Medieninformation

Statistiken auf den Grund gehen

Software der TU Darmstadt nutzt Daten aus dem Internet zur Interpretation von Statistiken

Kommunikation und Medien
Corporate Communications

Karolinenplatz 5
64289 Darmstadt

Ihr Ansprechpartner:
Christian Siemens
Tel. 06151 16 - 32 29
Fax 06151 16 - 41 28
siemens.ch@pvw.tu-darmstadt.de

www.tu-darmstadt.de/presse
presse@tu-darmstadt.de

Darmstadt, 16.07.2012. Informatiker der TU Darmstadt haben eine Software entwickelt, die mit Hilfe sogenannter Linked Open Data – enormen Sammlungen von semantisch vernetzten Daten im Internet – Korrelationen sowie Regeln findet und Hypothesen zur Interpretation von Statistiken aufstellt.

Die Interpretation statistischer Erhebungen, z.B. des Korruptionsindex von Transparency International, fällt häufig nicht leicht. „Es gibt zwar Verfahren, die Erklärungen von Statistiken finden. Allerdings können diese nur Daten nutzen, die in der Statistik enthalten sind; weitere Hintergrundinformationen werden dazu nicht hinzugezogen“, so Dr. Heiko Paulheim von der Knowledge Engineering Group am Fachbereich Informatik der TU Darmstadt. „Daraus entstand schließlich die Idee, Verfahren des Data-Mining, die hier erforscht werden, auf das Semantic Web anzuwenden, um zusätzliche Hintergrundinformationen zu erhalten und so mehr über Statistiken zu erfahren.“

Das von Paulheim entwickelte Tool „Explain-a-LOD“ greift auf Linked Open Data (LOD) – enormen frei verfügbaren Sammlungen von semantisch vernetzten Daten im Internet – zu und erstellt aus diesen Informationen automatisch Hypothesen zu beliebigen statistischen Daten. Dazu werden zunächst die zu interpretierenden statistischen Daten bei Explain-a-LOD eingegeben. Die Software sucht dann aus den Linked Open Data automatisch nach korrespondierenden Datensätzen und fügt diese den statistischen Ausgangsdaten hinzu. „Wenn also im Korruptionsindex das Land „Deutschland“ aufgeführt ist, werden Datensätze in Linked Open Data identifiziert, die Informationen zu Deutschland enthalten, und aus diesen zusätzliche Merkmale generiert, z.B. die Bevölkerungszahl, die Mitgliedschaft in der EU und OECD oder die Anzahl von Firmensitzen“, erklärt Paulheim. Um den Umfang der angereicherten statistischen Daten zu reduzieren, werden Merkmale, die voraussichtlich keine brauchbaren Hypothesen liefern, automatisch entfernt.

Nach der Datenaufbereitung erstellt Explain-a-LOD im zweiten Schritt aus den angereicherten Daten automatisch Hypothesen. Hierzu werden zum einen einfache Korrelationsanalysen und zum anderen Regellernverfahren eingesetzt, um komplexere Erklärungsansätze zu entdecken, die mehr als ein Merkmal beinhalten. Schließlich werden dem Nutzer die gefundenen Hypothesen präsentiert, z.B. in Form von Sätzen wie *Ein Land des Typs OECD-Mitglied hat einen niedrigen Korruptionswahrnehmungsindex*, wenn

eine positive Korrelation zwischen dem Merkmal OECD-Mitgliedschaft und dem Zielattribut Korruptionswahrnehmungsindex vorliegt. Dazu muss in der ursprünglichen Statistik nicht erhoben worden sein, ob es sich um OECD-Mitgliedsstaaten handelt oder nicht; dieses Hintergrundwissen wird von Explain-a-LOD automatisch hinzugezogen.

Überraschende und nützliche Hypothesen

Paulheim und seine Kollegen haben ihren Ansatz an verschiedenen Statistiken eingehend getestet, unter anderem an der Mercer-Studie zur Lebensqualität und dem Korruptionswahrnehmungsindex von Transparency International. „Man erhält eine Mischung aus naheliegenden und überraschenden Hypothesen, wie *Städte, in denen es im Mai nicht wärmer als 21°C wird, haben eine hohe Lebensqualität; Hauptstädte haben generell eine geringere Lebensqualität als Nicht-Hauptstädte*, oder *Staaten mit wenigen Schulen und Radiosendern haben einen hohen Korruptionswahrnehmungsindex*“, erläutert Paulheim. Eine Evaluierung der Ergebnisse durch Probanden konnte diesen Eindruck bestätigen. „Die Testpersonen empfanden die Hypothesen überwiegend als überraschend sowie als nicht-trivial und vielfach auch als nützlich“, so Paulheim. Größere Zweifel hätten die Probanden aber bei der Vertrauenswürdigkeit der Hypothesen gehabt. Dies sei auch darauf zurückzuführen, dass die Qualität der Daten in der Linked Open Data Cloud nicht immer zufriedenstellend sei, wie Paulheim bemerkt.

Explain-a-LOD wurde in den vergangenen Monaten auf mehreren internationalen Konferenzen vorgestellt. Ende Mai erhielt das Tool bei der Extended Semantic Web Conference 2012 auf Kreta Auszeichnungen als „Best In-Use Paper“ und „Best Demo“. Für die Zukunft sind einige Weiterentwicklungen an Explain-a-LOD geplant: u.a. sollen weitere Algorithmen zur Merkmalgenerierung implementiert und zudem weitere Datenbestände aus der LOD-Cloud herangezogen werden.

Weitere Informationen

Das Tool „Explain-a-LOD“ ist auf den Webseiten der Knowledge Engineering Group als Open Source verfügbar:
http://www.ke.tu-darmstadt.de/resources/explain-a-lod
Eine Infografik zum Tool kann unter www.tu-darmstadt.de/pressebilder heruntergeladen werden.

Pressekontakt

Dr. Heiko Paulheim, Tel. 06151 / 16-6634, paulheim@ke.tu-darmstadt.de

MI-Nr. 62/2012, pb/csi