Technology
Arts Sciences
TH Köln

Pressemitteilung

Nr. 9 vom 7. Februar 2017

Smarte Informationsextraktion für Literaturdatenbanken

DFG fördert Forschungsprojekt „Smart Harvesting 2“

Referat Kommunikation und Marketing
Presse- und Öffentlichkeitsarbeit
Monika Probst
0221-8275-3948
pressestelle@th-koeln.de

Technische Hochschule Köln

Postanschrift:
Gustav-Heinemann-Ufer 54
50968 Köln

Sitz des Präsidiums:
Claudiusstraße 1
50678 Köln

Freier, digitaler Zugang zu Fachliteratur ist eine Voraussetzung für hochwertige Forschungsarbeit und die Vermittlung von Wissen. Doch die immer größer werdende Publikationslandschaft macht es für Anbieter von Literaturdatenbanken schwierig, bibliographische Daten zu erheben, aufzubereiten und diese schnell an ihre Nutzer weiterzugeben. Im Forschungsprojekt „Smart Harvesting 2“ arbeiten Forscherinnen und Forscher der TH Köln, Universität Trier und des GESIS – Leibniz-Institut für Sozialwissenschaften jetzt an einer softwarebasierten Lösung zur Erfassung und Aufbereitung bibliographischer Daten aus dem World-Wide-Web. Das Projekt wird mit 414.000 Euro durch die Deutsche Forschungsgemeinschaft (DFG) gefördert. Die Software soll als Open Source für Betreiber aller Fachdisziplinen zur Verfügung stehen.

Bisher werden Internetseiten von Verlagen und Publikationsservern meist aufwendig manuell durchsucht, um bibliographische Daten für Literaturdatenbanken zu erheben. Durch die kontinuierlich steigende Zahl wissenschaftlicher Publikationen und Internetseiten stößt diese personal- und zeitintensive Arbeitsweise an ihre Grenzen. Automatisierte Verfahren bieten noch keine universelle Lösung, um Daten zu Fachliteratur effizient und qualitativ hochwertig zu sammeln: Bei der computergesteuerten Informationsextraktion suchen sogenannte Wrapper die Seiteninhalte nach strukturierten Datentexten ab. Dabei wird für jede Art von Datenstruktur ein passender Wrapper benötigt.

„Unsere bisherigen Untersuchungen haben gezeigt, dass die Entwicklung eines universell einsetzbaren, lernenden Algorithmus, der die Muster von Literaturangaben selbstständig erkennt, nicht fehlerfrei möglich ist“, sagt Prof. Dr. Philipp Schaer von der Fakultät für Informations- und Kommunikationswissenschaften der TH Köln. „Die Vielzahl der im Web verwendeten Technologien und Datenstrukturen sowie die sich dynamisch ändernden Seiteninhalte stellen immer noch eine große Herausforderung dar: Bereits nach drei Monaten ist ein bestehendes Wrappersystem veraltet und muss neu programmiert werden. Dieser Entwicklungsaufwand ist für die Einrichtungen einfach zu hoch, weshalb viele noch bei den manuellen Verfahren bleiben.“

Schwerpunkt des DFG-Projekts Smart Harvesting 2 ist deshalb die Entwicklung von wartungsarmen Wrappern, die von Nicht-Informatikern einfach bedient und laufend auf neue Website-Strukturen angepasst werden können. „Bei der Mustererkennung ist das menschliche Gehirn nämlich äußerst smart“, so Philipp Schaer. Die Idee ist, dass eine Informationsfachkraft den ersten Schritt der Mustererkennung übernimmt, in dem sie exemplarisch einen Titel, Autor, Seitenzahl etc. auf einer Internetseite markiert. Auf Grundlage der HTML-Struktur liest die Software aus diesen Angaben regelbasierte Muster für die übrigen Inhalte der Website aus.

Der Aufgabenschwerpunkt der TH Köln ist dabei, ein interaktives Interface für die Benutzerinnen und Benutzer zu bauen, mit dem sie auf beliebigen Webseiten Informationen extrahieren und diesen Prozess verwalten können. Als Basis dient die Infrastruktur der Universität Trier. Unter der Leitung von Dr. Michael Ley wurde hier mit der Computer Science Bibliography dblp ein Publikationsserver im Bereich der Informatik entwickelt, der die Daten bereits weitestgehend automatisch generiert. Das neue Interface

wird im ersten Schritt für die Weiterentwicklung von dblp und für GESIS – Leibniz-Institut für Sozialwissenschaften (Leitung Prof. Dr. Brigitte Mathiak) eingesetzt – um es anschließend in eine Open Source-Software zu überführen. So sollen die entwickelten Technologien und Lösungen auch für andere Disziplinen genutzt werden können.

Das Projekt wird von der der DFG über zwei Jahre gefördert Erste Ergebnisse für die Fachöffentlichkeit sind für Anfang 2018 geplant.

Die **TH Köln** bietet Studierenden sowie Wissenschaftlerinnen und Wissenschaftlern aus dem In- und Ausland ein inspirierendes Lern-, Arbeits- und Forschungsumfeld in den Sozial-, Kultur-, Gesellschafts-, Ingenieur- und Naturwissenschaften. Zurzeit sind mehr als 25.000 Studierende in über 90 Bachelor- und Masterstudiengängen eingeschrieben. Die TH Köln gestaltet Soziale Innovation – mit diesem Anspruch begegnen wir den Herausforderungen der Gesellschaft. Unser interdisziplinäres Denken und Handeln, unsere regionalen, nationalen und internationalen Aktivitäten machen uns in vielen Bereichen zur geschätzten Kooperationspartnerin und Wegbereiterin. Die TH Köln wurde 1971 als Fachhochschule Köln gegründet und zählt zu den innovativsten Hochschulen für Angewandte Wissenschaften.