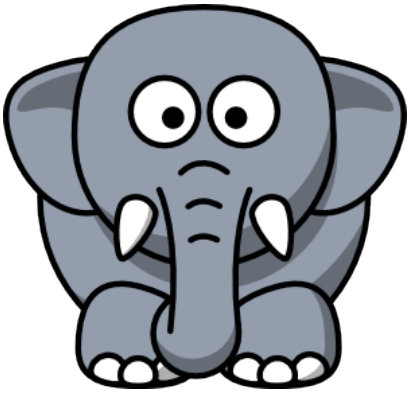

Elephant Search Anforderungen

Anforderungsdokument der Bachelorthesis

Studiengang:	Informatik
Autoren:	Sven Osterwalder, Mira Günzburger
Betreuer:	Dr. Jürgen Eckerle
Experte:	Leclerc Jean-Marie
Datum:	24. September 2014

Inhaltsverzeichnis

Versionen

Version	Datum	Status	Bemerkungen
0.1	08.06.2014	Entwurf	Anforderungsdokument erstellen
0.2	08.08.2014	Entwurf	Korrekturen und Ergänzungen
0.3	19.09.2014	Entwurf	Anpassungen an die Aufgabenstellung
0.4	24.09.2014	Entwurf	Anpassung der Milestones nach Besprechung mit Betreuer

1 Einleitung

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

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 im Rahmen der Bachelorthesis ein Dokument mit Tutorial-Charakter erstellt werden, welches einen leicht verständlichen Einstieg in das Thema der semantischen Datenbanken / Suche bietet. Als Hilfsmittel dafür kommt Apache Stanbol zum Einsatz.

Nachfolgend werden die einzelnen Elemente der Bachelorthesis beschrieben.

1.1 Aufgabenstellung

'Ziel dieser Arbeit ist die Entwicklung und Anwendung eines Systems zur Speicherung in einer semantischen Datenbank auf der Basis von Apache Stanbol. Dies schliesst die Erstellung einer Domänen-Ontologie mittels RDF/OWL ein und die Anwendung dieser Ontologie auf ein Anwendungsproblem, wie beispielsweise der Erlernung einer Programmiersprache ein. Exemplarisch soll aufgezeigt werden, wie dabei ein Knowledge Engineer vorgehen, um eine Problemdomäne systematisch zu modellieren und formalisieren. Besondere Bedeutung kommt dabei der Schnittstelle zwischen Mensch und System zu.' [2]

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 einen leicht verständlichen Einstieg in das Thema der semantischen Datenbanken / Suche bieten zu können, wird die Wissendomäne, mit welcher gearbeitet wird, stark eingeschränkt. Bei der gewählten Domäne handelt es sich um die Programmiersprache Prolog. Anhand der wird an einem exeplarischen Beispiel gezeigt, wie eine Wissendomäne modelliert wird, wie darin enthaltene Daten verknüpft werden und wie schlussendlich die logische Ableitung der Regeln der Domäne erfolgt.

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
- Interne und externe Schnittstellen zur Kommunikation

Die oben genannten Komponenten werden bereits durch die in der Projektarbeit evaluierten Lösung — Apache Stanbol — zur Verfügung gestellt. Allerdings hat sich gezeigt, dass diese keine dem Menschen leicht verständliche Möglichkeit bietet, um eine Fragestellung in ein für den Menschen logisches Resultat zu überführen.

So kann zum Beispiel die Frage 'Welches ist die Hauptstadt der Schweiz¿ nicht einfach so beantwortet werden. Die Lösung bietet Extraktion von Entitäten, deren Eigenschaften und Relationen sowie Abfragen dieser mittels einer speziellen Sprache (SPARQL). Die Überführung der extrahierten Entitäten in eine Abfrage, welche für den Menschen ein sinnvolles Resultat zurückliefert, ist nicht gegeben.

3.1 Architektur

Um die verschiedenen Teile zusammen zu nutzen, bietet Apache Stanbol eine frei konfigurierbare Verkettung dieser an. Dies geschieht mittels einer so genannten 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.

3.2 Abbildung der Umwelt mittels Wissensdatenbank

3.2.1 Objekte abbilden

Die Abbildung der Umwelt geschieht in Apache Stanbol mittels dem so genannten 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.

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 so genannte Enhancement Chain um von einem gegebenen Input, mittels verschiedenen Komponenten, zu einem Output zu gelangen [3]. Die Enhancement Chain basiert auf einer Graph-Struktur, welche sie zur Kommunikation nutzt.

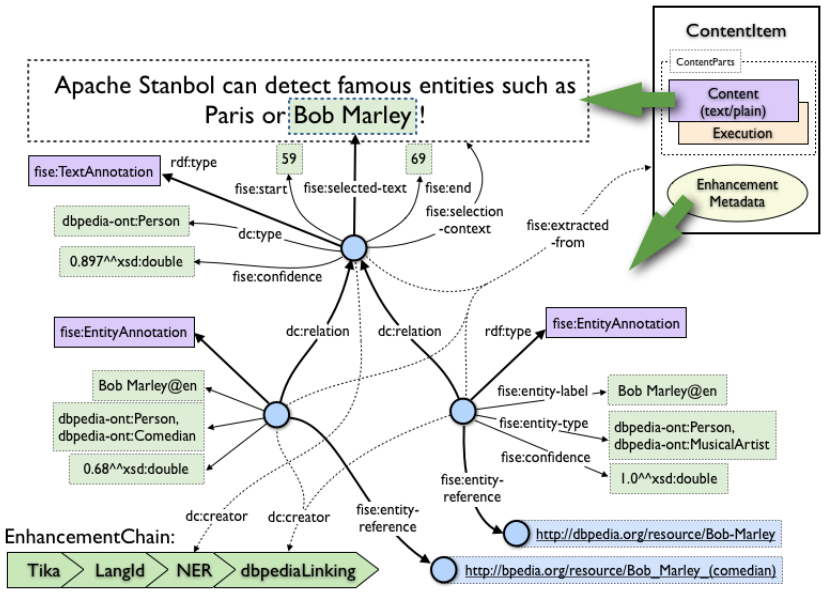

Abbildung 3.1: Apache Stanbol Enhancement Chain[1]

[1] [4]

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.

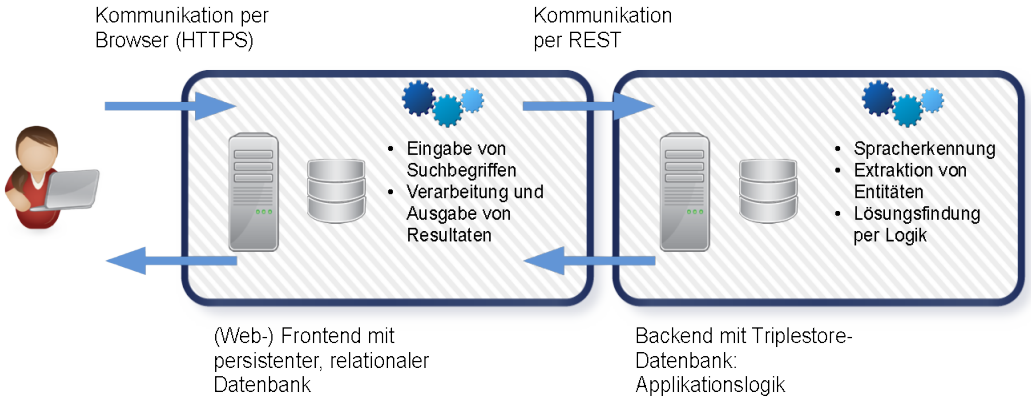

Abbildung 3.2: Kommunikation im Überblick[2]

[2]Eigene Darstellung mittels Libre Office Writer

4 Technische Umsetzung

Zur technischen Umsetzung wird, wie in der Abschlussarbeit des Moduls 7302 ‘Projekt 2‘ evaluiert, Apache Stanbol eingesetzt. Freundlicherweise steht von der BFH Infrastrukur zum Betrieb zur Verfügung. Der Server steht unter elephantsearch.bfh.ch zur Verfügung.

5 Ziel der Thesis

Als Endresultat der Thesis soll ein Dokument mit Tutorial-Charakter stehen, welches aufzeigt wie Knowledge Engineers vorgehen, um eine Problemdomäne systematisch zu modellieren und zu formalisieren.

Es soll also gezeigt werden, wie eine Speicherung von Daten, die Verknüpfung dieser sowie ihre logische Ableitungen in einer semantischen Datenbank für eine RDF/OWL-Ontologie umgesetzt wird. [2]

Dies wird mittels eines konkreten Anwendungsproblems, der Erlernung der Programmiersprache Prolog, umgesetzt.

5.1 Milestones

Die folgende Auflistung zeigt eine Übersicht, der in der Anfangsphase bereits erkennbaren Meilensteine der Arbeit:

- Anforderungsdokument
- Modellierung der Entitäten, Attribute und Regeln;
  Erzeugung der Ontologie mit dem Ziel einer Domänenspezifikation
- Einbetten der Sprache der Domänenspezifikation in die technische Lösung
- Analysieren und Anwenden der Sprache SPARQL
- Installieren der nötigen Infrastruktur auf dem BFH Server
- Realisierung einer einfach handhabbaren Anwenderoberfläche

Literaturverzeichnis

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

[2] Eckerle, Dr. J.: *Aufgabenstellung Bachelorthesis*. 2014

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

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

Abbildungsverzeichnis