1 Thema

Situation-Aware Semantic Service Discovery

2 Zeitraum

Sommersemester 2012 und Wintersemester 2012/2013

3 Umfang

Jeweils 8 Semesterwochenstunden

4 Veranstalter

Anna-Lena Lamprecht <anna-lena.lamprecht@cs.tu-dortmund.de>, Tel. 7736
Stefan Naujokat <stefan.naujokat@tu-dortmund.de>, Tel. 7734
Informatik Lehrstuhl 5, Otto-Hahn-Straße 14, Raum 131

5 Aufgabe

Motivation

In vielen wissenschaftlichen und nichtwissenschaftlichen Anwendungsgebieten werden Workflows oder Prozesse eingesetzt, um Aufgaben verschiedener Art service-orientiert zu automatisieren. Dementsprechend existieren zahlreiche Workflow- und Prozessmanagementumgebungen, die Softwareunterstützung für ihre Modellierung und Ausführung bieten. Sie verlassen sich jedoch zumeist darauf, dass der Nutzer die Services, aus denen er den Workflow oder Prozess konstruieren möchte, kennt und mit ihrer Handhabung vertraut ist. Das bedeutet häufig, dass der Nutzer, aus welcher Anwendungsdisziplin er auch kommen mag, sich mit den technischen Aspekten und Informatik-Begrifflichkeiten der Services beschäftigen muss, anstatt seine eigene Fachterminologie verwenden zu können.

In vielen Bereichen der klassischen und serviceorientierten Softwareentwicklung spielt in diesem Zusammenhang das Konzept von *Discovery* eine wichtige Rolle, bei der basierend auf abstrakten Beschreibungen konkrete Instanzen gefunden werden. Als einfachstes Beispiel kann hierfür zum Beispiel das *Reflection*-Konzept angesehen werden, das von vielen objektorientierten Programmiersprachen, wie z.B. Java, C# oder Python, unterstützt wird. Es ermöglicht, Klassen und Methoden aufgrund von Name und Signatur zur Laufzeit zu finden, ohne dass sie zur Kompilierzeit bereits bekannt sein müssen. Im Umfeld von *SOA* (Service-Oriented Architectures) werden Prozesse aus abstrakt definierten Services zusammengestellt (orchestriert), die dann von der Ausführungsumgebung mittels Discovery durch ausführbare Instanzen ersetzt werden. Beim *Semantic Web* werden im Rahmen einer ontologisch spezifizierten Domäne *Reasoner* oder *Rules Engines* eingesetzt, die zu Anfragen des Anwenders (Suchbegriffe, Tags, ...) gültige Dienste liefern.

All diesen Bereichen gemeinsam ist, dass qualitative Kriterien herangezogen werden, die zu einer klaren ja/nein-Auswertung führen. In der PG soll nun eine Service-Discovery entwickelt werden, welche die „Graustufen“ quantitativer Aspekte (wie z.B. statistische Verfügbarkeit, Zuverlässigkeit, Performance oder sogar Beliebtheit) mit einbezieht. Dazu sollen die Methoden von *Recommender Systems* [8, 9] eingesetzt werden.

Um die entwickelten Verfahren in der Praxis evaluieren zu können, sollen sie für die EMBOSS Toolsuite [2, 3] umgesetzt werden. Diese enthält mehr als 430 Bioinformatik-Tools[1] und ist als einzige große und frei vergügbare Bibliothek von Services bereits mit semantischen Annotationen (auf Basis der EDAM-Ontologie [4]) versehen. Auch die einzusetzende Workflowmanagementumgebung *Eclipse4Bio*,

[1] Anmerkung: Für die Anwendung dieses Beispiels und die PG im Allgemeinen werden keine Bioinformatik-Kenntnisse vorausgesetzt

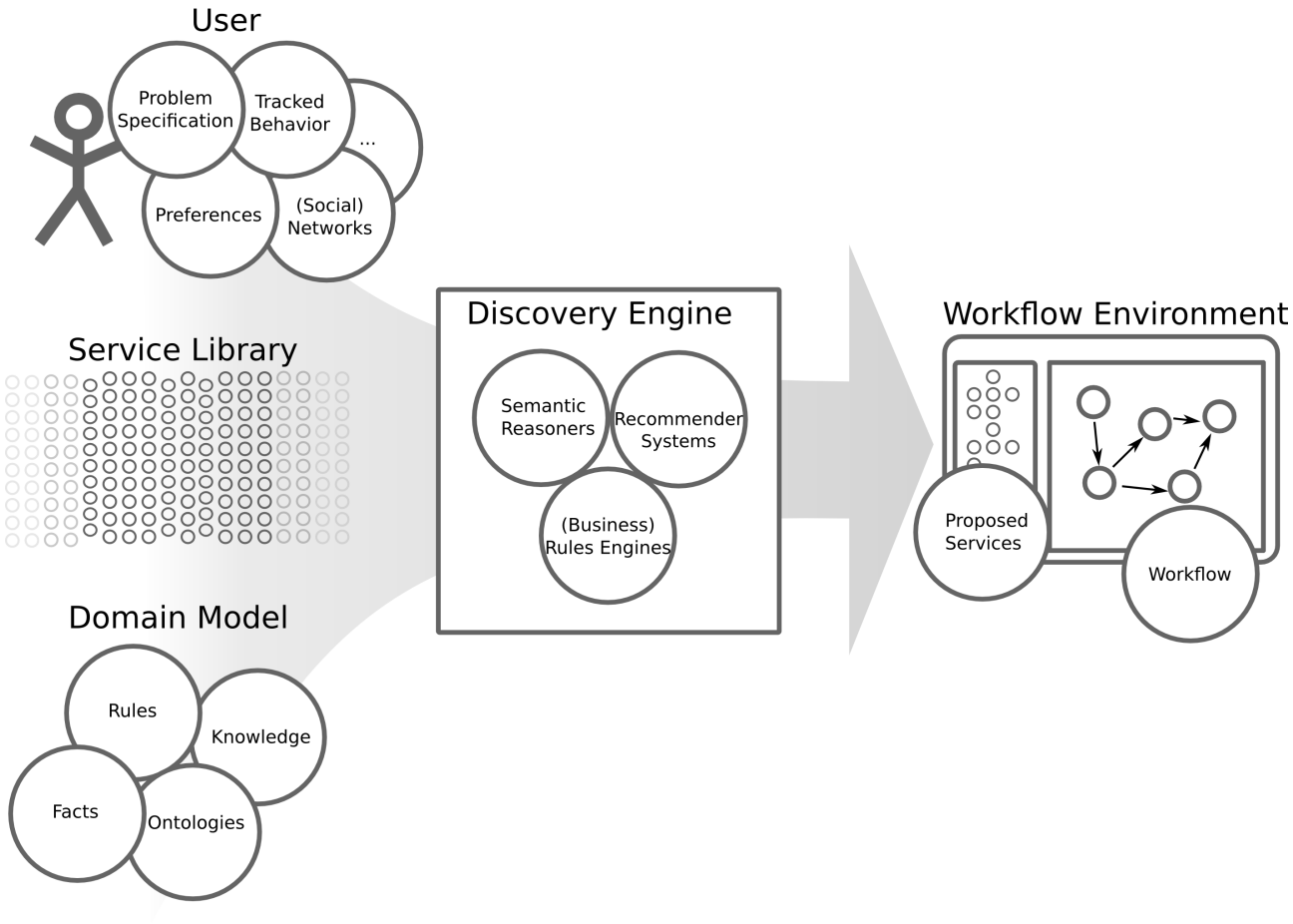

Abbildung 1: Zusammenspiel der Komponenten/Akteure

die im Rahmen einer vorhergehenden PG entstanden ist, ist bereits zielgerichtet auf das Vorhaben vorbereitet. Abbildung 1 zeigt eine Übersicht, wie die Komponenten des Systems ineinandergreifen sollen.

Technischer Hintergrund

Workflowmanagementumgebungen Im Rahmen von serviceorientierter Softwareentwicklung spielt die Modellierung von Workflows oder Prozessen aus (bestehenden) Bausteinen eine zentrale Rolle. Dabei sind folgende Aspekte von besonderer Bedeutung, die unterschiedlich stark von den verschiedenen Entwicklungsumgebungen unterstützt werden:

- Verwaltung von Servicebibliotheken und Integration bestehender Software
- (Graphische) Spezifikation/Modellierung von Workflows
- Direkte Ausführung oder Export des Workflows in eine Prozessengine

Im zweiten Semester ist für die PG vorgesehen, *Eclipse4Bio* [5] zu erweitern. Bei dieser durch eine vorherige PG entwickelte Workflowmanagementumgebung handelt es sich um eine auf Eclipse-Technologie [1] basierende Open-Source-Plattform für die Modellierung und Verwaltung von (Bioinformatik-)Workflows und Services. Sie erlaubt die Integration von Javamethoden, Kommandozeilenprogrammen und Web Services in ihre Bibliothek von Diensten, mit denen dann Workflows modelliert und anschließend ausgeführt werden können. Aufgabe dieser PG ist es somit, auf Basis der Schnittstellen für Servicedefinition, Prozessdefinition sowie Prozessausführung die zuvor entwickelte Discovery-Funktionalität zu integrieren.

Discovery Bei der Discovery werden (einzelne) Services anhand von gegebenen Kriterien identifiziert und dem Nutzer zur Verwendung vorgeschlagen. Die Kriterien können dabei direkt vom Benutzer angegeben werden (z.B. wenn Suchbegriffe eingegeben werden) oder auch aus dem Kontext abgeleitet werden (z.B. bei der Suche nach passenden Nachfolgern für Services in einem Workflow).

Für die Umsetzung der Discovery sind regelbasierte Methoden gängig. Beim *Semantic Reasoning* werden über eine Menge von Regeln und gegebenen Fakten logische Schlussfolgerungen gezogen. Ontologien bieten hierbei die Möglichkeit, das Wissen strukturiert auszudrücken. Reasoner (z.B. Jena oder Pellet) implementieren dann die logische Inferenz der Anfragen. Dahingegen stellen *Business Rules Engines* Funktionalität sicher, die durch die deklarative Beschreibung von Zusammehängen (z.B. WENN-DANN-Regeln) spezifiziert sein kann. Zusätzlich werden Ereignisse formuliert, bei deren Eintreten automatisch die entsprechende Regel angewendet wird, sofern ihre Bedingungen erfüllt sind.

Diese weit verbreiteten qualitativen Verfahren sollen nun um die quantitativen Aspekte von Recommender Systems erweitert werden. Üblicherweise schlagen Recommender Systems einem Nutzer in Abhängigkeit von über ihn bekannten Daten sowie einer gerade vorliegenden *Situation* eine Menge von Elementen vor, die dahingehend optimiert sind, wie nützlich sie potentiell für ihn sind. Dabei können die nutzerabhängigen Daten entweder explizit angegeben, oder aber implizit gesammelt (z.B. Nutzungshäufigkeit) werden. Ein prominentes Beispiel für diese Elemente sind Produktempfehlungen in zahlreichen Online-Shops: „Kunden, die diesen Artikel gekauft haben, kauften auch...“. Auch die Nähe in sozialen Netzwerken kann hierfür herangezogen werden, wenn man davon ausgeht, dass „befreundete“ Personen ähnliche Interessen verfolgen. In dem hier angestrebten System sollen vergleichbare Mechanismen für Services entwickelt werden.

Zusätzlich zur Discovery soll sich die PG mit der Fragestellung auseinandersetzen, wie mehrschrittige Empfehlungen mit den in der PG gewonnen Erkenntnissen verbessert werden können. Typischerweise wird das automatische Erzeugen von Sequenzen von Services mit Planning- oder Syntheseverfahren realisiert [10]. Dabei entsprechen die Entscheidungen, die diese Algorithmen beim Zusammenstellen der Sequenzen treffen müssen, gewissermaßen einer eingebetteten Discovery. Dieser Aspekt soll allerdings nicht im Rahmen der Projektgruppe implementiert werden, sondern ist für ggf. folgende Arbeiten (z.B. Diplom- oder Masterarbeiten) interessant.

EMBOSS und EDAM Die European Molecular Biology Open Software Suite (EMBOSS) [2, 3] ist eine umfangreiche Open-Source-Sammlung von Bioinformatik-Werkzeugen für die Analyse von biologischen Sequenzen. Die enthaltenen Tools können über verschiedene grafische und Kommandozeilen-Interfaces verwendet und in andere Anwendungen integriert werden. In sogenannten ACD-Dateien (Ajax Command Definition) werden die Tools und ihre Parameter strukturiert beschrieben, sodass sich daraus Schnittstellen zu den Tools automatisch generieren lassen.

Seit Release 6.4.0 (Juli 2011) enthalten die ACD-Dateien außerdem semantische Annotationen für die EMBOSS-Tools und ihre Parameter auf Basis der EMBRACE Data and Methods Ontology (EDAM) [4]. Die EDAM-Ontologie wurde mit dem Ziel entwickelt, kontrolliertes Vokabular für die Beschreibung von Bioinformatik-Services zur Verfügung zu stellen. So enthält sie beispielsweise Begriffe für die Klassifizierung von Datentypen, Datenformaten und Operationen.

Als sehr umfangreiche, semantisch annotierte Service-Sammlung ist EMBOSS gut geeignet, um semantische Verfahren zur Service-Discovery und zur automatischen Workflow-Erzeugung zu testen und zu evaluieren (vgl. [6, 7]).

6 Teilnahmevoraussetzungen

Für die Teilnahme an der PG sind keine Vorkenntnisse in Biologie/Bioinformatik notwendig!

Vorausgesetzt werden:

- Fundierte Kenntnisse in mindestens einer objektorientierten Programmiersprache, z.B. Java

- Kenntnisse über formale und logische Systeme, wie sie zum Beispiel in den Vorlesungen „Formale Methoden des Systementwurfs“, „Darstellung, Verarbeitung und Erwerb von Wissen“ und „Dienstleistungsinformatik“ vermittelt werden.

Wünschenswert sind zudem:

- Kenntnisse in Eclipse-Werkzeugen und -Programmierung
- Kenntnisse über Geschäftsprozesse, Web Services, Semantic Web

7 Minimalziel

Im Rahmen der PG sollen mindestens die folgenden Punkte umgesetzt werden:

- Entwicklung eines semantischen Discovery-Verfahrens für Services, das Methoden von Recommender Systems für quantitative Aspekte einsetzt
- Integration des Verfahrens in die existierende Workflowmanagementumgebung Eclipse4Bio
- Evaluation des Verfahrens anhand der EMBOSS-Toolsuite und der EDAM-Ontologie

8 Literatur

[1] Eclipse: http://www.eclipse.org/
[2] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics: TIG, 16(6):276–7, June 2000.
[3] EMBOSS Homepage: http://emboss.open-bio.org/
[4] EMBRACE Data and Methods Ontology (EDAM): http://edamontology.sourceforge.net/
[5] Eclipse4Bio: https://projekte.itmc.tu-dortmund.de/projects/eclipse4bio
[6] A.-L. Lamprecht, S. Naujokat, T. Margaria, B. Steffen: Semantics-based composition of EMBOSS services. Journal of Biomedical Semantics 2011, 2(Suppl 1):S5.
[7] A.-L. Lamprecht, S. Naujokat, B. Steffen, T. Margaria: Constraint-Guided Workflow Composition Based on the EDAM Ontology. SWAT4LS 2010.
[8] A. Klahold, Empfehlungssysteme, 2009, http://www.springerlink.com/content/978-3-8348-0568-3
[9] D. Jannach, Recommender Systems – An Introduction, 2011
[10] A.-L. Lamprecht, S. Naujokat, T. Margaria, B. Steffen: Synthesis-Based Loose Programming. 7th Int. Conf. on Quality of Information and Communications Technology (QUATIC), 2010.

9 Rechtlicher Hinweis

Die Ergebnisse der Projektarbeit und die dabei erstellte Software sollen der Fakultät für Informatik uneingeschränkt für Lehr- und Forschungszwecke zur freien Verfügung stehen. Darüber hinaus sind keine Einschränkungen der Verwertungsrechte an den Ergebnissen der Projektgruppe und keine Vertraulichkeitsvereinbarungen vorgesehen.