemeineren Attributwerte für eine bestimmte Domäne spezifiziert werden, da eine automatische Erkennung dieser Werte ohne lexikalische Analyse nicht möglich ist.

  Außerdem dient dieser Bezeichner zur Beschreibung von Formularelementen mit unbestimmter Domäne, die in Anfragen benötigt werden. Die meisten unbestimmten Domänen lassen sich leider nicht über reguläre Ausdrücke beschreiben. Die Zuhilfenahme von Ontologien zur Bestimmung möglicher Werte ist aber

[7] Voraussetzung ist, daß die Web-Schnittstelle das Weglassen des Wertes für Element B erlaubt.

[8] Hier wird davon von einer vollständigen Ergebnismenge ausgegangen.

sehr problematisch, so daß in diesem Ansatz versucht wird, zumindest einen Teil unbestimmter Domänen mit regulären Ausdrücken zu erfassen. Viele Web-Anfrageschnittstellen lassen bei der Eingabe Teilausdrücke zu, etwa bei der Suche nach einer Stadt, wo nur die Anfangsbuchstaben angegeben werden müssen, z.B. 'Ros'. Im Ergebnis werden dann alle Städte berücksichtigt, die mit diesen Buchstaben beginnen, wie 'Rosenheim', 'Rostock' usw.. Diese Teilmuster können über reguläre Ausdrücke dargestellt werden.

Ein weiteres Problem ist, das hier kein Vergleich von vorhandenen Zeichenketten mit den regulären Ausdrücken stattfindet, sondern das mit Hilfe dieser Ausdrücke mögliche Werte erzeugt werden sollen. Aus diesem Grund sind hier auch nur sehr einfache reguläre Ausdrücke gestattet, mit denen eine endliche Menge von Werten beschrieben werden kann. Der reguläre Ausdruck '[A-Z]%' wird vom Gatherer beispielsweise expandiert zu 'A%', 'B%', ..., 'Z%', wobei % in diesem Fall für einen erlaubten Platzhalter für das Formularelement steht.

- *Result*: Kommen mehrere Ergebnisattribute an einer bestimmten Stelle im Dokument vor, dann kann der gemeinsame Teil der Pfadausdrücke in den einzelnen Extraktionsregeln weggelassen und in diesem Bezeichner angegeben werden. Dies ist z.B. nützlich, wenn alle Attribute in einer Tabelle vorkommen: der Pfad für die Tabelle wird mit *Result* definiert, so daß die Regeln für die Attribute dann alle relativ zu dieser Angabe gelten.

- *Result.<Attributname>*: Für jedes Attribut, dessen Werte bestimmt werden sollen, ist eine W4F-Extraktionsregel anzugeben. In Abschnitt 3.3 befindet sich eine detaillierte Beschreibung dieser Regeln. Für bestimmte Domänen, bei denen sich die nach außen sichtbaren Werte schlecht ermitteln lassen (Radiobuttons, Checkboxen), können auch alle möglichen Werte durch Kommata getrennt aufgelistet werden.

- *NavigationURL*: Die Angabe der URL, über die eine Navigation durch die Ergebnisseiten erfolgt, mit dem Platzhalter * für die sich ändernden Parameter gewährleistet das korrekte Bestimmen der Ergebnismenge. Ansonsten muß der Gatherer alle Links auf der Ergebnisseite verfolgen oder über Heuristiken[9] versuchen, die weiteren Ergebnisseiten zu ermitteln.

In Abschnitt 2 wurde die Einbindung des CityWorld-Hotelführers für den kooperativen Fall gezeigt. Dieser soll

[9]Das verfolgen aller Links, die mit Zahlen bezeichnet sind, wäre eine mögliche Heuristik.

Abbildung 4: Anfrageergebnis

jetzt über Metadaten eingebunden werden. Das Anfrageformular des Hotelführers (siehe Abbildung 2) hat drei Eingabefelder: Region (interner Bezeicher `region`), Ort (`ort`) und Zimmerpreis (`preis`). Die Felder Region und Preis haben eine bestimmte Domäne, für den Ort können beliebige Angabe gemacht werden. Das wesentliche Anfrageelement ist die Region, d.h. der gesamte Datenbankinhalt läßt sich durch das Abfragen der einzelnen Regionen ermitteln. Die Angabe eines Zimmerpreises schränkt nur die jeweilige Ergebnismenge ein. Die Ergebnisseiten sind entsprechend der in Abbildung 4 enthaltenen Beispielseite aufgebaut. Aus diesen Inhalten sollen die Attribute Region, Ort und Hotel- bzw. Pensionsname extrahiert werden. Die automatische Bestimmung der Domäne für das Feld Region im Anfrageformular ist nicht möglich, da es sich um Radiobuttons handelt. Für die Ermittlung dieser Domäne sind 2 Varianten möglich: entweder werden die Werte aus den Ergebnisseiten extrahiert oder es erfolgt eine Auflistung der Werte im entsprechenden Meta-Tag. Bei den Attributen Ort und Name gibt es keine Wahl: hier ist nur die Angabe von Extraktionsregeln erlaubt. Zur Einbindung des Hotelführers müssen folgende Metadaten angegeben werden:

```
<META NAME="DB.QueryElement" CONTENT="region">
<META NAME="DB.Result.region"
 CONTENT="html.body->A[i]->pcdata[1].txt
 where html.body->A[i].getAttr(name) = 'Seitenanfang'">
<META NAME="DB.Result" CONTENT="html.body->table[1].">
<META NAME="DB.Result.ort" CONTENT="tr[j:*].td[1].txt
 where tr[j].td[1].font.getAttr(color) = '#FF8000'">
<META NAME="DB.Result.name" CONTENT="tr[k:*].td[1].b.txt
 where tr[k].td[0].input.getAttr(name) = 'hotel_id'">
<META NAME="DB.Navigation" CONTENT="NO">
```

Die Extraktionsregeln sind so definiert, daß sie bei kleineren Layoutänderungen nicht geändert werden müssen. Eine Ortsangabe im Ergebnis ist farbig hervorgehoben. Dies wird bei der Bestimmung von Orten ausgenutzt, da nur Angaben in der 2. Spalte als Ort anerkannt werden, die eine solche Hervorhebung haben. Hotels und Pensionen sind im Ergebnis mit einer Checkbox versehen, um Reservierunganfragen an diese zu schicken. Bei der Extraktion von Hotels und Pensionen werden deshalb solche Zeilen gesucht, die in der ersten Spalte eine Checkbox enthalten. Das Ergebnis umfaßt den Namen eines Hotels bzw. einer Pension und weitere Informationen. Da der Name fett hervorgehoben ist, kann er von den übrigen Informationen getrennt werden.

4 Zusammenfassung und Ausblick

In diesem Artikel wurden zwei Ansätze zur Indexierung von Datenbankinhalten durch Suchmaschinen vorgestellt, die von einer unterschiedlichen Kooperationsbereitschaft des Datenbankanbieters ausgehen. Der erste in der Suchmaschine SWING realisierte Ansatz setzt die volle Kooperationsbereitschaft des Anbieters voraus, da ein lokaler Summerizer installiert werden muß. Dieser Mechanismus ist sehr effizient, da nur zwei HTTP-Requests für die Indexierung notwendig sind. Viele Anbieter scheuen sich aber aus unterschiedlichen Gründen davor, ein solches Programm auf ihrem Server zu installieren. Aus diesem Grund wurde ein zweiter Ansatz vorgeschlagen, der die Indexierung der Datenbankinhalte über die normale Anfrageschnittstelle erlaubt. Hier beschreibt der Anbieter über Metadaten, welche Funktionalität diese Schnittstelle bietet und wie die Ergebnisseiten aufgebaut sind. Der Gatherer nutzt die Metadaten für die Generierung von Anfragen und die Extraktion der Attribute, die indexiert werden sollen. Zwei Hauptprobleme treten bei diesem Mechanismus auf: zum einen die Beschreibung unbestimmter Domänen und zum anderen die Vielzahl der Anfragen, die an die Datenbank gestellt werden müssen. Zur Definition unbestimmter Domänen wird momentan auf reguläre Ausdrücke zurückgegriffen. Somit lassen sich nur Domänen beschreiben, die ein bestimmtes Format voraussetzen oder bei denen in der Anfrage einfache Teilmuster verwendet werden können. Weitere ungelöste Probleme sind Interaktionen im Anfrageprozeß bzw. abgeschnittene Ergebnislisten bei zu vielen Treffern.

Als nächster Schritt ist die prototypische Implementierung des Gatherers vorgesehen, um Aussagen über die Leistungsfähigkeit des Ansatzes geben zu können. Dabei ist vor allem interessant, wie sich das Zeitverhalten bei steigendem Datenvolumen und bei steigender Komplexität der Anfrageschnittstelle entwickelt. Dann kann auch die Frage geklärt werden, ob die Einbindung von Datenbanken überhaupt Sinn macht, wenn Formularelemente mit unbestimmten Domänen beim Abfragen der Datenbankinhalte berücksichtigt werden müssen.

Literatur

[1] M. K. Bergmann et al. The Deep Web: Surfacing Hidden Value. Whitepaper, BrightPlanet.com LCC, July 2000. Available at `http://www.completeplanet.com/tutorials/deepweb/`.

[2] I. Bruder. Zugriff auf dynamische Web-Dokumente mittels Web-Mining-Analysetechniken. Preprint CS-04-01, Fachbereich Informatik, Universität Rostock, Mar. 2001.

[3] R. Gohla. Integrierte WWW-Anfragesichten. Master's thesis, Fachbereich Informatik, Universität Rostock, 2000.

[4] D. R. Hardy, M. F. Schwartz, and D. Wessels. Harvest User's Manual. Technical Report CU-CS-743-94, University of Colorado, Boulder, Jan. 1996. Available at `http://www.tardis.ed.ac.uk/harvest/docs/`.

[5] A. Heuer and G. Weber. SWING: Eine Suchmaschine mit Datenbankanschluß. In *Workshop Internet-Datenbanken*, number 12 (Preprint), Magdeburg, Sept. 2000. Fakultät für Informatik, Otto-von-Guericke Universität.

[6] S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 280(5360):98–100, 1998.

[7] S. Raghavan and H. Garcia-Molina. Crawling the Hidden Web. Technical report, Stanford University, Nov. 2000.

[8] A. Sahuguet and F. Azavant. WysiWyg Web Wrapper Factory (W4F). Technical report, University of Pennsylvania und Telecom Paris (E.N.S.T.), 1998. Available at `http://db.cis.upenn.edu/W4F/`.

[9] A. Sahuguet and F. Azavant. *World Wide Web Wrapper Factory: User Manual*, 2000. Available at `http://db.cis.upenn.edu/W4F/`.