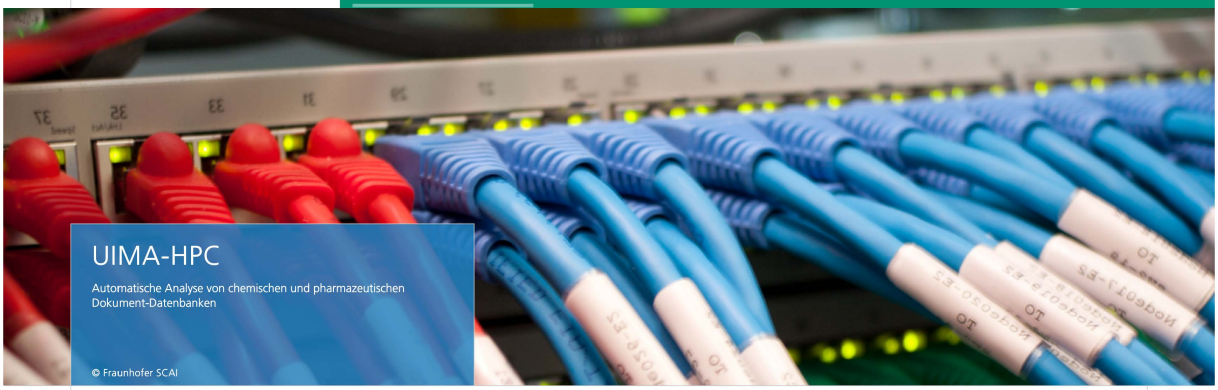
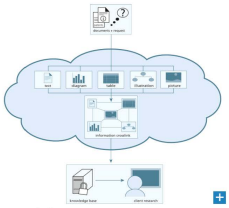


Die Suche nach der Wissens-Nadel im Daten-Heuhaufen

Über UIMA-HPC

Die unglaubliche und weltweit ständig wachsende Fülle von Fachartikeln, Patenten und anderen Nachrichtenquellen (wie z.B. Blogs) ruft sozusagen nach einem automatischen Lesen und Auswerten. So enthält die Literaturreferenzdatenbank PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) derzeit mehr als 20 Millionen Einträge auf dem biologisch-pharmakologischen Gebiet. Hier stößt die menschliche Fähigkeit, sich einen schnellen Überblick zu verschaffen, an ihre Grenzen. Die Idee dieses Projektes ist, Verfahren zu entwickeln, das bestehende Wissen in unstrukturierten Quellen schnell und effizient für neue Fragestellungen nutzbar zu machen.

Die Herausforderung in diesem Projekt ist die volle Breite der Quellen-Formate: Texte und Bilder, Spalten und Bildunterschriften, Tabellen und Diagramme, Kolumnen und Blogs, die alle automatisch, aber mit Sinn und Fachverstand interpretiert werden sollen. Deshalb werden neue Methoden der rechnerunterstützten Informationsextraktion benötigt, um Wissenschaftlern relevante Information in kompakter und strukturierter Form zur Verfügung zu stellen, welches über reine Stichwortsuchen hinausgeht.



© Fraunhofer SCAI

Der schematisierte UIMA-HPC Arbeitsablauf.

Automatische Analyse von chemischen und pharmazeutischen Dokument-Datenbanken

„Sind Strukturvariationen um dieses Grundgerüst in der Literatur bereits erwähnt worden und wenn ja: gibt es Hinweise auf deren Wirkungen (z.B. toxische oder unerwartete Wirkungen)?"; „Sind die Strukturvariationen mit Rechten Dritter belastet bzw. kann ich hierauf freien Zugriff erhalten?"; „Welche Unternehmen oder Forschergruppen beschäftigen sich seit wann und mit welcher Stoßrichtung auf meinem Arbeitsgebiet?"

Inbesondere spielt der Zeitfaktor bei der automatischen Verarbeitung eine wesentliche Rolle und erfordert deshalb eine neue Herangehensweise, welche sich moderne Rechnerarchitekturen (Mehrkernsysteme) zu nutze macht.

Die Partner – das Fraunhofer SCAI, das Jülich Supercomputing Centre, die Taros Chemicals GmbH & Co KG und die scapos AG – wollen das quasi Standardprotokoll für die Informationsextraktion UIMA in ein HPC Framework (UNICORE) einbetten, welches die effiziente Parallelisierung (Rechenzeit und Datenfluss) gewährleistet. UIMA (englisch: Unstructured Information Management Architecture) erlaubt ganz allgemein das Durchmustern von digitalen Datenströmen (Text, Audio, Bild, Video) nach Informationen.

Durch UIMA-HPC wird eine neue Klasse von Anwendungen für das Hoch- und Höchstleistungsrechnen (englisch: high performance computing – HPC) erschlossen. Mit dem neuen System für die vollständige und zeitnahe Informationsextraktion können Anwender, die bisher keinen Zugang zu HPC Ressourcen haben, diese nutzen. Die erworbene Expertise und das entwickelte System sollen Kunden als Dokumentenprozessierungsservice angeboten werden.

Hinweis:
Apache UIMA, UIMA sind registrierte Marken der Apache Software Foundation.

Kontakt

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Förderer

UIMA-HPC ist gefördert vom BMBF – Bundesministerium für Bildung und Forschung

(Förderkennzeichen: 01IH11012A).

Partner

Das Konsortium wird von Fraunhofer SCAI geleitet. Partner sind

- → Forschungszentrum Jülich
- → scapos AG
- → Taros Chemicals GmbH

Partner

→ Fraunhofer SCAI

→ Forschungszentrum Jülich GmbH (FZJ)

→ Taros Chemicals GmbH & Co. KG

→ scapos AG

Fraunhofer SCAI

Fraunhofer SCAI steht sowohl mit der speziellen Ausrichtung auf Fragestellungen und Herausforderungen der Informationsextraktion aus biomedizinischer Literatur als auch mit dem am SCAI entwickelten technologischen Ansatz zur biologischen Namenserkennung und Disambiguierung von Synonymen an international führender Stelle.

Fraunhofer SCAI bringt seine Expertise im Bereich der Informationsextraktion und die bereits entwickelten Softwarewerkzeuge in das Projekt ein. So ist die am SCAI etablierte Namenserkennung von biomedizinischen und chemischen Termen bei der Extraktion textueller Informationen von zentraler Bedeutung und Inhalt der Forschung biologisch geprägter Computerlinguistik. Nicht nur die Erkennung von Sprachvarianten, sondern auch die Zuordnung verschiedenster Synonyme zu definierten Entitäten (Disambiguierung), wie z.B. Genen oder chemischen Verbindungen, sind für die Pharma-Industrie relevant. Am Institut SCAI wurde in einem Kooperationsprojekt mit Aventis Pharma eine international konkurrenzfähige Plattform (ProMiner) zur Identifizierung und Normalisierung von Namensentitäten entwickelt. Neben der Extraktion aus Texten wurde in den letzten Jahren ein neuartiger Prototyp zur Rekonstruktion von chemischen Strukturen aus Bildern, wie sie in Patenten und anderen Veröffentlichungen vorkommen, entwickelt (chemoCR). Gerade in komplexen Dokumenten wie wissenschaftlichen Aufsätzen und Patenten spielt die Vorverarbeitung und Analyse des Layouts eine große Rolle. In Kombination von ProMiner (Text) und chemoCR (Bild) hat SCAI derzeit ein Alleinstellungsmerkmal.

Auf dem Gebiet des Text Minings konnte SCAI in öffentlichen Wettbewerben die exzellent Qualität der Lösung zeigen. Auf dem Gebiet des Image Minings in chemischen Abbildungen gibt es weltweit nur wenige akademische Gruppen, die an dem Problem forschen.

→ www.scai.fraunhofer.de

Das könnte Sie auch interessieren:

→ Forschungszentrum Jülich
→ scapos AG
→ Taros

Forschungszentrum Jülich

JÜLICH FORSCHUNGSZENTRUM

Das Jülich Supercomputing Centre (JSC) stellt für das Forschungszentrum Jülich und europaweit Rechenzeit auf Supercomputern, IT-Werkzeuge, -Verfahren und Knowhow zur Verfügung. Als nationales Höchstleistungsrechenzentrum verfügt das Jülich Supercomputing Center in der Forschungszentrum Jülich GmbH über ein breit gefächertes Wissen in der Entwicklung von Software für und im Betrieb von HPC Systemen und Hochleistungsnetzwerken.

Für UIMA-HPC stellt das JSC den Betrieb und die Wartung von UNICORE Services sowie Rechenzeit zur Verfügung. Die Grid Middleware UNICORE ermöglicht dabei die einfache und effiziente Nutzung von Grid- Ressourcen unter anderem unter Verwendung von anwendungsspezifischen Workflows.

→ http://www.fz-juelich.de/ias/jsc

Das könnte Sie auch interessieren:

→ Fraunhofer SCAI
→ scapos AG
→ Taros

Taros Chemicals GmbH & Co. KG

Taros Chemicals ist ein etablierter Anbieter für Auftragsforschung im Bereich organischer Synthesen, ansässig in Dortmund. Unsere mittlerweile 13jährige Erfahrung in der Arbeit für Pharma-, Biotech-, Agrar- und Chemiefirmen haben ihren Eindruck hinterlassen.

Taros Chemicals arbeitet hierbei nahtlos mit seinen Kunden zusammen, um Dienstleistungen abzudecken, die sonst intern durchgeführt werden müsste. Dazu gehört die Auftragsforschung und Auftragssynthese. Dabei sind wir in der Lage, die Kosten für den Kunden zu reduzieren und Entwicklungsphasen zu verkürzen. Dies ermöglicht besonders Biotech-Kunden schnell und effektiv ihre Ziele zu erreichen, ohne eigene Chemiekapazitäten aufbauen zu müssen. Damit ist gewährleistet, dass sich unsere Kunden auf ihre Kernkompetenzen konzentrieren können und ihre Ziele schnellstmöglich erreichen.

→ www.taros.de

Das könnte Sie auch interessieren:

→ Fraunhofer SCAI
→ scapos AG
→ Forschungszentrum Jülich

Teilen TEILEN TWEET MITTEILEN TEILEN

DRUCKEN

Folgen Sie uns

scapos AG

Die scapos AG wurde auf Initiative des Fraunhofer Instituts für Algorithmen und Wissenschaftliches Rechnen SCAI gegründet, um Marketing und Vertrieb der SCAI-Produkte zu verstärken. Die scapos AG, an der die Fraunhofer-Gesellschaft beteiligt ist, bietet ihre Dienstleistungen darüber hinaus auch anderen Fraunhofer-Instituten und Forschungsorganisationen an.

Alle von scapos vertriebenen Softwareprodukte zeichnen sich aus durch

- innovative mathematische Algorithmen
- industrienahe Entwicklung in Zusammenarbeit mit den Anwendern
- optimierte Performance auf modernen Hardwarearchitekturen
- nachgewiesene Reduktion von Kosten und Zeit bei Anwendern

Der Firmensitz der scapos AG befindet sich auf dem Campus des Fraunhofer-Institutszentrums Schloss Birlinghoven. Damit ist die Nähe und der kurze Draht zu Wissenschaftlern und Entwicklern sichergestellt.

→ http://www.scapos.de

Das könnte Sie auch interessieren:

→ Fraunhofer SCAI
→ Taros
→ Forschungszentrum Jülich

Teilen TEILEN TWEET MITTEILEN TEILEN

DRUCKEN

Folgen Sie uns

Publikationen

2012 | 2011

Paper von der eChallenges e-2012 Conference 17-19 Oktober, Lissabon,

UIMA-HPC – Application Support and Speed-up of Data Extraction Workflows through UNICORE

Publikationen

2012 2011

Artikel aus dem inSiDE Journal (Vol. 9 No. 2 • Autumn 2011) des Gauss Centre for Supercomputing (GCS)

UIMA-HPC: High-Performance Knowledge Mining

7th German Conference on Chemoinformatics, 2011 Goslar (FHG)

Download PDF, 1.1 MB

CGW11 - the Eleventh Cracow Grid Workshop, 2011 Cracow (FZJ)

Download PDF, 1.3 MB

Events

Event 2013

- International Supercomputing Conference ISC'13
  16-20 Juni Leipzig
- UNICORE Summit
  18 Juni Leipzig
- 3. HPC-Status-Konferenz Gauß-Allianz
  5-6 September Dresden

Events 2012

- 244th ACS National Meeting & Exposition
  19-23 Dezember Philadelphia, USA
- UNICORE Summit
  30-31 Mai Dresden
- ChemAxon's 8th European User Group Meeting
  22-23 Mai Budapest, Ungarn

Technik

Methods

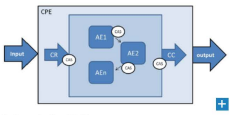


Fig. The input is converted into CAS by a collection reader (CR), further processed by a number of analysis engines (AE), and finally written back by a collection consumer (CC).

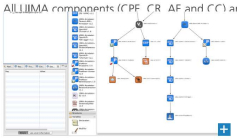
The distinctive feature of UIMA-HPC is the flexible generic approach which makes it applicable to any kind of UIMA-Pipelines and workflows thereof as well as any kind of compute resources, which are available.

UIMA pipelines are the basic building blocks of information extraction workflows. Apache UIMA provides a native Java framework for mining unstructured data. An UIMA application is organized as a Collection Processing Engine (CPE) that consists of an UIMA Collection Reader (CR), one or more UIMA Analysis Engines (AEs) and one Collection Consumer (CC). The analyzed artifact (e.g. text or binary data) is stored in the internal UIMA data structure Common Analysis Structure (CAS). The framework architecture also provides convenience methods for serializing CAS objects (XCAS) to store them persistently on hard disk. These stored XCAS files can then again be read by a CR. In our implementation we exploit this procedure to transport data between physically separated hardware nodes.

Information extraction from chemical patents

The goal of the research project UIMA-HPC is to automate and hence speed-up the process of knowledge mining in patents. Multi-threaded analysis engines, developed according to UIMA (Unstructured Information Management Architecture) standards, process texts and images in thousands of documents in parallel. UNICORE (UNiform Interface to COmputing Resources) workflow control and execution features capabilities make it possible to dynamically allocate resources for every given task to gain best cpu-time/real-time ratios in an HPC environment.

All UIMA components (CPE, CR, AE and CC) are specified via XML file format descriptors, which contain consistent predefined internal routes. For a UIMA-View converter is shown in dark-grey. Open Source software such as OpenNLP components are seamlessly integrated into one workflow together with proprietary software components (ProMiner, chemoCR) by sharing the same UIMA-TypeSystem.



Examples of implemented UIMA pipelines to process documents with medical and chemical content

INPUT	INTEGRATED 3RD PARTY SOFTWARE	FUNCTION	ANNOTATIONS	OUTPUT
PDF	CLI abbyy finereader	OCR	SourceDocument Information	XCAS
PDF	PDFbox, iText	Text extraction	SourceDocument Information	XCAS
XCAS	ProMiner	Dictionary based Annotation	Chemistry, Diseases, Genes	XCAS
XCAS	Linda	Machine Learning (ML) based Annotation	Diseases, Genes, IUPAC-terms	XCAS
XCAS	OSCAR	Dictionary and ML based annotation of chemical terms	Chemical terms	XCAS
XCAS	iText, PDFBox	Generating annotated PDF		Enriched PDF

UIMA and UNICORE

In order to make UIMA pipelines available on distributed heterogeneous resources to be accessible through UNICORE they have to meet certain requirements:

XCAS	OSCAR	Dictionary and ML based annotation of chemical terms	Chemical terms	XCAS
XCAS	iText, PDFBox	Generating annotated PDF		Enriched PDF

UIMA and UNICORE

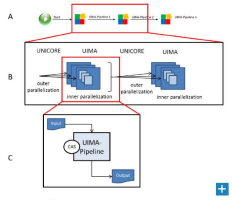


Fig. Complete architecture of the coupling between UIMA and UNICORE

In order to make UIMA pipelines available on distributed heterogeneous resources to be accessible through UNICORE they have to meet certain requirements:

- Installed on the target system,
- Executable as stand-alone applications,
- No hard-coded paths in file descriptors.

The overall architecture is shown in Figure 2. As UIMA is a native Java library it is cross platform compatible and can be installed on UNIX and Microsoft Windows based servers. The prerequisite is an installation of Java Virtual Machine 6 or higher.

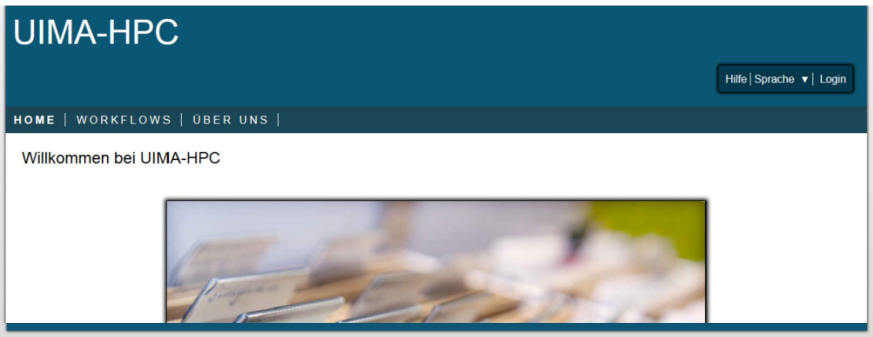
A UIMA pipeline is provided as a Java jar archive, which has to be available on a server's file system. The Input and output data format is defined in XML, it is called serialised CAS objects (C). This must be unified to be free in the choice of annotations and their order in a workflow. The Java archive is made available through UNICORE by defining it as an application resource (B). Upon execution the jar archive is called by UNICORE via a system call using the standard arguments of the Java virtual machine. The XML application configuration files support any number of

arguments that can be defined prior to execution separately for every job on the client side. UIMA provides multithreading of embedded components. This allows to exploit all cores of a node in the execution environment.

Demo

Falls Sie Interesse an einem kostenlosen Testzugang haben, dann kontaktieren Sie uns bitte per E-Mail.

WebPortal



Online Storage