KNOWLEDGE ANALYTICS FOR TECHNOLOGY INNOVATION – MASSGESCHNEIDERTE LÖSUNG FÜR DAS FRAUNHOFER INT

„KATI“ optimiert Recherchen und bietet Echtzeit-Auswertungen großer Publikationsmengen.

FRAUNHOFER-INSTITUT FÜR NATURWISSENSCHAFTLICH-TECHNISCHE TRENDANALYSEN INT

Innovation ist ein zentraler Bestandteil unserer Lebenswelt und insbesondere technologische Entwicklungen nehmen Einfluss auf unseren Alltag und unsere Zukunft. Menschen, die langfristige, strategische Entscheidungen treffen, benötigen belastbare und unvoreingenommene Information über aktuelle und erwartbare, innovative Angebote und gesellschaftliche Bedarfe.

Das Fraunhofer INT bietet diesen Entscheidern wissenschaftlich-basierte Unterstützung über das gesamte Spektrum technologischer Themen an - beispielsweise in den Bereichen Materialwissenschaften, Energie und IT sowie Luft- und Raumfahrt. Mit seiner Expertise insbesondere im Bereich der Technologievorausschau unterstützt das Institut Unternehmen damit bei ihrer strategischen Ausrichtung. Basis dafür ist die umfassende Kenntnis der wissenschaftlich-technischen Forschungslandschaft. Eine zentrale Herausforderung besteht darin, die Unmenge an zur Verfügung stehenden Informationen verarbeiten zu können. So erscheinen jede Woche mehr als 45.000 wissenschaftliche Publikationen. Um in dieser Menge relevante Veröffentlichungen, mögliche Durchbrüche oder Trends identifizieren zu können, wird ein Werkzeug benötigt, welches die Wissenschaftler am Fraunhofer INT in ihrer Arbeit effizient unterstützt.

Als ein Mehrwert für die Projektarbeit sollten darüber hinaus Auswertungen, die vorher manuell oder durch heterogenes und komplexes Tooling erstellt wurden, automatisiert werden, sodass bei Bedarf von jedem Wissenschaftler eigene, tagesaktuelle Auswertungen zu beliebigen Themen erstellt werden können, nach Möglichkeit in Echtzeit.

Fraunhofer INT

KATI
Knowledge Analytics for Technology and Innovation by Fraunhofer INT

IBM Platinum Business Partner

IBM WATSON TECHNOLOGIE ALS ERSTE GRUNDLAGE

KNOWLEDGE ANALYTICS FOR TECHNOLOGY INNOVATION – KATI

Zu Projektbeginn standen diverse Literaturdatenbanken für die wissenschaftliche Recherche zur Verfügung. Diese wiesen aber Mängel in Usability, spärliche Möglichkeiten für Auswertungen sowie verunreinigte Metadaten auf. Das Fraunhofer Institut entschied daher, einen geeigneten Datensatz für die eigene On-Premises Nutzung zu erwerben und auf dessen Grundlage ein maßgeschneidertes System zu implementieren, welches die eigenen, spezifischen Herausforderungen besser adressiert.

Die SVA hat zusammen mit dem Fraunhofer INT das KATI-System (Knowledge Analytics for Technology Innovation) entwickelt - eine komplexe Lösung, die neben umfangreichen Suchmöglichkeiten auch analytische Einsichten in die gewaltigen Mengen an Informationen bietet.

ARCHITEKTURFINDUNG

Nach einer Teststellung von **IBM Watson Explorer** sowie einem Datensatz aus dem „Web of Science“ in 2015 wurden in einem Proof of Concept Publikationen aus dem Korpus in Watson Data Explorer/Velocity und Watson Explorer Analytical Components indiziert. Untersucht wurde, inwieweit eine Search Engine bzw. Text Analytics einen Mehrwert gegenüber der bestehenden Lösung bieten konnten.

Nach ersten Tests und Bewertungen beschloss das Projektteam, sich zunächst stärker auf die analytische Komponente zu fokussieren, da die Suchtreffer, unter anderem bedingt durch mächtigere Syntax, bessere Ergebnisse lieferten. Außerdem stellten sich die zusätzlichen Auswertemöglichkeiten – schnelle Aggregationen von Metadaten über Suchtreffer hinweg – als vielversprechend heraus.

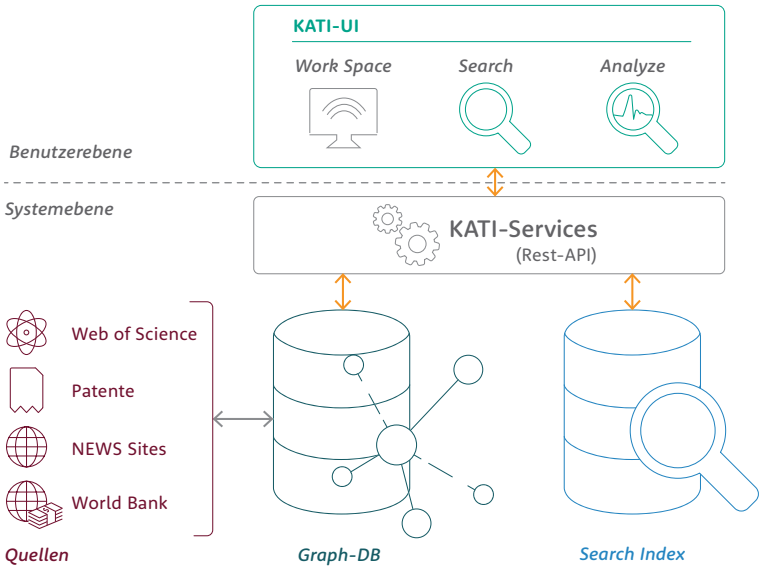
Die bibliometrischen Fragestellungen, die sich aus Zitationsbeziehungen zwischen Dokumenten oder auch Kollaborationen zwischen Personen oder Organisationen ergeben, ließen sich allerdings nicht mit einer Suchlösung beantworten. Hierfür war ein System notwendig, welches diese Beziehungen nicht nur im Datenmodell abbilden, sondern auch analytisch verwerten konnte. Das Team entschied sich strategisch für RDF als Graphentechnologie aufgrund der Möglichkeit, strukturiert und dokumentierbar eine Ontologie zu beschreiben. Außerdem verhinderten die offenen Standards ein Vendor Lock-In, so dass unterschiedliche Datenbankanbieter zur Verfügung standen.

LÖSUNGSENTWICKLUNG

In den darauffolgenden Jahren wurden in mehrmonatigen Projekten Sprints durchgeführt, die die Lösung sukzessive in einem Rapid Prototyping in die Produktion bringen konnten. Unter anderem wurden folgende Komponenten entwickelt:

- Eine umfangreiche Ontologie mit Konzepten und Beziehungen, die von generalisiert bis spezifisch die Entitäten beschreibt
- ETL-Routinen zur Transformation von XML-Daten in RDF
- Datenbereinigungsroutinen zur Deduplizierung von Ländern, Städten, Personen
- Update-Mechanismen für Scheduling, Automatisierung, Durchführung Teil-Updates von Graphen

- REST-Services für
  - Suche in Dokumenten
  - Abfragen im Graphen
  - Benutzer-Authentifizierung & Informationen
  - Verwaltung (ACL, Sharing, Datenstrukturen) von kollaborativen Workspaces
  - Verwaltung von Hintergrundaufgaben (wie Erstellung von Korpora)
- Datenschemata für IBM Watson Content Analytics, später Migration und Datenstrukturen, Clients, Datensynchronisationsroutinen für ElasticSearch
- Plug-Ins, eigene Query Parser für ElasticSearch
- Angular-basierte Front-Ends für Suche sowie Semantic Browsing und Dashboarding gegen SPARQL/RDF-Backend



TECHNOLOGISCHE BASIS

Der Kern der Lösung basiert auf zwei Softwareprodukten sowie speziell entwickelter Middleware und ETL-Tooling.

VIRTUOSO ALS ZENTRALE DATENBANK

Als zentrale Datenbank, die ohne Informationsverlust komplexe Daten speichert und Daten aus unterschiedlichen Quellen verknüpft, dient hier die Graphen-(RDF)-Datenbank **Virtuoso** von OpenLink. Diese wird wöchentlich aktualisiert und enthält mehrere hundert Millionen Entitäten mit ca. drei Milliarden Eigenschaften und vier Milliarden Verbindungen (Kanten). Die Entitäten werden über eine definierbare Logik zu logischen Dokumenten zusammengefasst und in einem ElasticSearch Cluster indiziert, der das Back End für die Suche bildet.

Eine zentrale REST-Services-Schicht (JAX-RS) bietet dabei eine einheitliche API für alle Oberflächen, die Informationen aus beiden Quellen zentralisiert zur Verfügung stellt und weitere fachliche Funktionalitäten (wie z. B. Collaboration/ACL, Request Logging, Custom Query Parsing) bietet. Diese wird von AngularJS-basierten, eigens entwickelten Oberflächen genutzt.

OPTIMIERTE SUCHE UND ANALYSE

DIE USER STORIES HINTER KATI

Anwenderseitig besteht KATI aus drei wichtigen Komponenten:

> Ein Frontend für die Suche/Recherche, mit dem Anwender einen durch wöchentliche Datenlieferungen wachsenden Datenbestand durchsuchen, filtern und sortieren können
> Kollaborationsmöglichkeiten – Suchabfragen und Suchergebnisse (Document Sets) können in hierarchischen Ordnerstrukturen (Workspaces) gespeichert und freigegeben werden. Dies dient als wichtiges Kollaborationswerkzeug für gemeinsame Recherchen.
> Entitäten-zentrische Dashboards für tiefe, interaktive Einblicke in komplexe Zusammenhänge in der Publikationswelt

Die Nutzer des Systems können sich über die Suchfunktionalität einen Überblick verschaffen und relevante Publikationen finden. Hierfür wurden spezielle Ranking-Kriterien entwickelt, die besonders informationsträchtige Dokumente herausstellen. Ergänzt wird dies durch facettierte Suche sowie State-of-the-Art-Features wie Phrasenvervollständigung.

Nach einer Vorselektion von Dokumenten (über Stichwortsuche oder Metadaten) können Korpora (oder „Document Sets“) gebildet werden, die sich in semantischen Dashboards analysieren lassen. Hierbei werden in Echtzeit (und damit tagesaktuell) graphenbasierte Auswertungen gefahren, die komplexe Zusammenhänge abbilden können.

Die Besonderheit dieser Lösung ist ein dahinterliegendes Graphensystem und eine Ontologie, die jedes Objekt im System und dazugehörige Verlinkungen navigierbar macht, so dass Entitäten wie Autoren, Publikationen, Universitäten, Länder und Organisationen in einer echten, semantischen 360°-Manier von Benutzern erkundet werden können.

ZAHLEN, DATEN, FAKTEN

Seit Dezember 2018 ist die inzwischen zweite Ausbaustufe des Tools in Produktion und wird täglich von den Mitarbeitern des Fraunhofer INT für Recherchen und Analysen genutzt, vom Institut beworben und auf Messen als echte Innovation im Bereich der Technologieforschung präsentiert.

Stand 2018 beinhaltete das System Informationen zu mehr als 60 Millionen Publikationen aus dem „Web Of Science“ sowie ca. 20.000 Indikatoren (statistische Kennzahlen) zu allen Ländern der Welt von der World Bank. Aus den 60 Millionen Publikationen entsteht eine vielfache Menge an logischen Entitäten – darunter Personen (Autoren, Editoren u. a.), unterschiedliche Dokumententypen (verschiedene Publikationsarten), Geolokationen (Adressen, Städte, Länder), Themen, Organisationen, Fachgruppen/Arbeitskreise mit mehreren Milliarden Datensätzen und mehr als vier Milliarden Beziehungen. Die Datenbank ist ca. 600 GB groß, genau wie die entsprechenden ElasticSearch-Indizes. Beide Systeme speichern Daten auf einem NetApp SSD-Array. Im Betrieb benötigt der Graph hierbei ca. 700 GB an Arbeitsspeicher, während die zwei Suchknoten mit jeweils 32 GB RAM eine sehr gute Suchperformance liefern.

KONTAKT
SVA System Vertrieb Alexander GmbH
Borsigstraße 14
65205 Wiesbaden
Tel. +49 6122 536-0
Fax +49 6122 536-399
mail@sva.de
www.sva.de

Fraunhofer INT

KATI Knowledge Analytics for Technology and Innovation by Fraunhofer INT

IBM Platinum Business Partner