Information Mining:
Aus dem Steinbruch der Wissenschaft

Ulrich Herb
scinoptica Wissenschaftsberatung & Publikationsberatung
http://www.scinoptica.com
u.herb@scinoptica.com

16.02.2013

[scinoptica]

Erschienen als: Herb, U. (2013). Information Mining:
Aus dem Steinbruch der Wissenschaft. irights.info, (14.02.2013).
http://irights.info/information-mining-aus-dem-steinbruch-der-wissenschaft/11499

Information Mining: Aus dem Steinbruch der Wissenschaft

Große Massen an Forschungsdaten werden mit Techniken des „Information Mining" maschinell ausgewertet, um neue statistische Muster zu entdecken. Das wirft technische und rechtliche Fragen und Probleme auf: In PDF-Dateien lässt sich schlecht nach Daten schürfen. Soll Information Mining in der Wissenschaft allgemein erlaubt oder lizenzrechtlich geregelt werden?

In nicht wenigen Wissenschaftsdisziplinen verschwimmen die Grenzen zwischen Text und Software, etwa wenn man *living documents* betrachtet, die Updates unterliegen, oder dazu übergeht, Texte in Umgebungen wie *Github* oder *Figshare* kollaborativ zu entwickeln. Wenn man Texte als eine Art kompilierte Software ansieht, sollte man auch deren Quelltexten, den Forschungsdaten, Aufmerksamkeit schenken. Denn wie Jenny Molloy von der Open Knowledge Foundation resümiert: „Science is built on data" (Molloy, 2011).
Textpublikationen dokumentieren die Schaffung eines Wissensstands, die in Form von Zitaten oder Projektbewilligungen belohnt wird. Die zugrundeliegenden Daten bleiben oft verborgen – es sei denn, man stellt sie im Open Access bereit. Dies birgt gewisse Risiken: Wissenschaftler, die keinen Beitrag zur Erhebung leisteten, könnten die Daten auswerten und den ursprünglichen Datenproduzenten zur Konkurrenz werden. Andererseits potenziert die offene Zugänglichkeit den wissenschaftlichen Fortschritt und die Verwertung der Daten, da unzählige Wissenschaftler sie auswerten können. Diese Crowd-Komponente der Datennutzung wird ergänzt durch die technischen Möglichkeiten des Data Mining. Digital vorliegende Forschungsdaten werden automatisiert und rechnergestützt ausgewertet – ob Datenreihen, Tabellen, Graphen, Audio- und Videodateien, Software oder Texte.

Muster in Datenbergen entdecken

Digitale Verfügbarkeit und maschinelle Auswertungen kennzeichnen den Aufstieg der *data-driven science*, die statistische Muster in schier unendlichen Daten ausmacht, um diese anschließend wissenschaftlich zu erklären. Dieser Ansatz ergänzt die traditionelle *theorie- und hypothesengetriebene Wissenschaft*, die von Theorien ausgeht, Hypothesen ableitet, Erhebungsinstrumente entwirft, dann Daten erhebt und anschließend analysiert. Um die Möglichkeiten der neuen Methoden auszuschöpfen, sollten die Daten jedoch offen verfügbar sein. So verlangen es zum Beispiel die *Panton Principles*, die fordern, dass Forschungsdaten auf jede mögliche Art offen genutzt, ausgewertet und weiterverbreitet werden dürfen, solange die Datenproduzenten genannt werden. Sogar diese Bedingungen entfallen, wenn die Resultate in die *public domain*, in die Gemeinfreiheit entlassen werden.

Stochern in PDF-Dateien

In der Praxis sind Forschungsdaten zwar teils verfügbar – sei es nur für Subskribenten wissenschaftlicher Journale oder auch für jedermann – offen sind sie jedoch nicht unbedingt: Weder rechtlich, denn selbst Informationen in auslesbaren Formaten stehen längst nicht immer unter einer Lizenz, die Data Mining ausdrücklich erlaubt. Noch technisch, denn oft finden sich Daten in versiegelten PDF-Dateien, die nicht maschinell ausgewertet werden können. Ein Umstand, den die Open-Science-Community pointiert mit der Analogie beschreibt, Daten aus einer PDF-Datei zu extrahieren gleiche dem Versuch, aus einem Hamburger wieder ein Rind zu machen. Gegen das Text- und Data-Mining positionieren sich kommerzielle Akteure, deren Geschäftsmodell auf der Verknappung von Information basiert: In einer Konsultation des Intellectual Property Office in Großbritannien sprachen sich zahlreiche dieser Informationsanbieter gegen eine Blankoerlaubnis zum Content-Mining copyrightbelasteter Inhalte zu wissenschaftlichen Zwecken aus – selbst wenn die Institution eines Forschers auf die Inhalte via Subskription zugreifen darf und obwohl die Forschungsergebnisse mit öffentlichen Geldern produziert wurden. Einige der Informationsanbieter schlu-

gen vor, den Zugang über Lizenzierungen zu regeln, die allerdings vermutlich – dem traditionellen Geschäftsmodell folgend – kostenpflichtig sein dürften. Dem Chemiker Peter Murray-Rust etwa gestattete ein Verlag nach zwei Jahren zäher Verhandlung das Text-Mining von Publikationen, jedoch nur wenn die Rechte an den Resultaten an den Verlag fielen und nicht öffentlich zugänglich gemacht würden.

Nutzen der Offenheit

Volkswirtschaftlich betrachtet haben Data- und Text-Mining jedoch ungeheures Potential: Ihre Anwendung in der Wissenschaft könnte nach einer McKinsey-Studie der europäischen Wirtschaft eine Wertschöpfung von 250 Milliarden Euro pro Jahr bescheren. Das setzt aber voraus, dass Informationen offen verfügbar sind, denn der Ausschluss kommerzieller Daten-Nutzung verhindert, dass neue Dienste und Produkte entwickelt werden.
Murray-Rust etwa entwickelte Techniken zum Data-Mining kristallographischer Daten, deren Ergebnisse sehr fruchtbar für die Schaffung neuer medizinischer Wirkstoffe sein können. Wenn es nicht erlaubt ist, die ausgewerteten Daten kommerziell zu verwerten, werden Pharmafirmen vor der Verwendung Abstand nehmen. Nicht zuletzt ermöglicht Text- und Data-Mining auch effizienteres Information Retrieval, etwa wenn Forschern Empfehlungsdienste nach einer Analyse relevante Daten oder Texte vorschlagen und aufwändige Recherchen abkürzen.

Literatur

Molloy, J. C. (2011). The Open Knowledge Foundation: open data means better science. *PLoS biology*, *9*(12), e1001195. doi:10.1371/journal.pbio.1001195