

Konzeption und Realisierung einer kontextbasierten Suche in externen Datenquellen im Bereich der Neurobiologie



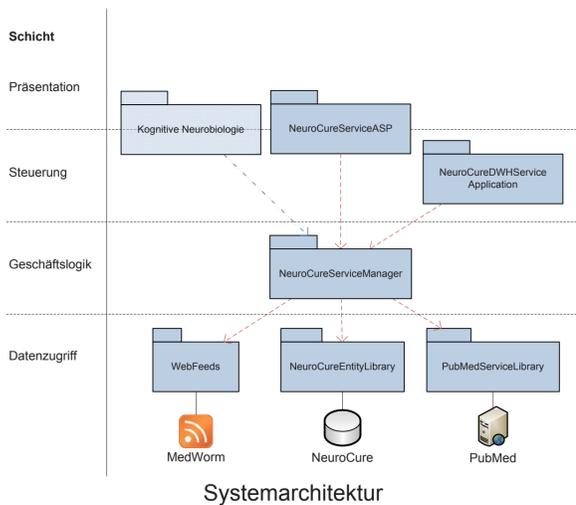
Masterarbeit, vorgelegt von Rainer Muth

Aufgabenstellung:

Die Arbeit umfasst die Konzeption und prototypische Implementierung einer kontextbasierten Suche zur Erweiterung des Systems NeuroCure. Das System NeuroCure wird verwendet um Daten für neurologische Erkrankungen und Testverfahren in Verbindung zu bringen. Die Suche wird in den externen Informationsquellen PubMed und MedWorm durchgeführt, wobei die Suche für relevante Ergebnisse optimiert werden soll.

Systemarchitektur:

Die Funktionen des Prototypen sind in verschiedene Bibliotheken aufgeteilt. Das vorhandene System der *Kognitiven Neurobiologie* kann durch Verwendung der *NeuroCureServiceManager* Bibliothek auf die Funktionen der kontextbasierten Suche zugreifen.



Konzeption der Suchfunktionen:

Die Suchfunktionen werden für PubMed und MedWorm unterschiedlich implementiert. PubMed bietet einen Webservice, der durch eine WSDL definiert und abgefragt werden kann. Bei MedWorm wird eine frei verfügbare Bibliothek verwendet, um Daten aus dem RSS-Feed abzufragen und dem System bereit zu stellen. Weiterhin werden zwei Suchverfahren vorgestellt, durch die mögliche Seiteneffekte vermieden werden sollen. Zum einen gibt es den direkten Aufruf, der Anfragen direkt an die externen Systeme weitergibt. Hierbei werden der Ausfall sowie die Überlastung der Systeme nicht beachtet.

Eine weitere Suchfunktion speichert die gefundenen Ergebnisse in der Datenbank ab. Die Suche liefert dann immer die aktuellsten Ergebnisse zurück, die in der Datenbank enthalten sind. Hierbei kann auf Ausfälle der externen Informationsquellen gut reagiert werden. Der Datenbestand muss allerdings in bestimmten Intervallen aktualisiert werden.

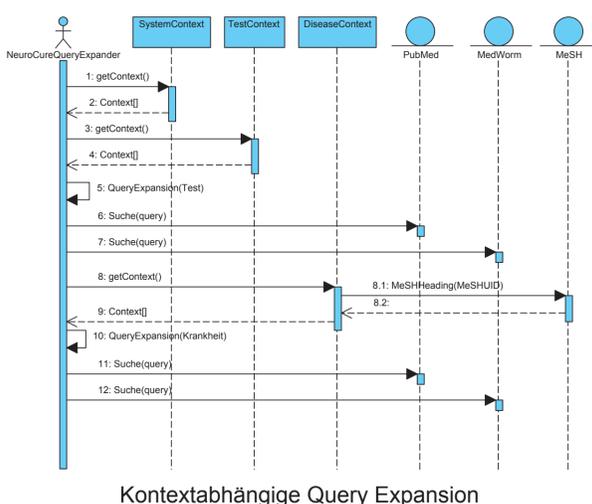
Implementierung des Prototypen:

Die Nutzung des Webservice von PubMed erfolgt über die in Visual Studio integrierte Webservice-Unterstützung. Hierbei werden Klassen basierend auf den Definitionen in der WSDL-Datei erstellt. Für den Zugriff auf die Daten des RSS-Feed von MedWorm wird die frei verfügbare Bibliothek WebFeeds verwendet.

Anpassung der Suchanfrage:

Zur Erhöhung der Anzahl von relevanten Ergebnissen wird das Verfahren der Query Expansion verwendet. Hierbei werden Suchanfragen mit zusätzlichen Informationen angereichert.

Für das System Neurocure werden kontextabhängige Begriffe verwendet. Diese sind zum einen systemabhängig (zum Beispiel "mouse" oder "mice") und zum anderen abhängig vom Suchbegriff. So werden bei Krankheiten, soweit vorhanden, Informationen aus dem online verfügbaren Thesaurus MeSH verwendet und bei Testverfahren entsprechende im System definierte Synonyme.



Performanz:

Die Performanz des Systems wurde zusätzlich untersucht, um die Entscheidung für eines der beiden Suchverfahren zu unterstützen. Zwei Messungen wurden durchgeführt, wobei die durchschnittlichen Antwortzeiten von PubMed und MedWorm ermittelt wurden.

	PubMed		MedWorm	
	M1	M2	M1	M2
Anzahl der Anfragen	124	124	124	124
Ausführungszeit gesamt	6 Min.	6 Min.	8 Min.	4 Min.
Durchschnittliche Abfragezeiten	> 3 Sek.	< 3 Sek.	> 4 Sek.	> 2 Sek.
Minimale Abfragezeit	0,6 Sek.	0,7 Sek.	0,5 Sek.	0,5 Sek.
Maximale Abfragezeit	6,9 Sek.	6,5 Sek.	16,5 Sek.	21 Sek.

Vergleich der Abfragezeiten für zwei Messungen

Die Messungen ergaben bei der direkten Anfragen an PubMed und MedWorm eine durchschnittliche Antwortzeit von 3 Sekunden. Im schlechtesten Fall muss ein Benutzer allerdings über 20 Sekunden auf eine Antwort warten.

Die Ergebnisse der Messungen belegen, dass die Daten der kontextbasierten Suche automatisch erfasst und in bestimmten Intervallen aktualisiert werden sollte.

Ergebnisse:

Das bestehende System NeuroCure wurde so erweitert, dass die Daten von PubMed und MedWorm integriert werden können. Dabei werden relevante Informationen durch die angepassten Suchanfragen ermittelt und gespeichert.

Die Untersuchung der Performanz hat ergeben, dass eine Integration der Daten in NeuroCure die bessere Variante darstellt. Hierbei kann die Aktualisierung der Daten über eine entsprechende Automatisierung durchgeführt werden. Weiterhin hat der Einsatz der kontextbasierten Suche ergeben, dass die Konfiguration der Query Expansion einen großen Einfluss auf die Relevanz der Ergebnisse hat.

Die kontextbasierte Suche lässt sich zusätzlich in ähnliche Systeme mit einem entsprechend definierten Kontext einbinden.

Quellen:

- [1] <http://upload.wikimedia.org/wikipedia/commons/f/fb/US-NLM-PubMed-Logo.svg>
- [2] <http://www.anh-europe.org/files/images/MedWorm-Logo.png>