Projektvertrag

Verantwortlicher: Daniel Alexander

Daniel Alexander, Jakob Kusnick, Richard Khulusi, Christopher Schott, Tobias Turke, Nikolas Oberhäuser, Hans Dieter Pogrzeba, Felix Conrads

Inhaltsverzeichnis

1. Zielbestimmung

Ziel des Projekts ist die Erstellung einer Anwendung zum Sammeln und zielgerichteten Auswerten von Schachdaten.
Zu diesem Zweck müssen drei Dienste erstellt werden, die das Internet nach Schachdaten im PGN-Format, Informationen zu Schachspielern und Events durchsuchen. Diese gesammelten Daten werden dann in RDF-Triples konvertiert, die sich in einem Triplestore speichern lassen. Mit Mitteln der Data-Fusion werden die gesammelten Daten schrittweise in die Menge der bereits vorhandenen Daten eingefügt.
Weiterhin wird eine Anwendung erstellt, mit der die im Triplestore hinterlegten Datensätze ausgewertet und abgefragt werden können.
Am Ende wird diese Anwendung dann Fragen wie beispielsweise die Folgende in Form einer SPARQL-Query beantworten können: „Welche Spieler aus Deutschland mit einer ELO von über 2500 haben im Jahr 2007 in China gespielt?“

2. Voraussetzungen

Eine Voraussetzung, um das Projekt erfolgreich durchzuführen, ist u.a. der Konverter von PGN zu RDF, welchen die Gruppe SWP13-SC im Vorjahr erstellt hat.
Weiter wird der durch den Benchmark Test als Vorprojekt gefundene Triplestore benötigt.
Um das Projekt möglichst effizient zu bearbeiten, ist ein grundlegendes Verständnis von LIMES und dessen Benutzung sowie Wissen über die Funktionsweise eines Crawlers unabdingbar.

3. Wichtige Rollen

3.1 User

Der größte Anteil von Anwendern wird der **User** sein. Dieser besitzt keinerlei Schreibrechte an der Datenbank. Er darf lediglich lesende Anfragen stellen.
Der User hat die Möglichkeit über eine Eingabezeile Suchanfragen direkt in SPARQL an den Triplestore zu stellen.
Es werden dem User generelle Informationen über den Datenbestand, wie etwa wie viele Schachpartien enthalten sind, präsentiert.

3.2 Administrator

Der **Administrator** ist ein Benutzer mit erhöhten/erweiterten Rechten, die ihm zusätzlich zu den öffentlich verfügbaren Funktionalitäten auch die Möglichkeit der Bearbeitung der Daten, sowie Zugriff auf die Funktionalitäten des Crawlers gestatten. Er kann Zeitintervalle der Crawlersuche direkt beeinflussen, die Konfigurationsdatei für den Crawlersuchlauf ändern und den Speicherinhalt des Triplestore verwalten.
Zusätzlich hat er die Befugnisse weitere Administratoren zu ernennen, was über eine eingebundene MySQL-Datenbank realisiert wird.

4. Arbeitspakete und Funktionalität

Nr.	Arbeitspaket	Aufwandsanteil
1.1	**Schnittstelle zu Triplestores**	10%
1.2	**Benchmark (Vorprojekt)**	20%
1.3	**Crawler**	30%
1.4	**Sammeln/Data Fusion**	40%
1.5	**Oberfläche (User und Administrator)**	10%
	Gesamtes Projekt	***110%***

Tabellarische Darstellung der Arbeitspakete und deren Anteil

4.1 Schnittstelle zu Triplestores

Ein wichtiger Grundstein für die Auswertung und Verarbeitung der mit dem Crawler gefundenen Daten ist deren Speicherung in einem passenden Triplestore. Hierfür müssen zunächst die Daten in ein kompatibles Format gebracht werden, dies wurde durch die Vorgängergruppe bereits zu unserem Nutzen ausgearbeitet. Anschließend ist es Aufgabe des Teams, die Daten effizient in den Triplestore zu laden. Um dies zu ermöglichen, soll eine Schnittstelle programmiert werden, so dass Daten komfortabel aus dem Triplestore herausgelesen und hineingeschrieben werden können. Alternativ können vorgefertigte Frameworks dazu dienen, die Schnittstelle zu vereinfachen, indem die Frameworks für das Projekt angepasst werden.

4.2 Benchmark (Vorprojekt)

Um eine passende Datenbank für die erstellten Triple zu finden, ist es nötig einen Benchmark Test durchzuführen.
(Für detaillierte Informationen siehe Entwurfsbeschreibung Benchmark)

4.3 Crawler

Die Crawler bilden das zentrale Element des Projektes. Sie sollen Schachdaten, wie Informationen zu Schachspielern, Events und Partien, im Internet suchen.
Dabei werden sowohl ZIP-komprimierte wie auch nicht komprimierte Daten betrachtet.
Diese Daten werden zusammengetragen und anschließend zur Weiterverarbeitung zugänglich gemacht. Hierzu wird ein Zwischenspeicher aufgebaut, um effizienteres Abarbeiten der gefunden Datenbestände zu ermöglichen.
Besonderes Augenmerk gilt es auf das richtige „Durchwandern“ des Internets zu werfen, sodass nötige Webseiten besucht und nur relevante Inhalte heruntergeladen werden.
Hierzu wird eine dynamische Seed-Erweiterung implementiert, welche den Wert einer Domain an den gefundenen Schachpartien misst. Außerdem wird auch die Integration einer Blacklist von Internetdomains umgesetzt. Die Crawler werden auf Basis des Crawlerframeworks Crawler4J entwickelt. Nachdem die aus den PGN-Files gesammelten Daten in den Triplestore eingepflegt wurden, startet ein weiterer Service, der zusätzliche Informationen zu Schachspielern von der Fide-

Homepage ermittelt und Randdaten für Events und sonstige Turniere aus dem Internet gewinnt. Diese werden ebenso in den Triplestore eingelesen.

4.4 Sammeln / Data Fusion

Da eine möglichst lückenlose Datenlandschaft erstrebenswert wäre, ist es wichtig, die bei verschiedenen Quellen gefundenen Daten zu sammeln und anschließend sinnvoll zu fusionieren. Gegebenenfalls kann das Fehlen von einzelnen Informationen durch die Rekonstruktion aus vorhandenen, korrespondierenden Inhalten behoben werden.
Hierzu wird das Framework LIMES benutzt, um die Inhalte der vorliegenden PGN-Dateien mit den Daten im Triplestore zu vergleichen.
Um einen konsistenten Datenbestand zu gewährleisten, werden Regeln definiert, so dass zum Beispiel Spieler mit gleichem Namen unterschieden werden, wenn sie in verschiedenen Jahrhunderten gespielt haben.

4.5 Oberfläche (User und Administrator)

Ein weiteres Ziel ist es, eine Benutzeroberfläche zu schaffen, über die ein User bequem auf die Daten zugreifen kann. Dies umfasst eine Eingabemöglichkeit der Suchanfragen mittels SPARQL-Queries. Dafür ist auch ein gesonderter Verwaltungsbereich für Administratoren vorgesehen. Zusätzlich werden noch einige visuelle Features eingebunden, wie zum Beispiel die übersichtliche Ansicht der Daten im Allgemeinen und das Anzeigen von eventuellen Spielerbildern.
Des Weiteren wird eine Art Statistik der gestellten Anfragen oder der wichtigsten Schachdaten im Dashboard angezeigt.

5. Qualitätssicherung

Diese Qualitätsanforderungen sind während des Projektes von den Teammitgliedern zu beachten und zu erfüllen:

Produktqualität	Sehr gut	Gut	Normal	Nicht relevant
Funktionalität			x	
Zuverlässigkeit		x		
Benutzbarkeit		x		
Effizienz	x			
Änderbarkeit		x		
Übertragbarkeit		x		

Begründung:

Bei dem Projekt steht vor allem die Effizienz im Vordergrund. Sowohl die Crawler als auch der Triplestore müssen schnell die jeweiligen Daten verarbeiten, da es sonst bei den gewaltigen Datenmengen zu großen Verzögerungen kommen kann.
Weniger wichtig sind dagegen Zuverlässigkeit, Benutzbarkeit, Änderbarkeit und Übertragbarkeit, dennoch dürfen sie aus verschiedenen Gründen nicht vernachlässigt werden. Die Crawler sollten möglichst ohne Ausfälle arbeiten, weil eine bestimmte Zeit benötigt wird um eine umfangreiche Suche durch zu führen. Daher muss eine gewisse Zuverlässigkeit gewährleistet sein. Die Benutzbarkeit darf nicht vernachlässigt werden, denn die Crawler sollten schnell und unkompliziert gestartet werden können. Die Änderbarkeit muss gut sein, um kurzfristig Änderungen an den Suchbedingungen ausführen zu können. Um auf verschiedenen Servern zu laufen, muss Übertragbarkeit in einem gewissen Maße gegeben sein.
Die Funktionalität sollte keine überschwänglichen Ausmaße annehmen, da sowohl die Crawler als auch die Schnittstelle zum Triplestore nur ihre spezielle Aufgabe erfüllen und keinen erweiterten Funktionsumfang bieten müssen.

6. Glossar

siehe externes Dokument