

SWING: Die Suchmaschine des Landesinformationssystems MV-Info

Andreas Heuer Holger Meyer Gunnar Weber

Lehrstuhl Datenbank- und Informationssysteme
Fachbereich Informatik, Universität Rostock
Postadresse: D-18051 Rostock
Tel.: 0381/498-3401, -3402 und -3405
Fax: 0381/498-3443
E-Mail: swing@informatik.uni-rostock.de

Abstract. SWING ist der Anfrage- und Suchdienst des Landesinformationssystems MV-Info, das die Internet-Informationsangebote Mecklenburg-Vorpommerns integriert und klassifiziert. SWING umfaßt neben klassischen Suchtechniken auch diverse neue Merkmale wie Konsistenzchecks, Profildienste und insbesondere das Ausnutzen von Strukturen in Informationen, die aus Datenbanken stammen. In diesem Beitrag wird ein Überblick über die Weiterentwicklung von SWING gegeben und ein Konzept für ein Autorenwerkzeug vorgestellt, das die redaktionelle Arbeit für MV-Info erleichtern soll.

1 Einführung

Das Projekt SWING (Suchdienst für WWW-basierte Informationssysteme der nächsten Generation) ist Teil des Landesinformationssystems Mecklenburg-Vorpommern (MV-Info), das verschiedene Unternehmen und Forschungseinrichtungen des Landes entwickeln und betreiben. MV-Info bietet eine einheitliche Schnittstelle auf Informationen, die im Internet über Firmen, Institutionen und Personen des Landes angeboten werden. Ziel des u.a. auf der CeBIT 99 vorgestellten Teilprojekts SWING ist es, diese Informationen strukturierter zu verwalten und insbesondere schneller auffindbar zu machen. Der Kern des Projekts ist die Entwicklung einer neuartigen Suchmaschine, die Datenbanktechniken (die in vielen firmeninternen Informationssystemen seit Jahren angewendet werden) und die Suche in elektronisch bereitgestellten Texten intelligent miteinander verbindet. SWING hat gegenüber den weltweit im WWW (World Wide Web) verfügbaren internationalen Suchmaschinen diverse neue Merkmale:

- Beschränkung der durchsuchten Informationen auf die Region Mecklenburg-Vorpommern.
- Verteilte, parallele (und damit weitaus schnellere) Suche nach Informationen auf verschiedenen Rechnern im Land (Gatherer). Danach Zusammenführung der Teilergebnisse (Broker).
- Zugriff auf strukturierte Datenbanken: im Gegensatz zu allen anderen Suchmaschinen werden nicht nur Texte, sondern auch Datenbanken durchsucht. Letztere werden in sogenannte "dynamische WWW-Seiten" eingebunden, die bisher von jeglicher Suche ausgeschlossen waren.
- Spezielle Meta-Tags wie *official-homepage* führen zu einer höheren Ergebniswichtung bei der Bewertung der zu einem Suchbegriff gefundenen Informationen. Bei der Generierung dieser Tags wird der Nutzer durch ein eingebautes MetaTool unterstützt.
- Profildienste: durch die Kenntnis von Nutzerspezifika können dem Anwender bestimmte Suchumgebungen bzw. zusätzliche Dienste bereitgestellt werden. Dazu zählen das Abonnieren von Anfragen durch den Nutzer und ein URL-Reminder, der den Anwender über Veränderungen in wichtigen WWW-Dokumenten informiert.
- Benutzerunterstützung durch Ausnutzung von linguistischem Wissen: ein Wörterbuch stellt alternative Begriffe wie Synonyme, allgemeinere und speziellere Begriffe bereit.
- Überprüfung aller durchsuchten Informationen durch regelmäßige Konsistenzchecks.

In diesem Beitrag sollen insbesondere neuere Entwicklungen im Projekt SWING II vorgestellt werden, die im Laufe des vergangenen Jahres konzipiert und umgesetzt wurden. Beispielsweise unterstützt ein MetaTool die Definition sogenannter Meta-Tags, die WWW-Seiten so beschreiben, daß sie von SWING und anderen Suchmaschinen (wie Altavista) besser gefunden werden. Ein weiteres Werkzeug ermöglicht die Einbindung mehrerer, autonom bei verschiedenen Anbietern gehaltenen heterogenen Datenbanken in die Suche. Die Profildienste der SWING-Suchmaschine zur Benachrichtigung über Neuerungen in bestimmten Bereichen des Landesinformationssystems werden ebenfalls vorgestellt.

Das Landesinformationssystem ist als Katalog aufgebaut, d.h. der Nutzer kann in hierarchisch aufgebauten Themengebieten navigieren. Ein Informationsanbieter kann sich in verschiedenen Gebieten registrieren lassen, die er bei der Anmeldung seines Dokumentes bestimmen kann. Anschließend erfolgt dann die redaktionelle Bearbeitung durch den Verantwortlichen für das entsprechende Themengebiet. Um diese Arbeit zu vereinfachen, soll in nächster Zeit ein Redaktionssystem entstehen, daß die Verwaltung von Inhaltsverzeichnissen (strukturierten Linksammlungen) im Internet ermöglicht.

Die SWING-Suchmaschine ist über die Internet-Adresse <http://swing.m-v.de> oder <http://swing.informatik.uni-rostock.de> für jeden Internet-Nutzer zugänglich. Für nähere Informationen zu technischen Details sei auf [LDHM97] verwiesen, eine Kurzfassung wurde auf einem nationalen Workshop als [DLHM97] veröffentlicht. Auf den ersten IuK-Tagen des Landes-Mecklenburg-Vorpommern wurde die Funktionalität der ersten Version der Suchmaschine (SWING I) vorgestellt [HMDL97]. Die dort bereits beschriebenen Features werden hier nicht wiederholt. Für weitere Informationen zu SWING I sei daher auf die Originalartikel verwiesen.

2 Meta-Daten und MetaTool

2.1 Definition und Nutzen von Meta-Daten

Mit Hilfe von Meta-Daten kann man bestimmte Eigenschaften eines HTML-Dokuments beschreiben, die nicht vom WWW-Browser angezeigt werden. Suchmaschinen können diese Informationen interpretieren und sie bei der Bewertung des Dokuments bezüglich der Suchanfrage sowie bei der Ausgabe des Suchergebnisses verwenden. Meta-Daten werden mit Hilfe von *Meta-Tags* definiert, die folgende allgemeine Syntax haben:

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

Diese Meta-Tags müssen im Kopf eines HTML-Dokuments stehen. Damit auch Anwender, die nicht mit HTML vertraut sind, die Vorteile von Meta-Daten in ihren Dokumenten nutzen können, wurde ein MetaTool entwickelt, das die Generierung der Meta-Tags übernimmt.

Es existieren verschiedene Meta-Tags, von denen für Suchmaschinen das *description*- und das *keywords*-Tag die größte Bedeutung haben. Das *description*-Tag enthält eine Beschreibung der HTML-Seite, die von den Suchmaschinen bei der Ausgabe des Dokuments im Suchergebnis verwendet wird. Fehlt dieses Tag, generieren viele Suchmaschinen eine Zusammenfassung des Dokuments, die in den meisten Fällen aber wenig aussagekräftig ist. Wenn im *keywords*-Tag Suchbegriffe angegeben sind, unter denen das Dokument gefunden werden soll, dann erhöht sich nicht nur die Trefferwahrscheinlichkeit, sondern auch der Rang der Seite in der Ergebnisliste.

SWING sieht außerdem die Meta-Tags *official-homepage* und *official-homepage-aliases* vor, die zur Bewertung eines Dokuments herangezogen werden. Mit dem *official-homepage*-Tag werden die in relationalen Datenbanken bekannten Primärschlüssel realisiert: in einer Datenbank kann eine Eigenschaft von Objekten angegeben werden, die diese eindeutig identifiziert (wie Personalausweisnummer für Personen und Matrikelnummer für Studenten). Damit eignet sich dieses Tag für die Kennzeichnung des Einstiegspunktes in eine Dokumentenhierarchie, von dem aus jedes Dokument in der Hierarchie über Links erreichbar ist, und es sollte auf der Startseite einer Firma oder Institution mit der offiziellen Bezeichnung (Firmen- bzw. Institutsname) hinterlegt werden. Die Definition des Tags *official-homepage* hat zwei Vorteile:

- Findet SWING eine zweite Seite in MV-Info mit dieser Kennung, wird auf die Nicht-Eindeutigkeit aufmerksam gemacht.
- Findet SWING mehrere Seiten mit dem Firmen- bzw. Institutsnamen als Suchbegriff, dann wird die so gekennzeichnete “official-homepage” am höchsten bewertet und somit als erstes angezeigt.

Weitere Schreibweisen für die *official-homepage*-Angabe (z.B. vollständig ausgeschriebener Firmenname oder mehrdeutige Abkürzung) können mit dem Tag *official-homepage-aliases* angegeben werden.

Die Meta-Daten sind bei der SWING-Suche abfragbar. Somit wird das Anfrageergebnis reduziert und die Trefferrelevanz der zurückgelieferten Dokumente erhöht sich.

2.2 Einfluß der Meta-Daten auf die Bewertung eines Dokumentes

Das Ergebnis einer Anfrage ist eine Liste von bewerteten Dokumentreferenzen. Diese Bewertung wird auch als Ranking bezeichnet. In diesem Abschnitt werden die Regeln vorgestellt, nach denen die SWING-Suchmaschine die Relevanz eines Dokumentes bezüglich der Suchanfrage bewertet. Im Mittelpunkt der Betrachtung soll dabei die Beantwortung der Frage stehen, wie Meta-Daten die Bewertung eines Dokumentes beeinflussen.

Der Ranking-Algorithmus, der für die Suchmaschine implementiert wurde, weist jedem einzelnen Dokument aus der Menge der mit SWING gefundenen Dokumente eine Wertigkeit zu. Die Bewertung kann maximal 100% betragen und wird folgendermaßen bestimmt:

- *Datenbank-Anfrageformular*: Handelt es sich bei dem Dokument um ein Datenbank-Interface in Form eines Anfrageformulars, dann wird dieses Dokument mit 100% bewertet (alle weiteren Regeln werden in diesem Fall nicht betrachtet). Somit erscheinen Datenbank-Anfrageformulare immer am Anfang der Ergebnisliste, basierend auf der Annahme, daß man durch die größeren und vor allem strukturierten Datenmengen in den zugrundeliegenden Datenbanken die besten Suchergebnisse erreicht.
- *Qualität des Dokumentes*: Die Qualität eines Dokumentes ist durch das Vorhandensein von Meta-Tags gekennzeichnet. Ist einer der Suchbegriffe in einem Meta-Tag enthalten, dann wird das Dokument höher bewertet als andere, wobei die Einstufung am höchsten ist, wenn der Suchbegriff im *official homepage*-Tag vorkommt. Desweiteren werden HTML-Dokumente mit dem Namen *index.html* höher gewertet als andere, da diese in der Regel Einstiegsseiten für Dokumentensammlungen darstellen.
- *Vorkommen des Suchbegriffs im Titel*: Dokumente, bei denen der Suchbegriff im Titel vorkommt, erhalten eine höhere Bewertung als andere.
- *Häufigkeit des Suchbegriffs*: Bezüglich des gefundenen Dokumentes wird das anzahlmäßige Auftreten des Suchbegriffs ausgewertet. Ein Dokument wird hoch eingestuft, wenn der Suchbegriff häufig im Dokument vorhanden ist.
- *Pfadlänge*: Der Pfadname eines gefundenen Dokumentes wird hinsichtlich seiner Länge analysiert. Kurze Pfadnamen werden höher bewertet als lange, da die Länge des Pfadnamen die Hierarchie auf dem Server widerspiegelt, d.h. je kürzer der Pfadname, desto höher ist das Dokument in der Serverhierarchie angeordnet.
- *Pivotisierung*: Alle Dokumente auf den Servern des Landesinformationssystems erhalten eine höhere Bewertung als andere, da es sich bei MV-Info wie bereits erwähnt um ein Katalogsystem handelt. Somit können in dem gefundenen Dokument weitere interessante Links enthalten sein, die vielleicht nicht den Suchbegriff enthalten, aber zum gesuchten Themenkreis passen.

Einige dieser Bewertungsregeln werden von den meisten Suchmaschinen verwendet (Vorkommen des Suchbegriffs im Titel, Häufigkeit des Suchbegriffs), andere wurden speziell für die SWING-Suchmaschine eingeführt.

Einige der genannten Bewertungskriterien sind von jedem Informationsanbieter in seinen eigenen Dokumenten realisierbar. Es ist somit möglich, Dokumente so zu gestalten, daß sie zu bestimmten Stichworten durch den Ranking-Algorithmus hoch bewertet werden. Dies soll im folgenden an einem Beispiel demonstriert werden. Dazu wurde die Homepage des Lehrstuhls zweimal kopiert und unter den Pfaden `/CeBit99-1/index.html` bzw. `/CeBit99-2/index.html` auf dem WWW-Server `wwwdb.informatik.uni-rostock.de` abgelegt. Das erste Dokument enthält das *official-homepage*-Tag mit dem Wert "Lehrstuhl Datenbank- und Informationssysteme". Ansonsten stimmen beide Dokumente überein. Sucht man mit SWING nach Dokumenten, die die Phrase "Datenbank- und Informationssysteme" enthalten, dann erscheint das Dokument unter `/CeBit99-1/index.html` an erster Stelle mit 100%, während das andere Dokument erst an der dreizehnten Position mit dem Rankingwert 39% auftaucht.

Warum wurden die beiden Dokumente so unterschiedlich bewertet, obwohl sie doch fast identisch sind? Nach dem obigen Ranking-Algorithmus liefern alle Bewertungsregeln bis auf eine Ausnahme die gleichen Resultate für die Dokumente. Unterschiedliche Resultate ergeben sich nur bei der Bewertung der Qualität der beiden Seiten, da der Suchbegriff "Datenbank- und Informationssysteme" im *official-homepage*-Tag des ersten Dokuments enthalten ist. An diesem Beispiel sieht man, daß ein Dokument, auch wenn es noch so gut zur Suchanfrage paßt, nicht so hoch bewertet (und das nicht nur bei SWING, sondern auch bei allen anderen gebräuchlichen Suchmaschinen wie Altavista) wird, wenn die entsprechenden Meta-Daten fehlen. Aus diesem Grund wurde für die SWING-Suchmaschine ein Werkzeug entwickelt, das die Definition von Meta-Tags auf sehr einfache Art und Weise erlaubt.

2.3 MetaTool

Das MetaTool gestattet dem Anwender das Erzeugen, das Ändern sowie das Löschen von Meta-Tags für eine bestimmte HTML-Seite. Für die HTML-Seite, die analysiert werden soll, kann entweder der URL angegeben werden, oder es kann auch ein Upload der HTML-Datei vom lokalen Dateisystem des Anwenders erfolgen. Ein Upload der HTML-Quelle wird angeboten, da der HTML-Code eines Dokuments von einigen WWW-Server bei der Bereitstellung des Dokumentes verändert werden kann (z.B. Roxen Challenger). Bei der Abspeicherung des Dokumentes mit dem neu erzeugten Meta-Datenblock würde das Dokument dann den vom WWW-Server modifizierten HTML-Code enthalten, was in den meisten Fällen aber sicherlich nicht gewünscht ist. Mit dem MetaTool können folgende Meta-Tags definiert werden:

- *language*: Sprache, in der das Dokument vorliegt.
- *official-homepage*: Offizielle Bezeichnung des Dokuments innerhalb des Landesinformationssystems, die eindeutig sein sollte. Die Eindeutigkeit dieses Tags kann über eine eingebaute Funktion getestet werden.
- *official-homepage-aliases*: Weitere Bezeichnungen für das Dokument.
- *theme*: Themengebiet(e) des Landesinformationssystems, denen das Dokument zugeordnet werden kann.
- *author*: Autor des Dokuments.
- *keywords*: Stichwörter, unter denen das Dokument gefunden werden soll.
- *description*: Beschreibung des Dokuments, die von der Suchmaschine bei der Ausgabe des Dokuments verwendet werden kann.

Weitere Meta-Daten, die im Dokument bereits definiert sind, können ebenfalls verändert bzw. gelöscht werden. Die Meta-Tags *language*, *author*, *keywords*, *description* werden auch von vielen anderen Suchmaschinen verwendet, die anderen sind speziell für die SWING-Suchmaschine eingeführt worden.

Nach der Angabe der Meta-Daten wird dann entweder das gesamte Dokument neu erzeugt oder nur ein Meta-Datenblock generiert, der anschließend über die Copy&Paste-Funktion des WWW-Browsers in das HTML-Dokument eingefügt werden kann.

3 Einbindung heterogener Datenbanken in die Suche

Damit die Suchmaschine systembekannte Datenbanken referenzieren kann, müssen ihr die Datenbank-Inhalte bekannt gemacht werden. In der ersten Version von SWING [HMDL97] wurden Datenbanken über *Schemainformationen* angesprochen: Wurde beispielsweise der Name einer in SWING eingebundenen Datenbank oder Datenbanktabelle (etwa `Hotel`) im Suchausdruck entdeckt, so wurde das zugehörige Datenbankanfrageformular angesprochen. In [HMDL97] wurden verschiedene Möglichkeiten aufgezeigt, wie auf diese Art gefundene Datenbanken dann technisch an die SWING-Suche angekoppelt werden können.

Die in SWING II vorgenommene Erweiterung erlaubt es nun sogar, nach *Datenbank-Inhalten* die zu einem Suchauftrag passende Datenbank zu finden. Es existieren bereits sehr einfache Mechanismen, die es Suchmaschinen erlauben, Datenbank-Inhalte zu indexieren. Beispielsweise ist es möglich, HTML-Texte aus Datenbank-Inhalten zu generieren, die dann von der Suchmaschine verarbeitet werden können. Dieser Ansatz hat aber entscheidende Nachteile:

- Die strukturierten Werte der Datenbank kommen im HTML-Text nur als unstrukturierte Zeichenketten vor. Somit läßt sich auch nicht feststellen, in welchem Attribut der Suchbegriff vorkommt.
- Als Ergebnis der Suchanfrage erhält man einen Link auf den HTML-Text, der den Datenbank-Inhalt repräsentiert und keinen Link auf ein Anfrageformular, mit dem dann eine Suche in der Datenbank möglich ist.

In diesem Abschnitt soll nun ein Mechanismus vorgestellt werden, der speziell für SWING entwickelt wurde und die eben genannten Nachteile beseitigt.

3.1 Aufbau der Suchmaschine

Der Basisbestandteil der SWING-Suchmaschine ist das Informationsmanagementsystem *Harvest*. Die zwei Hauptkomponenten des Harvest-Systems [HSW96] werden als *Gatherer* und *Broker* bezeichnet. Gatherer sammeln Meta-Informationen über Dokumente, die im Internet verfügbar sind. Anschließend werden diese Meta-Informationen vom Broker indexiert und als Ergebnis auf Suchanfragen geliefert.

Um die Aktualität der Daten zu gewährleisten, prüfen die Gatherer in periodischen Abständen, ob die verfügbaren Dokumente verändert wurden. Ein Gatherer ermittelt aus den Dokumenten Informationen, die den Inhalt betreffen (z.B. Titel des Dokuments), sowie Meta-Informationen, zu denen beispielsweise URL, Größe und Datum der letzten Änderung zählen. Diese Informationen werden in regelmäßigen Abständen vom Broker angefordert, der diese in sein Retrievalsystem einspeist.

Der Gatherer kann neben Dokumenten auf HTML-Basis auch andere Formate wie PostScript analysieren, da das Harvest-System verschiedene Filter bereitstellt, die in Abhängigkeit vom Dokumenttyp angewendet werden. In Abbildung 1 wird dieser Prozeß der Datenextraktion grob skizziert.

Die offenen Schnittstellen des Gatherers erlauben die Einbindung weiterer Filterprogramme für beliebige Dokumenttypen. Dadurch ist es möglich, eigene Filter für die Integration von Datenbank-Inhalten zu realisieren.

3.2 Integration von Datenbank-Inhalten

Für die Meta-Daten, die vom Gatherer extrahiert werden, definiert Harvest das *Summary Object Interchange Format* (SOIF). Das SOIF besteht aus einer Liste von Attribut/Wert-Paaren. Gebräuchliche Attribute sind beispielsweise *author*, *description*, *keywords* und *title*. Andere Attribute können frei hinzugefügt werden. Die uneingeschränkte Erweiterbarkeit der Attribute wird für die strukturierte Abspeicherung der Datenbank-Inhalte genutzt.

Die Extraktionsroutinen für die Datenbanken, die in die Suche einbezogen werden sollen, müssen in Abhängigkeit vom zugrundeliegenden Datenbanksystem leicht modifizierbar sein, da beispielsweise unterschiedliche SQL-Dialekte

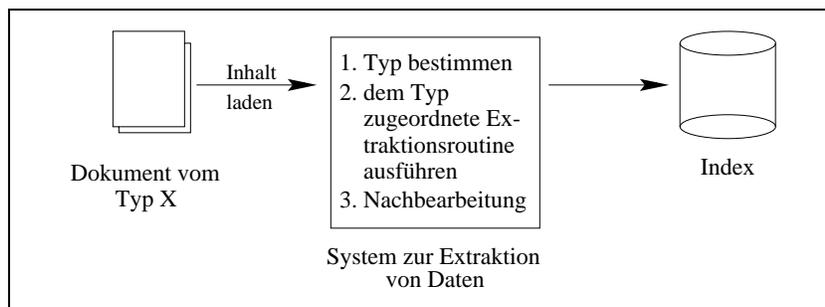


Abbildung 1: Der Prozeß der Datenextraktion

zu berücksichtigen sind. Damit ein Gatherer erkennen kann, um welches Datenbanksystem es sich handelt, wurden neue Dokumenttypen eingeführt, z.B. *ing-db* für Ingres- und *ora-db* für Oracle-Datenbanken. Diese Dokumente müssen auf den WWW-Servern, die den Zugriff auf die einzubindenden Datenbanken ermöglichen, vorhanden sein und enthalten Informationen, die für die Integration einer Datenbank benötigt werden:

- Die Angabe *Summarizer* gibt Auskunft darüber, wo das Programm zu finden ist, das aus den Datenbank-Inhalten ein SOIF-Fragment erzeugt. Der Summarizer wurde in DB-Perl implementiert und muß auf dem WWW-Server, der die Schnittstelle für den Zugriff auf die Datenbank bereitstellt, installiert werden. Bei DB-Perl handelt es sich um ein Perl-Interface, das den einheitlichen Zugriff auf eine Vielzahl existierender Datenbanksysteme ermöglicht.
- *URL* enthält die Adresse für ein Anfrage-Interface, das den Zugriff auf die Datenbank ermöglicht.
- *Title*, *Description* und *Keywords* enthalten Informationen, die dem Datenbank-Anfrageformular zugeordnet werden.
- *Database* gibt den Namen der Datenbank an, die eingebunden werden soll.
- Über die Angaben *Table* und *Attribut* kann der Datenbank-Administrator steuern, welche Attribute aus welcher Relation indiziert werden sollen. Dadurch wird sichergestellt, das nur Inhalte, die öffentlich gemacht werden sollen, an die Suchmaschine weitergegeben werden.

In die SWING-Suchmaschine wurde beispielsweise der CityWorld-Hotelführer integriert, dem eine Oracle-Datenbank zugrunde liegt. Auf dem WWW-Server, über den der Zugriff auf diese Datenbank erfolgt, ist das Dokument *cityworld.ora-db* vorhanden, das folgende Informationen bereitstellt:

```
Summarizer:    http://www.cityworld.de/.../Ora-DB.pl
URL:           http://www.cityworld.de/.../auswahl.pl
Title:         CityWorld - Hotels und Pensionen
Description:   Hotelführer M-V
Keywords:      Hotel, Zimmer, Unterkunft, Pension, Motel, 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.

Die folgenden Schritte werden vom Gatherer durchgeführt, um Datenbank-Inhalte in das Retrievalsystem der SWING-Suchmaschine einzuspeisen:

1. Lesen des Dokuments, das die benötigten Informationen für die Integration der Datenbank enthält (der URL muß SWING bekannt sein).
2. Bestimmen des Datenbanksystems über den Dokumenttyp und Aufruf der entsprechenden Extraktionsroutine.
3. Die Extraktionsroutine ruft dann den Summarizer auf, der unter dem im Dokument angegebenen URL zu finden ist.
4. Der Summarizer generiert ein SOIF-Fragment aus den Schemainformationen der Datenbank und den Attributwerten, die extrahiert werden sollen, und liefert es an den Gatherer zurück.
5. Der Gatherer vervollständigt dieses Fragment und nimmt ein URL-Mapping vor, d.h. der URL für das Dokument mit den Integrationsinformationen wird durch den URL für das Anfrageformular ersetzt.

Nach dem Einsammeln der Daten liegen die Datenbank-Inhalte in strukturierter Form im Suchmaschinen-Index vor, so daß jederzeit feststellbar ist, in welchem Attribut ein Suchbegriff enthalten ist.

3.3 Bereitstellung der relevanten Datenbanken

Als Suchergebnis auf Anfragen, für deren Beantwortung eine Datenbank als relevant eingestuft wurde, wird eine Referenz auf eine Interface-Seite in Form eines Anfrageformulars generiert, die der Nutzer im Dialog vervollständigen kann. In der ursprünglichen Anfrage enthaltene Stichworte werden nach Möglichkeit in das Formular übernommen. Der Nutzer kann somit hochrelevante Informationen in kompakter Form erhalten. Die eigentliche Datenbankanfrage bleibt dem Suchmaschinennutzer hierbei vollkommen transparent.

In nächster Zeit ist die folgende Weiterentwicklung geplant: Falls zu einer Suchthematik mehrere gleichartige bzw. ähnliche Datenbanken existieren, die sich in ihrem Schema nur geringfügig unterscheiden, dann sollen diese Datenbanken gemeinsam zur Bereitstellung der gesuchten Informationen genutzt (geplant ist die Kopplung mehrerer Hoteldatenbanken) werden. Die beteiligten Datenbanken können hierbei auf unterschiedlichen Rechnersystemen beheimatet sein.

4 Profildienste

In der heutigen Zeit reichen Suchmaschinen allein nicht mehr aus, um die Informationsflut im Internet zu bewältigen. Aus diesem Grund gehen immer mehr Anbieter dazu über, Informations-Abonnementdienste bereitzustellen. Je nach der Form des Dienstes unterscheidet man dabei zwischen Alerting- und Profildiensten [Glo98]:

- *Alertingdienst*: In einem föderativen, service-orientierten System zur Verbreitung und Nutzung von Informationen treten Ereignisse auf, von denen nicht bekannt ist, wann sie auftreten, die jedoch für den Nutzer von Interesse sind. Ein Alertingdienst ist ein Benachrichtigungsdienst, der den interessierten Nutzer über das Eintreten von Ereignissen informiert. Die Ereignisse werden dabei immer anbieterseitig ausgelöst.
- *Profildienst*: Ein Profildienst filtert auftretende Ereignisse hinsichtlich nutzerspezifischer Interessen und informiert den Benutzer über relevante Ereignisse. Die Definition des erforderlichen Nutzerprofile kann auf verschiedene Arten erfolgen, zum Beispiel durch Beobachtung des Benutzerverhaltens, anhand von Benutzereigenschaften oder durch Beispielvorgaben. Ein Profildienst kann reaktiv sein, das heißt er kann auftretende Ereignisse hinsichtlich der definierten Profile filtern. Er kann aber auch aktiv nach auftretenden Ereignissen bei Dienstanbietern fragen.

Im Umfeld von Suchmaschinen existieren zur Zeit aber noch keine oder nur sehr rudimentäre Abonnementdienste. In diesem Abschnitt soll deshalb der Profildienst der SWING-Suchmaschine vorgestellt werden, der mit Hilfe von Agententechnologien gezielt nach Informationen sucht, diese entsprechend den Wünschen des Nutzers aufbereitet und in periodischer Form zur Verfügung stellt.

4.1 Aufbau von Abonnementdiensten

Der wesentliche Aufbau eines Abonnementdienstes, unabhängig von seiner Organisationsform (entweder beim Anwender oder beim Dienstanbieter organisiert) entspricht der Abbildung 2 [Her96].

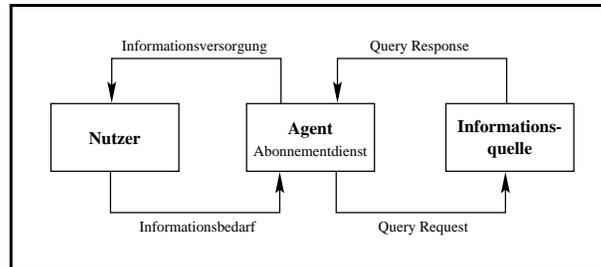


Abbildung 2: Aufbau eines Abonnementdienstes

Der Agent bildet die zentrale Rolle in dieser Drei-Schichten-Architektur. Die Nutzer haben einen bestimmten Informationsbedarf, zu deren Deckung sie sich eines Agenten bedienen. Der Agent verfügt nun nicht unbedingt selbst über die nachgefragte Information und bedient sich wiederum Informationsanbietern. Das Anfrageergebnis der Informationsanbieter wird vom Agenten unter Umständen noch einer Filterung unterzogen, bevor es dem Nutzer bereitgestellt wird.

4.2 Der SWING-Abonnementdienst

In der Konzeptionsphase wurden verschiedene Abonnementdienste wie Ariadne, The Informant, Pointcast, Business Network und Elsevier ContentDirect betrachtet ([Por98]), um Erkenntnisse über positive und negative Eigenschaften dieser Systeme in den Entwurf des SWING-Abonnementdienstes einfließen zu lassen. Das Anforderungsprofil, das der Abonnementdienst einer Suchmaschine abdecken sollte, umfaßt die folgenden Punkte:

- Benachrichtigung des Nutzers über Veränderungen im System,
- Bearbeitung zyklisch wiederkehrender Anfragen,
- Bereitstellung turnusmäßiger Informationen zu einem speziellen Thema.

Der SWING-Abonnementdienst ist so konzipiert, daß er einerseits diese Anforderungen erfüllt und andererseits möglichst viele der Eigenschaften besitzt, die bei der Untersuchung der verschiedenen Systeme als positiv bewertet wurden. Im folgenden soll die Funktionalität beschrieben werden, die der Profildienst der SWING-Suchmaschine bietet. Auf Elementarfunktionen wie das Authentifizieren oder die Neuanmeldung eines Nutzers soll hier nicht näher eingegangen werden.

Die *Anfragebearbeitung* ermöglicht es dem Nutzer, Anfragen an die Suchmaschine in einem von ihm festgelegten Zeitintervall zu abonnieren. Der Unterschied zu herkömmlichen Suchmaschinen liegt in der autonomen Initiierung der Anfrageausführung durch den Abonnement-Agenten. Neben der Abonnementfunktion werden Funktionen zum Löschen, Editieren und manuellem Wiederholen von Anfragen angeboten.

Innerhalb des *URL-Reminders* können durch den Anwender Internetadressen spezifiziert werden, die durch den Abonnementdienst auf Änderungen beobachtet werden. Dabei kann das gesamte Dokument betrachtet werden oder aber nur ausgewählte Bereiche (Tabellen, Listen usw.).

Der *Newschannel* sammelt Nachrichten an einen Nutzer. Bei den Nachrichten kann es sich sowohl um abonnierte Anfrageergebnisse bzw. Mitteilungen des URL-Reminders als auch um SWING-Systemnachrichten handeln. News können im Moment nur in eine Richtung vom SWING-System an Abonnementnutzer übertragen werden, denkbar ist aber auch der Austausch von Nachrichten zwischen Nutzern.

Der *Themenkatalog* ist eine Art Browsing-System, in dem Dokumentenreferenzen nach Themen sortiert sind. Der Nutzer kann durch Auswahl von Themen Neuigkeiten direkt auf seinem Profil sehen. Aufgrund der gleichen Begriffswelt des Themenkatalogs sind Interessensanalysen zwischen Profilen unterschiedlicher Nutzer möglich.

Mit Ausnahme des Themenkatalogs sind alle anderen Komponenten bereits im Profildienst der SWING-Suchmaschine realisiert. Die Architektur dieses Dienstes wird in Abbildung 3 dargestellt, wobei die oben beschriebene Schichtenarchitektur deutlich zu erkennen ist.

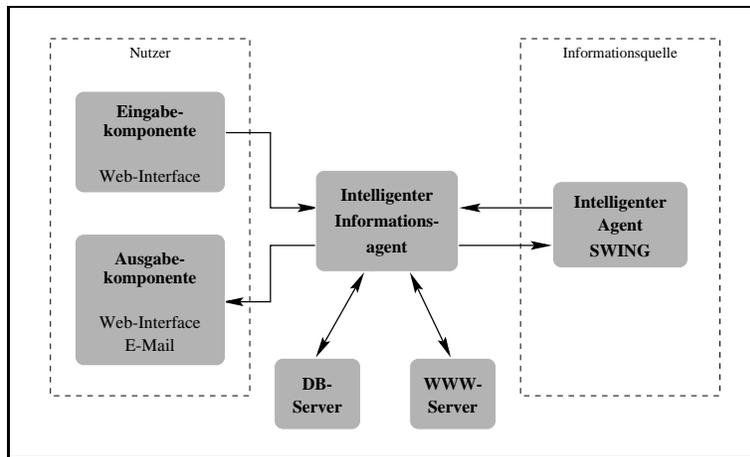


Abbildung 3: Architektur des SWING-Abonnementdienstes

Mit Hilfe der Eingabekomponente stellt der Nutzer des Abonnementdienstes Kontakt zum Informationsagenten her. Wichtigstes Werkzeug ist für ihn in diesem Zusammenhang der WWW-Browser. Die Ausgaben der personalisierten Informationen des Agenten können dem Nutzer sowohl in Form von E-Mails als auch über das WWW bereitgestellt werden. Der Informationsagent verwaltet alle Daten der Nutzer, wie Name, Passwort usw.. Im besonderen verwaltet er auch die vom Nutzer gestellten Anfragen, die zurückgelieferten Dokumentenreferenzen und die zu überwachenden URLs. Zusätzlich übernimmt er die Visualisierung der Daten. Dies geschieht in diesem Profildienst mittels HTML-Formularen. Zur Informationsgewinnung bedient sich der Informationsagent des SWING-Agenten, indem er über eine Schnittstelle Informationen zu einer Nutzeranfrage anfordert. Die Antwort, in Form eines HTML-Formulars, beinhaltet neben den Dokumentenreferenzen auch Relevanzbewertungen und Kurzbeschreibungen der Referenzen.

5 Autorenwerkzeug zur Verwaltung von Inhaltsverzeichnissen

Der Online-Dienst benutzt zur Verwaltung, Pflege und Konsistenzwahrung der Link-Sammlung ein Autorensystem. Dieses wird auf Basis des intelligenten Hypermedia-Systems HyperWave implementiert. Wir beschreiben im weiteren dieses Autorensystem und warum HyperWave zur Realisierung genutzt wird.

MV-Info stellt eine in verschiedene Kategorien unterteilte Link-Sammlung mit Information aus und über Mecklenburg-Vorpommern da. Damit ist es vergleichbar mit anderen sogenannten Katalog- oder Verzeichnisdiensten wie Yahoo.de¹, LEO.org² oder Entry.de³. Für die einzelnen Kategorien, wie Tourismus, Städte und Regionen, sind Redakteure zuständig. Sie haben folgende Aufgaben:

- Neuaufnahme von Verweisen auf neue Server, Web-Präsentationen oder einzelne Web-Seiten mit Bezug zu M-V, Initiator kann sein
 - ein Server-Betreiber, der seinen Server im System anmeldet
 - ein Web-Nutzer, der Hinweise auf M-V relevante Informationen an den Redakteur direkt oder den Dienstbetreiber allgemein schickt
 - der Redakteur selbst, der mittels weiterer Suchdienste neue Inhalte “aufspürt”
- Zuordnung zu einer oder mehreren Kategorien
- Entscheidung, ob der Inhalt aufgenommen wird oder nicht, Kriterien sind dabei Qualität der Gestaltung, Eignung des Inhaltes⁴, Themenbezug etc.
- Überprüfung der Inhalte auf regelmäßiger Basis, evtl. muß eine thematische Neuordnung, zusätzliche Aufnahme in weitere Kategorien oder ein Entfernen der Verweise erfolgen
- Kontrolle der Verfügbarkeit der jeweiligen Inhalte (Link-Überprüfung), aktuell nicht zugreifbare Verweise (Server ist für einen längeren Zeitraum nicht erreichbar) werden deaktiviert und gegebenenfalls wieder aktiviert

Ziel des Redaktionssystems ist es, den Redakteur bei den genannten Aufgaben zu unterstützen, Teilaufgaben zu automatisieren, bestimmte Abläufe (zyklische Wiedervorlage, Link-Überprüfung) zu erzwingen und damit insgesamt die Konsistenz der Link-Sammlung zu wahren. Technisch sind damit Forderungen nach der Unterstützung einer verteilten Arbeit in der Redaktionsgruppe, damit verbundener ablaufenden parallelen Arbeit (Transaktionen), einer Versionskontrolle und einer möglichst automatischen Erzeugung der eigentlichen Web-Präsentation verbunden.

Diese genannten Forderungen werden heutzutage von verteilten Datenbanklösungen realisiert. Da es aber keine kommerziell nutzbaren, zugeschnittenen Lösungen gibt, bestand die Alternative in einer Neuimplementierung auf Basis z. B. eines relationalen oder objekt-relationalen Datenbanksystems oder im Einsatz geeigneter Web-Technologie.

HyperWave [Mau96] ist solch eine Web-Server-Lösung, die Hypermediatechniken mit Datenbankunterstützung verknüpft. Das System verwaltet strukturiert Web-Dokumente in einem Dokumenten-Server und Verweise separat in einer Link-Datenbank. Darüberhinaus bietet das System verbesserte Strukturierungsmöglichkeiten (Kollektionen, Cluster, linearer Hypertext, Guided tours, etc.) zur Ablage von Hypermediadokumenten. Über einen objekt-orientierten Mechanismus lassen sich leicht Erweiterungen in das System einbinden. Zur Datenbankverwaltung kann alternativ zum HyperWave-eigenen System das relationale DBMS ORACLE benutzt werden, so daß speziell bei großen Anwendungen die Performance garantiert werden kann.

Durch vorgefertigte Extern-Link-Objekte, einen Mechanismus zur dynamischen Generierung von Web-Inhalten (PLACE-Templates), ein durchdachtes Rechtevergabe- und Nutzerkonzept sowie die existierende Versionskontrolle für Dokumente ist HyperWave einer selbst zu realisierenden, individuellen Datenbankapplikation vorzuziehen.

¹<http://www.Yahoo.de>

²<http://www.LEO.org>

³<http://www.Entry.de>

⁴Z. B. fremdenfeindliche oder pornografische Inhalte werden nicht in das System aufgenommen. Notfalls wird eine Entscheidung in der Redaktionsgruppe bei “Grenzfällen” herbeigeführt.

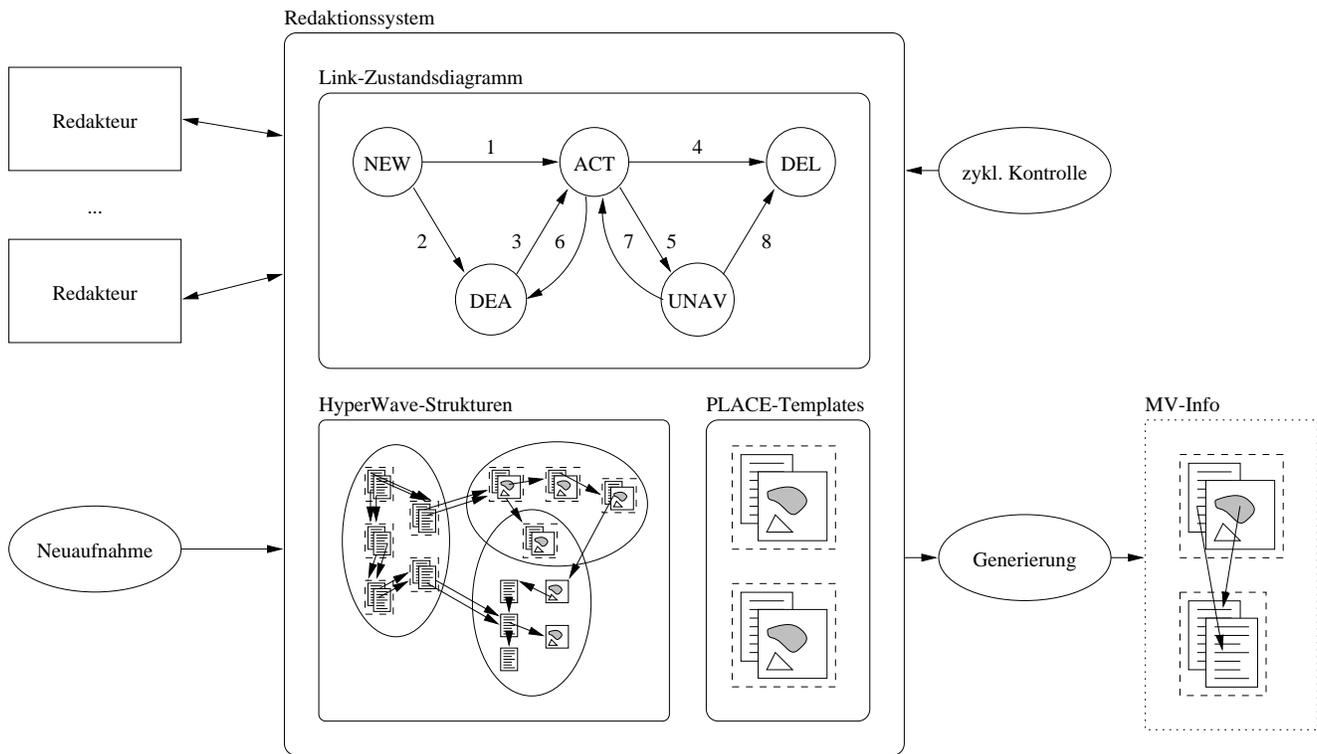


Abbildung 4: Workflow im Autorensystem

Das Redaktionssystem (siehe Abbildung 4) ist mit den von HyperWave bereitgestellten Mechanismen relativ einfach zu implementieren. Mit speziellen Link-Objekte werden alle benötigten Informationen für die in das Informationssystem MV-Info aufgenommenen Verweise verwaltet. Die Redakteure können entsprechend den ihnen zugewiesenen Rechten diese Objekte manipulieren, z. B. sie in bestimmte Kategorien (realisiert mit HyperWave-Collections) verschieben. Bestimmte Vorgänge (Link-Überprüfung) laufen durch Skripte automatisiert ab.

Der HyperWave-Template-Mechanismus gestattet es, die Gestaltung der MV-Info-Web-Präsentation unabhängig von den eigentlichen Inhalten zu verwalten. Die Präsentation wird automatisch aus dem System heraus erzeugt, und sie kann auf den Web-Servern des MV-Info-Systems auch von anderer Server-Technologie (Apache, Netscape, Roxen, Microsoft-Web-Server) verwaltet werden.

Durch die Versionsverwaltung lassen sich jederzeit verschiedenste Inhalte neu erzeugen oder wiederherstellen. Sei es, weil Fehler auftraten, Redakteure in Widerspruch stehende Aktionen auslösten, oder um neue, noch nicht freigegebene Web-Präsentationen auszuprobieren.

6 Zusammenfassung

SWING ist eine moderne Suchmaschine, die viele exklusive Merkmale wie die Integration von Datenbanken bei der Suche im WWW aufweist. SWING ist damit eine der entscheidenden Funktionalitäten im Landesinformationssystem MV-Info, das im Internet die WWW-Angebote Mecklenburg-Vorpommerns integriert und klassifiziert. Das SWING-Projekt soll hier auch als gelungenes Beispiel für die Synergie-Effekte stehen, die ein Verbundprojekt mit lokalen Firmen (insbesondere Software-Häusern und EDV-Dienstleistern) und Universitäten (hier speziell der Fachbereich Informatik der Universität Rostock) erzeugen kann — und dies in einem für die Entwicklung des

Landes entscheidenden Bereich beim Übergang in die Informationsgesellschaft: bei den *multimedialen Informationssystemen im Internet*.

Literatur

- [DLHM97] A. Düsterhöft, U. Langer, H. Meyer und A. Heuer. SWING: Konzept einer Suchmaschine für das regionale Informationssystem MV-Info. In *9. Workshop "Grundlagen von Datenbanken", Forschungsbericht 643 der Universität Dortmund, 1997*.
- [Glo98] Global-Info: *Vorschlag der technischen Arbeitsgruppe in Schwerpunkt 4 zur Rahmenarchitektur*. 1998.
- [Her96] Hermans, B.. *Intelligent Software Agents on the Internet*. Diplomarbeit an der Universität Tilburg, Niederlande, Juli 1996.
- [HMDL97] A. Heuer, H. Meyer, A. Düsterhöft und U. Langer, . SWING: Der Anfrage- und Suchdienst des Regionalen Informationssystems MV-Info. In: *1. IuK-Tage Mecklenburg-Vorpommern, Schwerin, 27.-28. Juni 1997*, Wirtschaftsministerium Mecklenburg-Vorpommern, Juni 1997.
- [HSW96] D.R. Hardy, M.F. Schwartz, and D. Wessels. HARVEST - Effective Use of Internet Information. TR CU-CS-743-94, University of Colorado at Boulder, Dept of CS, Januar 1996.
- [LDHM97] U. Langer, A. Düsterhöft, A. Heuer und H. Meyer. SWING: Ein Anfrage- und Suchdienst im Internet. In *Rostocker Informatik-Berichte (1997), Heft 21*.
- [Mau96] H. Maurer. *HyperWave — The Next Generation Web*. Addison-Wesley Longman, 1996.
- [Por98] B. Porst. Konzept eines Internet-Abo-Dienstes für die SWING Suchmaschine. Studienarbeit, Universität Rostock, Fachbereich Informatik, Oktober 1998.
- [Tit98] P. Titzler. Realisierungsvorschlag für die Implementierung einer verteilten Suchmaschine. Studienarbeit, Universität Rostock, Fachbereich Informatik, Oktober 1998.