PORTAL – ein System zur Informationsextraktion und -aufbereitung im „World Wide Web“

Praktische Hausarbeit im Fach Linguistische Datenverarbeitung
Autor: Johannes Kiehl
Seminar: Textklassifikation und Informationsextraktion (SS 1999)
Leitung: Dr. H.-J. Weber

Übersicht

Die Trierer Studentenzeitung wünscht sich eine „dynamische“ Begrüßungsseite (portal site), die bei jedem Abruf vom Server aktualisiert oder mit vom Benutzer gewünschten Informationen ergänzt werden kann.

Dafür wurde das System PORTAL entwickelt, das zu festgelegten Zeiten aus verschiedenen Datenquellen im Internet Nachrichten und Informationen abruft, auswertet und zu einer „Bildschirmzeitung“ zusammenstellt. Diese ist in grafische „Module“ eingeteilt, die folgende Themenbereiche abdecken:

- Wettervorschau für den Tag
- Zeitungsschlagzeilen des Tages über Uni und Stadt Trier
- Speiseplan der Mensa für den Tag
- Abfahrtszeiten der nächsten 8 Busse
- Aktuelle Kleinanzeigen aus dem „Gopher“-Angebot des Uni-Rechenzentrums

Das System wurde modular in der Programmiersprache „Perl“ entwickelt. Es besteht aus einer Online- und einer Offlinekomponente sowie einer kleinen Datenbank. Die Onlinekomponente fungiert als cgi-Server, der auf die Seitenabrufe durch die Benutzer reagiert und mit Inhalten aus der Datenbank dynamisch eine HTML-Seite generiert. Die Offlinekomponente ist das eigentliche Content-Pflege- und Managementsystem, das nach einem festgelegten Zeitplan Inhalte abruft, auswertet und sie als Datenbank bereitstellt.

Im Rahmen dieser Hausarbeit soll nur das Offline-System im Detail vorgestellt werden.

Ein HTML-Frontend, mit dem das System getestet werden kann, ist derzeit unter der Adresse

http://ldv33.uni-trier.de:8080/~kiehl/hst/

installiert. In näherer Zukunft ist eine Installation im Web der Campuszeitung unter

http://www.nuweb.de/

vorgesehen.

1. Systemarchitektur

PORTAL besteht aus einer Offline- und einer Online-Komponente und einer Datenbank, auf die beide Komponenten gemeinsam zugreifen. Da die Online-Komponente nie schreibend zugreift, konnte für die Datenbank vorerst eine einfache Listenform mit SGML-Tags gewählt werden. Im Konfliktfall schlägt das Einlesen durch die Online-Komponente fehl. Diese gibt entsprechende Hinweise aus, dass die Seite durch den Benutzer neu aufgerufen (reload) werden muss. (vgl. Tabelle 1.1).

Architektur des HST-Systems

SGML-Datenbank		
Online-Komponente: hstWeb.pl	Offline-Komponente: hstOffline.pl	
Layoutmodul: hstPage.pm	Internetmodul: hstRequest.pm	Auswertungsmodul: hstContent.pm

Tabelle 1.1: Systemarchitektur

Die Offline-Komponente stellt, mit Hilfe der Perl-Bibliothek „libwww“[1], nacheinander die Verbindung zu den gewünschten Datenquellen (URLs) her und liest deren Inhalte (im HTML-Format) ein. Die entsprechenden Funktionen sind im Modul hstRequest.pm gekapselt. Danach werden die Inhalte mit Funktionen aus dem Auswertungsmodul, hstContent.pm, gefiltert und schließlich in die Datenbank eingetragen.

Die Online-Komponente wird über den cgi[2]-Mechanismus jedesmal aktiviert, wenn ein Benutzer die Portal-Seite aufruft. Sie generiert HTML-Code, der dem Benutzer genauso präsentiert wird, als ob er eine statische HTML-Seite aufgerufen hätte.

Das Online-Modul (hstWeb.pl) liest eine Layout-Datei ein, in der die Platzierung und Größe der einzelnen Nachrichtenmodule[3] eingetragen ist. Aufgrund dieser Daten und, sofern diese Information verfügbar ist[4], der Fenstergröße des Anzeigeprogramms (Browser), wird dann eine HTML-Seite mit den entsprechenden grafischen Elementen (z.B. Scrollbalken, wenn ein Nachrichtenmodul aus Platzgründen nicht vollständig dargestellt werden kann) generiert und die Inhalte aus der Datenbank eingetragen. Die Objekte „PAGE“ und „BOX“, die das Seitenlayout repräsentieren und schließlich erzeugen, sind im Modul hstPage.pm gekapselt.

2. Beschreibung der einzelnen Nachrichten-Module

Die Funktionen „Wettervorschau“, „Schlagzeilen“, „Mensa“, „Busfahrplan“ und „Schwarzes Brett“ haben eine bestimmte „Inhaltserwartung“, die aufgrund der dargebotenen Daten möglichst genau erfüllt wird. Sie

[1] http://www.sn.no/libwww-perl/

[2] Common Gateway Interface: Ein Mechanismus, der den Aufruf ausführbarer Dateien über das WWW erlaubt, vgl. http://www.apache.org

[3] Der Begriff „Modul“ im Zusammenhang mit Seitengestaltung ist der Terminologie der Zeitungsgestaltung entnommen und bezeichnet ein rechteckiges Element auf der Seite. Ein Modul in diesem Sinne kann etwa ein Bild, einen Kommentar-Kasten, eine Grafik oder einen Artikel enthalten.

[4] Beim Seitenaufruf kann die Fenstergröße als Argument übergeben werden. Dieses Feature können jedoch nur bestimmte Web-Browser nutzen.

komplettieren also jeweils spezifische semantische „Frames“, mittels spezifischer heuristischer Verfahren, aus den festgelegten Informationsquellen. Im Falle der „Schlagzeilen“-Funktion müssen inhaltsähnliche Texte erkannt und ausgefiltert werden.

2.1 Wettervorschau

Das System erhält die Wettervorschau der Mess-Station Nürburgring des Deutschen Wetterdienstes in Form einer Tabelle. Diese wird mit Hilfe der Funktion parseContent in eine einfache Liste heruntergebrochen, aus der durch Schlüsselwortsuche die relevanten Informationen herausgelesen werden:

> Eine verbale Beschreibung des aktuellen Zustands („sonnig“),
>
> Kurze, sprachlich gefasste Vorschauen für den Vor- und Nachmittag des folgenden Tages.
>
> Temperaturvorschauen für den folgenden Tag

Um diese in einer Zeile zusammenfassen zu können, werden die verbalen Vorschauen in (klein geschriebene) Phrasen umgewandelt, es sei denn, sie beginnen mit einem der im konstanten Feld weatherCaseNouns definierten Nomina:

```
%weatherCaseNouns = (Regen,",Sonne,",Hagelschauer,",Schauer,",
Gewitter,", Schneefall,",Regenschauer,",Nebel,");
```

Sind die Vorschauen gleich, so werden sie in einem Satz („Morgen ganztägig Regen.“) zusammen-gefasst.

2.2 Schlagzeilen

Das System stellt einmal täglich eine Suchanfrage über Trier[5] an einen elektronischen „Ausschnittdienst“ (www.paperball.de). Dieser liefert jeweils eine Schlagzeile, den Textanfang, den Namen der Zeitung sowie einen Link, über den der Rest der Nachricht direkt bei der betreffenden Zeitung abgerufen werden kann. Der folgende Ausschnitt zeigt das Ausgabeformat des Dienstes:

Alle Rubriken Artikel 1-10 von 48 Treffern

Heilbronner Stimme Sport 27.9.2000 16:7
Fussball
Regionalliga-Statistik
  Süd FC Bayern München Am. - TSV 1860 München Am. 0:0
Zuschauer: 1500. Gelb-Rote Kartte: Diara (Bayern/75. Foulspiel). VfB
Stuttgart Am. - FC Car ...
Nur Artikel aus dieser Zeitung zeigen | Keine Artikel aus dieser Zeitung zeigen

Heilbronner Stimme Sport 27.9.2000 16:4
Handball
Ergebnisse und Tabellen
  Handball Bundesliga Frauen, 2. Spieltag: Germania List - BVB
Dortmund 18:26, Buxtehuder SV - SG Minden/Minderheide 22:16, TV
Lützellinden - SG He ...
Nur Artikel aus dieser Zeitung zeigen | Keine Artikel aus dieser Zeitung zeigen

Bocholter Borkener Volksblatt Kultur 27.9.2000 15:50
Rock \& Pop Termine
Comet wird auf der Expo verliehen
Frankfurt/New York (AP). Wer die Mischung aus deftigem Rock und
heftigem Rap von SUCH A SURGE mag, hat in den kommenden ...
Nur Artikel aus dieser Zeitung zeigen | Keine Artikel aus dieser Zeitung zeigen

Die Ausgangsdaten werden mit Hilfe der Funktion parseContent in eine Liste von Ausdrücken der Form „Klartext CTX: HTML-Kontext“ gebracht. Für den HTML-Kontext kann spezifiziert werden, ob dieser auch Links (<a href>-Tags) enthalten soll; im Fall der Schlagzeilen wird hiervon Gebrauch gemacht – schließlich soll den Benutzern ja auch das Weiterklicken zur Nachrichtenquelle möglich sein.

Der folgende Ausschnitt zeigt das Ergebnis der Funktion (zu obigem HTML-Text):

Alle Rubriken CTX:font
Artikel 1-10 von 48 Treffern CTX:font
Sport CTX:font
16:7 CTX:i
Heilbronner Stimme 27.9.2000 CTX:font
Fussball CTX:a /service/forward-paperball.fcg?
action=aforward&rurl=%2Fservice%2F
paperball%2Dall%2Efcg&furl=http%3A%2F%2Fwww%2Estimme%
2Ede%2Fsport%2F
nachrichten%2Fartikel%2Findex%2Ecfm%3Fid%3D%
2D1039525477&ahdl=Fussball
&cat=Sport&anp=heil
Regionalliga-Statistik CTX:b
Süd FC Bayern München Am. - TSV 1860 München Am. 0:0
Zuschauer: 1500. Gelb-Rote Kartte: Diara (Bayern/75 Foulspiel). VfB
Stuttgart Am. - FC Car ... CTX:font
Nur Artikel aus dieser Zeitung zeigen CTX:a /service/paperball.fcg?
action=query&pg=detail&fmt=.&r=trier&q=%2Btrier+%2Dlars&stq=1
&d0=&d1=&what=german_web&pp=heil&rankedBy=date&categories
=all&papers=all

…

Häufig übernehmen Tageszeitungen Artikel von einer Agentur. Auf diese Weise treten, gerade im Angebot von paperball.de, zahlreiche Dubletten und inhaltlich stark ähnliche Texte auf. Um solche Artikel ausfiltern zu können, wurde ein einfaches Maß zur Textdistanz entwickelt. Das System führt mit diesem Distanzmaß eine Clusteranalyse durch und markiert alle diejenigen Texte, die auf einem bestimmten (heuristisch festgelegten) Clusterniveau als „ähnlich“ ermittelt wurden.

Als Distanzmaß wurde die prozentuale Übereinstimmung des Lexikons gewählt. Um die zunächst wenig befriedigenden Ergebnisse zu verbessern, wurde

a. das Lexikon auf „Inhaltswörter“ reduziert, und
b. jedes Wort des Lexikons mit einem numerischen Gewicht [0,1] ergänzt, das seine Nähe zum Textanfang ausdrückt.

Mit diesen Ergänzungen arbeitete das System zufriedenstellend.

Zu a.: Die Unterscheidung „Inhalts-“ vs. „Funktions-wörter“ übernimmt die Funktion „isFWord“, die auf eine Liste aus derzeit 93 hochfrequenten Wörtern zugreift. Diese wurden durch Auswertung eines Märchen-Textes (ca. 15.000 Token) ermittelt (5% häufigste Types).

Zu b.: Journalistische Texte weisen eine strukturelle Besonderheit auf: Das wichtigste kommt in der Regel zuerst. Die Gewichte machen es möglich, auf diese Besonderheit Rücksicht zu nehmen. So wird die Distanz zweier Texte stärker durch eine lexikalische Übereinstimmung zwischen etwa dem ersten und dem

[5] Die Anfrage lautet „+Trier –„Lars von“. Damit soll die oft große Zahl von Meldungen über den dänischen Filmregisseur gleichen Namens unterdrückt werden. Auf diese Weise kann die Spezifizität der Ausgangsdaten im Laufe der Zeit noch feiner adjustiert werden, wenn dies nötig erscheint.

siebten Wort reduziert, als durch eine Übereinstimmung zwischen etwa dem 27. und dem siebten.

Zwei Faktoren wurden heuristisch bestimmt:

BIASFC = 1.41

EQUITHRESH = 0.2

BIASFC gibt den Stellenwert der Gewichte bei der Distanzberechnung an. Ein Wert von 1 bedeutet „kein Stellenwert“, ein Wert von 2 „Erhöhe den Stellenwert des ersten Wortes um Faktor 2“.

EQUITHRESH stellt die Ebene dar, oberhalb derer zwei Cluster als getrennt angesehen werden. Die Clusteranalyse wird also auf der Ebene 0.2 abgebrochen.

Die verbleibenden Texte werden in Schlagwort, Medium, Datum, URL, Überschrift, Unterüberschrift und Textbeginn zerlegt und in die Datenbank eingetragen. Der folgende Ausschnitt zeigt den entsprechenden Ausschnitt aus der Datenbank:

```
<NEWS TIME=20000927:1622>
<KEY TIME=20000927:1622>Kultur</KEY>
<MED TIME=20000927:1622>Bocholter Borkener Volksblatt</MED>
<DAY TIME=20000927:1622>27.9.2000</DAY>
<LNK TIME=20000927:1622>http://www.bbv-
net.de/news/kultur/2000-0927/rock_termine.html</LNK>
<HDL TIME=20000927:1622>Rock \&amp;amp; Pop Termine</HDL>
<SHD TIME=20000927:1622>Comet wird auf der Expo
verliehen</SHD>
<BDY TIME=20000927:1622>Frankfurt/New York (AP). Wer die
Mischung aus deftigem Rock und heftigem Rap von SUCH A SURGE
mag, hat in den kommenden ...</BDY>
```

Dieses Prinzip erhält das System offen für weitere Datenquellen, die dann auch untereinander auf Übereinstimmungen hin abgeglichen werden können.

2.3 Mensa-Speiseplan

Wie die Nachrichten wird auch der Mensaspeiseplan als HTML-Seite eingelesen, auf eine Liste heruntergebrochen und schließlich analysiert.

In einem ersten Analyseschritt werden mögliche Hauptgerichte aufgrund der Nähe zu Tabellen-Tags oder der Schriftauszeichnung identifiziert und mit dem Schlüssel HGH:[6] versehen. Der zweite trifft die abschließende Entscheidung über die auszugebenden Hauptgerichte. Dabei muß eines der folgenden Kriterien erfüllt werden:

(1) Anfangsbuchstabe versal

(2) Erstes Wort beginnt auf „ge“ oder „vege“

Anschließend werden eine Reihe von Kürzungsregeln abgearbeitet:

```
(1) ($wline =~ /(.*) dazu/) && ($wline = $1);
(2) ($wline =~ /(.* mit .*) oder/) && ($wline = $1);
(3) ($wline =~ /(.* mit .*) und/) && ($wline = $1);
(4) ($wline =~ /(.*) mit \w+ \w+ \w+/) && ($wline = $1);
(5) ($wline =~ /(.*) nach/) && ($wline = $1);
(6) ($wline =~ /(.*) Dessert/) && ($wline = $1);
(7) ($wline =~ /(.*) \(/) && ($wline = $1);
```

Diese kürzen Hinweise nach „dazu“ (1), Ergänzungen zu Mahlzeiten, die bereits ein „mit“ enthalten (2,3), sie entfernen komplexe Angaben mit drei oder mehr Wörtern (4), Angaben nach „nach“, „Dessert“ oder in Klammern (5,6,7). Schließlich werden die Zusatzangaben „Stammessen“, „Wir verwöhnen Sie“ und „Fitmenü“ identifiziert.

Tabellen 2.1 bis 2.3 zeigen beispielhaft die HTML-Eingabe, das annotierte Zwischenergebnis (Ausgabe der Funktion mensaMacro) und schließlich den resultierenden Datenbankeintrag:

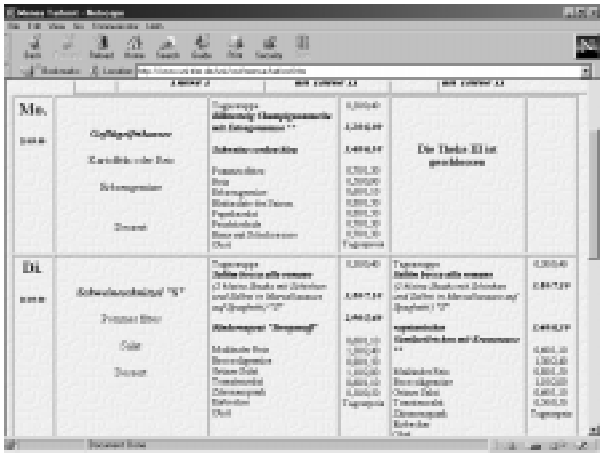

Tabelle 2.1: Speiseplan, HTML-Eingabe

```
MAGIC_DAY:0Mo.
MAGIC_DATE:25.09.00
Geflügelfrikassee
Blätterteig Champignontasche mit Estragonsauce
Schweine cordon bleu
MAGIC_DAY:1Di.
MAGIC_DATE:26.09.00
Schweineschnitzel
Saltim bocca alla romano
Rinderragout "Stroganoff
...
```

Tabelle 2.2: Speiseplan, Ausgabe von „mensaMacro“

```
<MENSA TIME=20000928:1536>
<DAY TIME=20000928:1536>1</DAY>
<DATE TIME=20000928:1536>20000925</DATE>
<DISH TIME=20000928:1536>Gefl&uuml;gelfrikassee </DISH>
<DISH TIME=20000928:1536>Bl&auml;tterteig
Champignontasche mit Estragonsauce </DISH>
<DISH TIME=20000928:1536>Schweine cordon bleu </DISH>
<DAY TIME=20000928:1536>2</DAY>
<DATE TIME=20000928:1536>20000926</DATE>
<DISH TIME=20000928:1536>Schweineschnitzel </DISH>
<DISH TIME=20000928:1536>Saltim bocca alla romano </DISH>
<DISH TIME=20000928:1536>Rinderragout &quot;Stroganoff
</DISH>
...
```

Tabelle 2.3: Speiseplan, Datenbankeintrag

2.4 Busfahrplan, Schwarzes Brett

Um den Stadtbus-Fahrplan für den ganzen Tag zu ermitteln, müssen mehrere Anfragen an den Internet-Dienstleister der Rheinland-Pfälzischen Verkehrsunternehmen gesandt werden (Funktion busCollect, Modul hstContent). Die Auswertung erfolgt mittels eines Automaten (state machine), der die Fahrplandaten in Blöcken von Abfahrtszeiten und Buslinie/Fahrtziel erwartet.

Beim Schwarzen Brett (gopher.uni-trier.de/Brett) liegt eine strenge Formatierung vor, so dass eine Auswertung mittels Mustervergleich (regular expression matching) möglich ist.

[6] Haupt-Gerichts-Hypothese