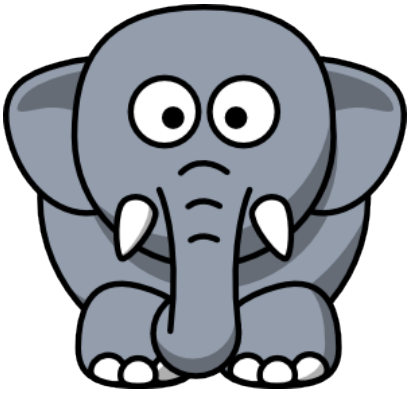

Elephant Search Anforderungen

Anforderungsdokument der Bachelorthesis

Studiengang:	Informatik
Autoren:	Sven Osterwalder,Mira Günzburger
Betreuer:	Dr. Jürgen Eckerle
Datum:	08.06.2014

Inhaltsverzeichnis

Versionen

Version	Datum	Status	Bemerkungen
0.1	08.06.2014	Entwurf	Anforderungsdokument erstellen
0.2	08.08.2014	Entwurf	Korrekturen und Ergänzungen

1 Einleitung

Das nachfolgende Dokument beschreibt die Anforderungen der Bachelorthesis von Sven Osterwalder und Mira Günzburger. Als Vorarbeit der Bachelorthesis dient die Arbeit, welche im Rahmen des Moduls 7302 „Projekt 2“ bereits erstellt wurde.

In der Bachelorthesis soll ein Werkzeug für die semantische Suche in einer Wissensdatenbank implementiert werden.

Wie in der Abschlussdokumentation der Projekt 2 Arbeit beschrieben, handelt es sich bei semantischen Suchmaschinen um „Werkzeuge, die in der Lage sind, auf Fragen mit Hilfe einer Datenbank oder des Internets Antworten zu generieren. Solche Werkzeuge können insbesondere dann eine sehr wertvolle Unterstützung für den menschlichen Experten sein, wenn unter extremer Zeitnot komplexe Entscheidungen getroffen werden müssen, wie beispielsweise in der medizinischen Diagnostik. Die Firma IBM hat vor nicht allzu langer Zeit für eine Überraschung gesorgt, als sie die Leistungsfähigkeit von „Watson“ im Quiz Jeopardy demonstriert hat. In diesem Quiz, wo schwierige, oft zweideutig formulierte Fragen aus beliebigen Bereichen unter Zeitdruck beantwortet werden müssen, konnte sich Watson überlegen gegenüber zwei bisher sehr erfolgreichen menschlichen Champions durchsetzen. [1]

Wie im Fazit der Projektarbeit beschrieben, muss der Fokus der Thesis verschoben werden. So soll der Schwerpunkt der Arbeit rein auf der technischen Umsetzung einer semantischen Suche mit Hilfe von Apache Stanbol gesetzt werden. Dies entgegen der ursprünglichen Intention, der Entwicklung eines kindergerechten Frontends.

Nachfolgend werden die einzelnen Aufgaben der Bachelorthesis beschrieben.

2 Wissensdomäne

Wie sich in der Vorarbeit herausgestellt hat, ist es notwendig die Domäne, in welcher Anfragen gestellt werden sollen, sehr detailliert abzubilden. Zudem ist die technische Umsetzung der Suche mittels Apache Stanbol weniger weit ausgearbeitet als ursprünglich angenommen.

Um die Komplexität in einem angemessenen Rahmen zu halten, gilt es die Entitäten, also die Modellierung der Umwelt, stark einzuschränken.

Als Folge dieser Erkenntnisse wird die Wissendomäne, mit welcher gearbeitet wird, eingeschränkt. Bei der gewählten Domäne handelt es sich um die Grundlagen der Programmierung am Beispiel der Programmiersprache Java.

Dies erlaubt es, dass der Aufbau der Wissensdatenbank überschaubar bleibt, hat aber den Nachteil, dass nur eine beschränkte Anzahl von Fragestellungen beantwortet werden können.

3 Komponenten

Bei den Recherchen der Projektarbeit hat sich die Erkenntnis ergeben, das eine erfolgreiche Verarbeitung von Anfragen die folgenden Komponenten benötigt:

- Abbildung der Umwelt mittels Wissensdatenbank
- Definieren von Regeln zur Ableitung von Schlüssen mittels Logik
- Erkennung von Sprache in Schriftform
- Interne und externe Schnittstellen zur Kommunikation

Die oben genannten Komponenten werden bereits im Ansatz durch die in der Projektarbeit evaluierte Lösung — Apache Stanbol — zur Verfügung gestellt. Allerdings hat sich gezeigt, dass diese grössere Erweiterungen benötigen um die gewünschten Ergebnisse zu liefern.

3.1 Architektur

Um die verschiedenen Teile zusammen zu nutzen, bietet Apache Stanbol eine frei konfigurierbare Verkettung dieser an. Dies geschieht mittels einer sogenannten Enhancement-Chain. Konkret heisst dies, dass eine beliebige Eingabe dieser Kette übergeben werden kann, worauf dann die erste Softwarekomponente der Kette die Eingabe verarbeitet und das Resultat an die nächste Komponente weiterreicht. Diese Vorgang wird durch sämtliche Komponenten der Kette fortgeführt bis schlussendlich das Endresultat an die anfragende Entität zurückgegeben wird.

Die Arbeit der Bachelorthesis besteht also darin, die Enhancement-Chain und die einzelnen Entitäten zu konfigurieren und zu erweitern.

3.2 Abbildung der Umwelt mittels Wissensdatenbank

3.2.1 Objekte abbilden

Die Abbildung der Umwelt geschieht in Apache Stanbol mittels dem sogenannten Entity Hub. Dieser stellt Informationen zu Entitäten und Objekten einer spezifischen Wissensdomäne zur Verfügung. Die Beziehungen werden in Apache Stanbol in Form von Relationen zwischen den Entitäten abgebildet, analog dazu werden die Eigenschaften als Attribute erfasst.

Die konkrete Arbeit mit dem Entity Hub besteht also darin Objekte, der für die Arbeit gewählten Domäne, als Entitäten abzubilden.

3.2.2 Definieren von Regeln zur Ableitung von Schlüssen mittels Logik

Um nun aus der Wissensdatenbank Schlüsse ziehen zu können, werden Regeln benötigt. Regeln werden verwendet um mittels Bedingungen auf weitere Eigenschaften schliessen zu können. Apache Stanbol unterstützt auf Prädikatenlogik basierende Regeln, welche innerhalb der Rule Store Komponente als Rezepte gespeichert werden. Diese sind nichts anderes als eine Zusammenfassung von Regeln, welche eine ähnliche Objektkategorie betreffen.

3.3 Spracherkennung

Um einen Eingabesatz auszuwerten, muss dieser erst als solcher erkannt werden. Konkret müssen also, die einzelnen Wörter als Tokens identifiziert und einer Sprachkategorie zugeorndet werden. Dies geschieht mittels der Spracheerkennungskomponente OpenNLP (open natural language processing). Diese Spracherkennung wird für die Englische Sprache von Stanbol schon weitgehend angeboten.

Nach unseren Erkenntinissen ist dies mit der Deutschen Sprache genau so möglich, jedoch nicht gegeben. Dies muss also noch erarbeitet werden. Nach ersten Recherchen wird dazu der Tokenizer sowie der Parser von OpenNLP verwendet.

3.4 Interne und Externe Schnittstellen zur Kommunikation

Um die oben genannten Komponenten ansprechen und nutzen zu können, sind Schnittstellen zur Kommunikation unumgänglich.

3.4.1 Interne Kommunikation

Intern nutzt Apache Stanbol, wie bereits beschrieben, die sogenannte Enhancement Chain um von einem gegebenen Input, mittels verschiedenen Komponenten, zu einem Output zu gelangen [2]. Die Enhancement Chain basiert auf einer Graph-Struktur, welche sie zur Kommunikation nutzt.

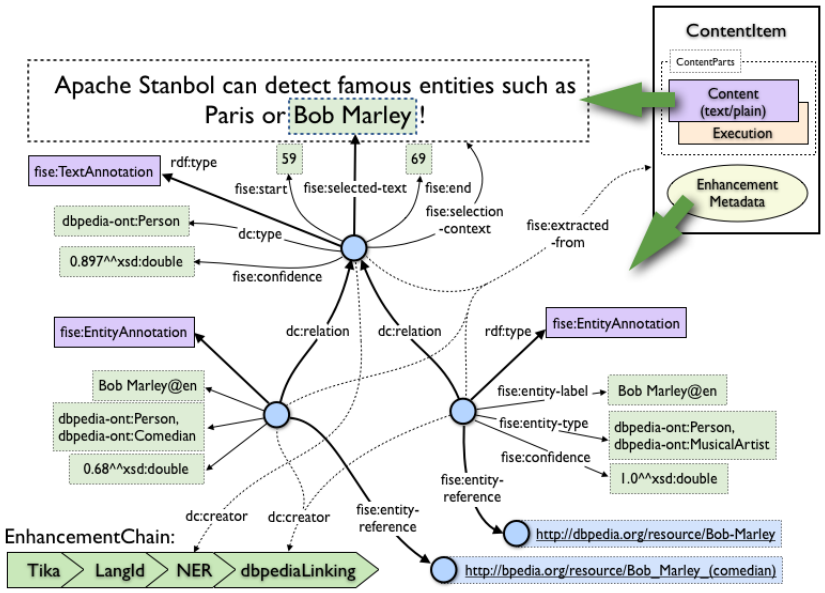

Abbildung 3.1: Apache Stanbol Enhancement Chain[1]

[1] [3]

3.4.2 Externe Kommunikation

Jede Komponente von Apache Stanbol, so z.B. auch die Enhancement Chain sowie deren Einzelkomponenten, verfügt über ein REST-Interface. Dies dient zur Kommunikation gegen aussen. Ein schematischer Ablauf der Kommunikation wird in Abbildung 3.2 grob dargestellt.

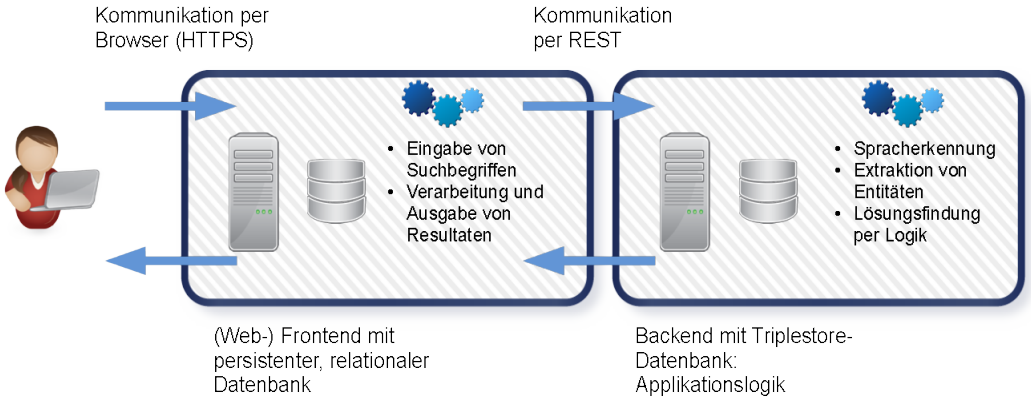

Abbildung 3.2: Kommunikation im Überblick[2]

[2]Eigene Darstellung mittels Libre Office Writer

4 Ziel der Thesis

Als Endresultat der Thesis soll eine Applikation zur Verfügung stehen, welches es erlaubt eine Frage in deutscher Sprache zur Domäne der Programmierung anhand der Programmiersprache Java zu stellen. Dies kann dank der gegebenen REST-Schnittstelle z.B. direkt per Konsole oder aber per ansprechendem Web-Interface geschehen, welches aber nicht Teil der Thesis ist.

Die Applikation soll in der Lage sein mittels der aufgebauten Wissensdatenbank, deren Relationen und schlussendlich Regeln die Frage zu beantworten. Kann eine Frage nicht eindeutig beantwortet werden, sollen zumindest Satzteile (Tokens) extrahiert und der entsprechende Inhalt zu diesen zurückgegeben werden. Eine Antwort ist dabei die Rückgabe einer Entität mit all deren Feldern, welchen dann von dem anfragenden Objekt entsprechend verarbeitet werden kann.

Literaturverzeichnis

[1] Sven Osterwalder, Mira G.: *Requirements of Elephant Search – A semantic search engine for children*. 2014

[2] *Apache Stanbol Enhancement Chain*. `http://stanbol.apache.org/docs/trunk/components/enhancer/chains/`

[3] *Apache Stanbol Enhancement Graph*. `https://stanbol.apache.org/docs/trunk/components/enhancer/enhancementstructure.png`

Abbildungsverzeichnis

Tabellenverzeichnis