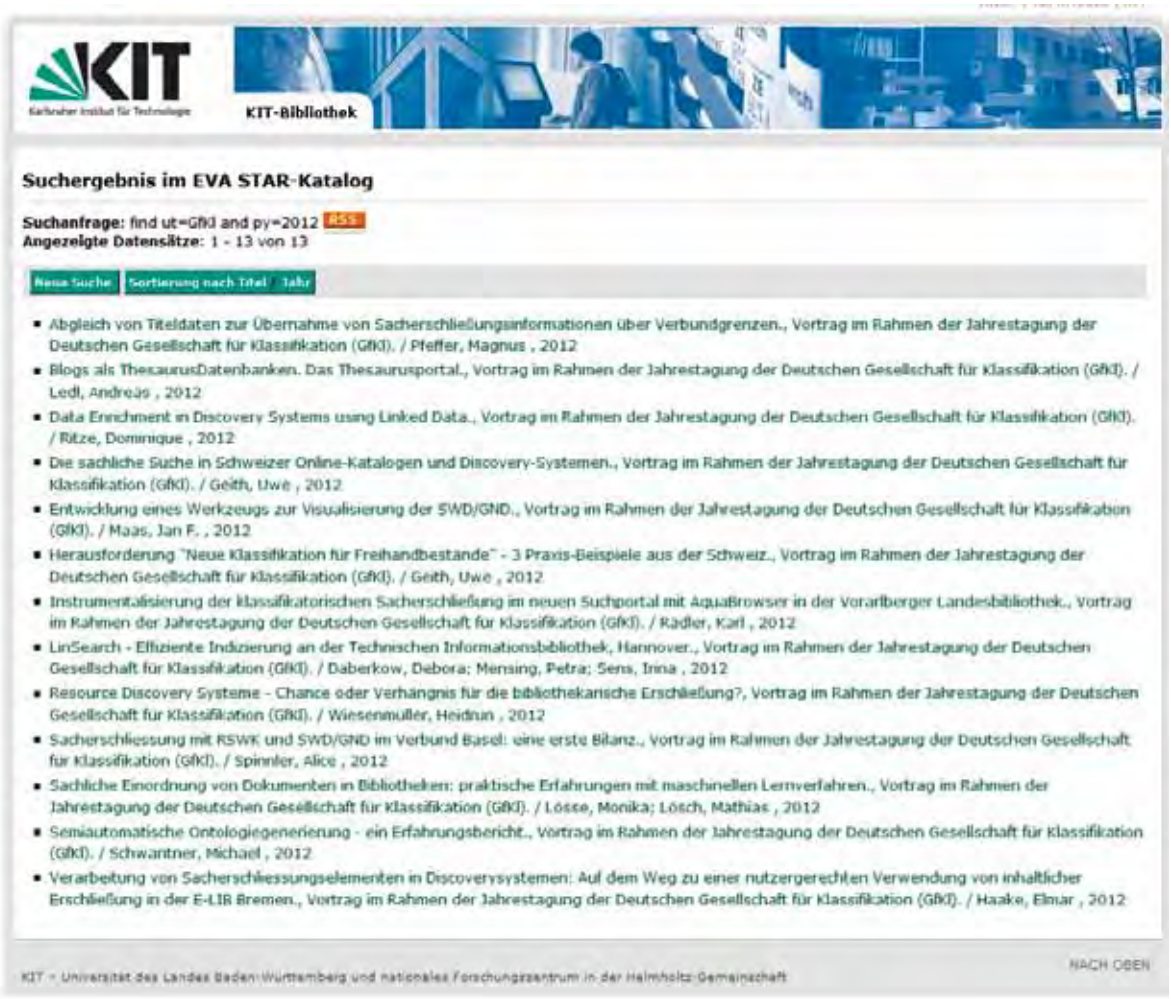
GfKl 2012: Die Konferenz der Datenverwertungsforscher

Bericht über die 36. Jahreskonferenz der Deutschen Gesellschaft für Klassifikation (GfKl) mit integriertem Workshop on Classification and Subject Indexing in Library and Information Science (LIS'2012), Hildesheim, 1. - 3. August 2012

Vera Münch

Man könnte sie auch Informations- oder Wissensentdecker nennen, die rund 200 Datenexperten aus Deutschland, Japan, USA, Canada, Zypern, Polen, Italien, Frankreich und Skandinavien, die Anfang August in Hildesheim Methoden, Ansätze und Lösungen zum automatisierten Sammeln, Ordnen, Kombinieren und Weiterverwerten von Daten diskutierten. Die Wissenschaftlerinnen und Wissenschaftler observieren die Datenflüsse in den Netzen und suchen nach Möglichkeiten, diese dabei so mit Beschreibungen zu ersehen, dass später eine gezielte inhaltsbezogene Analyse möglich wird. Zudem treibt sie dabei der Wunsch, Voraussetzungen für die Berechnung vielfältiger Prognosen zu schaffen; zum Beispiel, um Marktentwicklungen und Käuferverhalten vorherzusagen oder verschiedene Möglichkeiten zur Lösung komplexer Aufgaben planerisch durchspielen zu können. Das alles würden sie am liebsten automatisch von Maschinen erledigen lassen, die zu entwickeln sie sich ebenfalls auf die Fahnen geschrieben haben. Unter der Überschrift „Data Analysis, Machine Learning and Knowledge Discovery" diskutierten sie auf der GfKl 2012 den aktuellen Stand der Erkenntnisse ihres Fachgebietes.

Mittendrin in diesem abstrakten Konferenzumfeld beschäftigten sich rund 40 Bibliothekswissenschaftlerinnen und -wissenschaftler auf dem Workshop LIS'2012 mit der praktischen Seite der Informations- und Wissensorganisation; unter anderem mit automatischer und teilautomatischer Indizierung, Sacherschließung und Ontologieerstellung.



Vortragsfolien-KIT-BIB.png. Titelliste der LIS-Vorträge, Folien sind dort verlinkt

Wenn man als Laie verstehen möchte, an welchen Aufgaben Wissenschaftlerinnen und Wissenschaftler einer Community forschen, fragt man sie am besten nach exemplarischen Anwendungen. Wenn aus ihren Erkenntnissen Methoden und Werkzeuge für Empfehlungsmarketing im eCommerce, zur Operationsplanung in Finanzinstituten und Krankenhäusern, zur Verkehrslenkung und zum Spritsparen, für den Kopierschutz von Musikstücken, für Statistiken in der Bioinformatik und in der Bildungsforschung, für den Erkenntnis- und Erfahrungsaustausch in der Wissenschaft und für personalisierte Medizin ebenso entstehen wie für die elektronischen Sammel- und Ordnungssysteme zum Bestandsmanagement in Bibliotheken, muss der Forschungsgegenstand Grundlagenforschung sein. In den 130 Vorträgen der GfKl 2012, 16 davon im Rahmen des LIS'2012 Workshops, tauchten alle aufgezählten Anwendungsfelder mindestens einmal auf. Das gemeinsame Interesse der Konferenzteilnehmerinnen und Teilnehmer ist, Registrier- und Verarbeitungssysteme zu schaffen, die es ermöglichen, die riesige Mengen der heute weltweit verfügbaren Daten so auszuwerten, dass damit spezifische Fachfragen aus unterschiedlichsten Anwendungsgebieten beantwortet werden können und – so der Schwerpunkt der aktuellen Forschungsarbeiten – neue Algorithmen und Softwaremaschinen immer mehr Aufgaben ohne menschliches Zutun automatisch bearbeiten. Der Konferenzband „Data Analysis, Machine Learning and Knowledge Discovery", Proceedings of the 36th Annual Conference of the German Classification Society (GfKl) mit der Dokumentation der Vorträge wird bei Springer Science + Business Media in der Reihe „Studies in Classification, Data Analysis and Knowledge Organization (http://www.springer.com/series/1564) erscheinen. Die Folien zu den Vorträgen im Rahmen des LIS'2012 Workshops sind, soweit von den Referenten freigegeben, von der Bibliothek des Karlsruher Instituts für Technologie (KIT) im Netz breitgestellt: http://tinyurl.com/bs6ewnv

Idyllischer Konferenzort: Die Stiftungsuniversität Hildesheim nutzt seit ein paar Jahren die historischen Domäne Marienburg im Süden der Stadt für Lehrveranstaltungen. Die umfangreichen Umbauten und Renovierungsarbeiten sind noch nicht abgeschlossen.

Maschinen erkennen Muster in den Daten

„Datenanalyse und Maschinelles Lernen sind wie die meisten Technologi-

Der Präsident der Universität Hildesheim, Professor Dr. Wolfgang-Uwe Friedrich, begrüßt die Teilnehmer.

en anwendungsneutral: Man kann Prozesse für Menschen verbessern oder sie mit Werbung überhäufen“, so Professor Dr. Dr. Lars Schmidt-Thieme, Leiter des Bereiches Wirtschaftsinformatik und Maschinelles Lernen an der Universität Hildesheim. Schmidt-Thieme hatte gemeinsam mit seiner Kollegin Prof. Dr. Myra Spiliopoulou von der Otto von Guericke Universität Magdeburg die Programmleitung für die GfKl 2012 übernommen und die Konferenz in Hildesheim ausgerichtet. Ruth Janning, Informatikerin (M.A.) am Lehrstuhl von Schmidt-Thieme, unterstützte bei der lokalen Organisation. Janning ist begeistert vom Forschungsgebiet Maschinelles Lernen: „Datenvolumen wachsen – da sind so viele Informationen versteckt, die man alleine gar nicht finden kann. Also kommen Maschinen ins Spiel, erkennen Muster, werten Datenmengen aus, klassifizieren sie.“ So könnten zum Beispiel Instrumente aus Musikstücken erkannt oder die Kreditwürdigkeit von Bankkunden erklärungsfähig bewertet werden.

Regelungen zum Umgang mit personenbezogene Daten angemahnt

Datenanalyse sei unter dem Stichwort „Big Data“ zur Zeit in allen Unternehmensbereichen von Bedeutung, schrieb die Stiftung Universität Hildesheim in der Presseinformation zur GfKl 2012. Und weiter: in jüngster Zeit werde zudem über Datenauswertung zur Gesichtserkennung, für den Adresshandel und Krankenkassen-Cards debattiert. Schmidt-Thieme weist in diesem Zusammenhang darauf hin, dass die gesellschaftliche Debatte darüber, wie mit den Datenmengen umgegangen wird, sehr wichtig sei: „Bei der Anwendung dieser Technologien, insbesondere auf personenbezogene Daten, bedarf es klarer gesellschaftlicher, zum Beispiel gesetzlicher Regelungen.“ (Siehe auch Interview ab Seite 486.)

Ist der Versuch der Wissensorganisation ein Kampf gegen Windmühlenflügel?

Auch wenn die Systeme für Empfehlungsmarketing, Finanzmarktanalysen oder semantisch verknüpfte Informationssuche in den letzten zehn Jahren deutlich fortgeschritten sind, steht die Wissenschaft im Bezug auf effiziente und zuverlässige Datenverwertung noch vor ganz großen Herausforderungen. Während die Forscherinnen und Forscher nach Lösungen für eine zeitgemäße, zu den heute genutzten Informations- und Kommunikationstechnologien passende Wissensorganisation suchen und immer bessere Maschinen und Werkzeuge entwickeln, explodieren die Datenmengen, bringen Verlage und Unternehmen der neuen Netzwirtschaft ständig neue Werkzeuge und Plattformen auf den Markt, immer neue Softwaretechnologien setzen sich durch und zu guter Letzt publiziert die Wissenschaft im Zuge von Open Access auf unzähligen Universitäts-, Instituts- und Privatservern direkt.
Sie nutzen dafür frei verfügbare Plattfor-

men und Software wie Wordpress-Blog oder Mendeley, aber auch proprietäre Programme und immer neue Formate, die vielleicht an der Universität selbst entwickelt wurden.

Vor diesem Hintergrund muss die Frage gestellt werden, ob man diese chaotischen Datenfluten durch formale Beschreibung und die Entwicklung von Softwaremaschinen unter Kontrolle bekommen kann, oder ob sich die Welt von ihrem Versuch der kontinuierlichen Dokumentation wissenschaftlichen Wissens verabschieden muss. Auf der GfKl wurde diese Frage in dieser deutlichen Form nicht diskutiert. Aber Zweifel an der Machbarkeit zuverlässiger automatischer Wissensorganisation blitzten in den Diskussionen nach den Vorträgen immer wieder einmal auf. Die Forschung steht hier noch ziemlich am Anfang.

Ist maschinelles Lernen ohne Vorwissen möglich?

Einig sind sich alle Experten, dass die von Menschen und Maschinen produzierten Datenmengen nur noch durch Automatisierung bewältigt werden können. Wie weit diese gehen kann, darüber ist man sich noch nicht schlüssig. Bisher ist es den Forscherinnen und Forschern noch nicht gelungen, universelle Werkzeuge zu schaffen, die Daten unabhängig vom Fachgebiet, also ohne vorgegebenes oder trainiertes Domänenwissen, sinnvoll analysieren können. Shai Ben-David, Professor an der University of Waterloo, Kanada ging in seinem Hauptvortrag „Is learning possible without prior knowledge“, Untertitel: „Do Universal learners exist” der Frage nach, ob es theoretisch überhaupt möglich sein könnte, einen universell einsetzbaren – maschinellen – „Lerner“ zu kreieren. Mit seiner Kollegin Ruth Urner in Waterloo und in Kooperation mit dem Toyota Technological Institute in Chicago arbeitet er an dieser Fragestellung. Ben-David lieferte in seinem Vortrag verschiedene mögliche Definitionen für „universelles Lernen“. Unter anderem berichtete er über Tierexperimente mit Tauben und Ratten, bei denen die Tiere dazu gebracht werden sollten, inhärentes Wissen auf einen neuen Zusammenhang zu übertragen. Dafür wurden bestimmte Standardsituationen bei der Futtersuche mit neuer Information ergänzt. Es konnte nachgewiesen werden, dass die Konditionierung mit den neuen Informationen nicht dazu führte, dass inhärentes Wissen auf die neue Situation zur Lösung von Problemen resp. zum Erkennen von Problemen übertragen worden wären. Weiterentwickelt für die Informatik zog Ben-David daraus den Umkehrschluss, dass „Uniform Learning“ eigentlich möglich sein müsste. Er stellte die These auf, dass es universelle maschinelle Lerner geben kann, allerdings mit der Einschränkung, dass diese nur entweder zur Laufzeit oder zur Musterkomplexität (Sample Complexity) funktionieren könnten. Beides gleichzeitig ginge nicht. Sein Vortrag löste intensive Diskussionen unter den Experten aus.

Stream Data Mining, Anytime Algorithmen und Data Clumping

Auf ähnlich hoch abstrakter Ebene mathematischer und informatischer Forschung bewegten sich die weiteren Hauptvorträge und auch viele Einzelvorträge. Es ging um Theorien und Methoden für die automatische (permanente) Auswertung von Datenströmen in stationären und mobilen Netzwerken (z.B. mit Anytime-Algorithmen, die Prof. Dr. Thomas Seidl, RWTH Aachen im Vortrag „Stream Data Mining and Anytime Algorithms“ vorstellte), um die Überlappung von Datenanalyse und Graphentheorie zur Charakterisierung von Objekten (Prof. Dr. Wolfgang Gaul, KIT Karlsruhe) und um Szenariobäume, die bei der Optimierung von Finanztransaktionen (Prof. Alois Geyer, Uni Wien, Prof. Dr. Michael Hanke, Universität Lichtenstein, Dr. Alex Weissensteiner, FU Bozen) helfen. Weiter wurden Data Mining aus verteilten Maschinen im Umfeld der Fahrzeug-zu-Fahrzeug- und Maschine-zu-Maschine-Kommunikation diskutiert ebenso wie autonome Roboter und ihr Lernverhalten im Roboterschwarm und in der Interaktion mit Menschen (Prof. Dr. Michèle Sebag, University Paris-Sud, CNRS, Frankreich). Data Clustering wurde versus Data Clumping gestellt – wo immer auch der Unterschied liegen mag.

Nicht nur die Forschungsinhalte der verschiedenen Anwendungsbereiche, auch die Terminologie zur Beschreibung der Wissensorganisation forderte die Zuhörenden. Von den

Professor Dr. Ulrich Müller-Funk vom Institut für Wirtschaftsinformatik der Universität Münster hinterfragt die Ansätze von Professor Shai Ben-David, University of Waterloo, Kanada zu einer möglichen universellen Methode.

Wissenschaftlerinnen und Wissenschaftlern der verschiedenen Fachrichtungen werden zwar oft die gleichen Begriffe verwendet. Sie haben allerdings im jeweiligen Fach eine abweichende Bedeutung; beispielsweise die Worte Klassifikation und Indexierung. Schmidt-Thieme sagte dazu: „Klassifikation ist ein ganz abstraktes Setting und es gibt unendlich viele Anwendungen." In der Datenanalyse und dem Maschinellen Lernen versteht man nach seiner Erklärung unter Klassifikation, „dass man in einem technischen System Informationen protokolliert und entsprechende Handlungsoptionen hat, entsprechende Aktionen ausführen kann und dass man versucht, daraus Zusammenhänge zwischen Sensorinformationen und möglichen Aktionen zu lernen". Ein Sensor (oder Prädiktor, je nachdem, von welcher Seite man ihn betrachtet) ist für die Datenverwertungsforscher übrigens ein Daten- respektive Informationsabgriffspunkt. Der Sensor beobachtet die Datenströme, kann von Abfragewerkzeugen ausgelesen werden oder meldet Abweichungen von vorgegebenen Toleranzen.

LIS-Integration als Versuch einer Wiederannäherung

Bibliothekarinnen und Bibliothekare sind traditionell von Anfang an auf der GfKl-Jahreskonferenz vertreten. Allerdings hatten sich Konferenz und Workshop in den letzten Jahren so weit voneinander entfernt, dass die Veranstaltungen praktisch nur noch parallel stattfanden. In diesem Jahr nun sollte die Trennung durch die Integration ins Konferenzprogramm wieder aufgehoben und, wie es der Präsident der GfKl, Professor Dr. Claus Weihs von der TU Dortmund in seiner Ansprache zur Eröffnung des LIS Workshops ausdrückte, „darüber gesprochen werden, wie es mit der Annäherung weitergehen kann". Vor allem der Leiter der Bibliothek des Karlsruher Instituts für Technologie (KIT), Dr. Frank Scholze, macht sich für eine wieder engere Verbindung mit der GfKl stark. Scholze hat den LIS'2012 Workshop mit organisiert und eine Sitzung moderiert. Der Bibliotheksleiter sieht viele Parallelen bei den Aufgabenstellungen, mit denen wissenschaftliche Bibliotheken konfrontiert sind, und den Fragen, die von GfKl-Wissenschaftlern bearbeitet werden – Tendenz steigend. Für ihn steht außer Frage, dass ein gegenseitiger Wissensaustausch erfolgversprechend und fruchtbar ist.

Was bringt die GfKl den Bibliothekswissenschaftlern?

Durch die Einbindung ins Programm konnten sich die LIS-Teilnehmer die Hauptvorträge anhören, ohne einen Workshopbeitrag zu verpassen. Sie hatten aber auch die Möglichkeit, andere Vorträge zu besuchen. Angesichts der in den GfKl-Vorträgen diskutierten Themen durfte man sich aber schon fragen, was diese Bibliothekswissenschaftlerinnen und -wissenschaftlern bringen. „Doch, das sind die Themen, die uns auch gerade beschäftigen", antwortete Dr. Tamara Pianos von der Zentralbibliothek Wirtschaftswissenschaften (ZBW) in Kiel. Dr. Irina Sens, stellvertretende Direktorin der Technischen Informationsbibliothek (TIB) in Hannover und federführend bei vielen TIB-Entwicklungen in den letzten Jahren, wies pragmatisch darauf hin, dass „man ja nicht jedes Detail der Forschungsarbeiten verstehen muss". Wichtig sei es zu sehen, wie andere Experten an die Aufgabenstellung herangingen und welche Methoden sie zur Lösung vorschlagen würden. Inhaltlich biete die Konferenz durch ihre interdisziplinäre Ausrichtung gute Chancen, neue Forschungsansätze für die eigene Arbeit oder auch für mögliche neue Kooperationen kennen zu lernen, auf die man im eigenen Fachgebiet vielleicht nicht gestoßen wäre.

Automatische Zuordnung von Metadaten im LIS-Workshop

Sens stellte auf dem LIS-Workshop im Vortrag „LinSearch – Effiziente Indizierung an der Technischen Informationsbibliothek (TIB) Hannover" ein semiautomatisches Verfahren für die Erschließung von Einträ-

Dr. Frank Scholze, Leiter der Bibliothek des Karlsruher Instituts für Technologie (KIT), macht sich für den Wissensaustausch zwischen den Forscherinnen und Forschern in der GfKl und den Anwendern in Bibliotheken stark.

GfKl 2013 – wieder mit LIS'2013 – in Luxemburg

Im kommenden Jahr findet die Konferenz als „European Conference on Data Analysis" vom 10. bis 12. Juli in Luxemburg an der University of Luxembourg (UL) statt. Sie wird von der Deutschen Gesellschaft für Klassifikation (GfKl) und der Französischen Fachgesellschaft für Klassifikation (SFC) gemeinsam veranstaltet. Der LIS'2013 Workshop „Klassifikation und Sacherschließung" ist für den 10. und 11. Juli vorgesehen. Der Aufruf zur Einreichung von Beiträgen für LIS'2013 ist bereits veröffentlicht. Abstracts können über die Konferenzwebseite eingereicht werden. Konferenzsprache ist diesmal englisch.
http://gfkl2013.lu/

gen ins Fachportal GetInfo vor. Bereits jetzt weist GetInfo 45 Millionen Objekte nach. „Wegen der exponentiell anwachsenden Menge an verfügbaren Informationen ist es kaum noch möglich, alle Objekte manuell zu klassifizieren, deshalb setzen wir jetzt zur Klassifizierung der Metadaten ein vierstufiges, semiautomatisches Verfahren ein", berichtete sie. In der ersten Stufe ordnet das Verfahren Datenbanken wie beispielsweise den RÖMPP pauschal einem der sechs Schwerpunktfächer der TIB (Architektur, Chemie, Informatik, Mathematik, Physik, Technik) zu. In der zweiten Stufe werden alle in den Datensätzen vorhandenen Klassifikationsangaben (z.B. DDC, MSC u.a.) genutzt, um eine weitere fachliche Zuordnung zu ermöglichen. In der dritten Stufe werden ISSN- und Konferenzangaben für die weitere Zuordnung herangezogen. Wenn bis zu diesem Zeitpunkt keine automatische Verarbeitung möglich ist, übergibt das Verfahren die betroffenen Datensätze an eine Plattform zur Klassifizierung (averbis extraction platform). Die ersten drei Stufen sind Eigenentwicklungen der TIB, die eigens für diesen Zweck erzeugte lexikalische Ressourcen nutzen und im Rahmen eines vom Bundeswirtschaftsministerium (BMWi) geförderten Projektes entstanden sind. Die vierte Stufe basiert auf Methoden des automatischen Lernens und wurde gemeinsam mit der Firma averbis aufgebaut.

Ein Blog will zu 500 freien Thesauri in 47 Sprachen verlinken

Ein Thesaurusportal, das ein Blog ist, präsentierte Andreas Ledl von der Universitätsbibliothek Basel. Er hat die interessante Idee, Web 2.0-Technologie dafür einzusetzen, das Problem der leidigen im Web verstreuten Linklisten mit den URLs von frei zugänglichen Online-Thesauri, Online-Klassifikationen und Online-Ontologien zu lösen, selbst umgesetzt. „Bei kleineren Datenmengen eignen sich Blogs ideal dazu, unhandliche, statische Linklisten zu ersetzen. Man kann sie als Open Access-Datenbank benutzen und Web 2.0-Funktionalitäten einbinden", erläuterte er in Hildesheim. Die erste Auflage ist unter http://thesaurusportal.blogspot.com im Web und kann von allen Interessenten kostenlos benutzt werden. Ledl will den Blog sukzessive weiter ausbauen, so dass er „einmal über 500 frei zugängliche Thesauri, Klassifikationen und Ontologien in 47 Sprachen enthalten wird" – soweit er das heute schon abschätzen kann. Vielleicht werden es auch mehr.

Erste Erfahrungen mit semiautomatischer Ontologiegenerierung

„Es wäre schön gewesen, wenn es den Thesaurusblog 2009 schon gegeben hätte, als wir unsere Arbeit an dem Projekt NanOn begonnen haben", gab Dr. Michael Schwantner zu Beginn seines Vortrags „Text Mining für den Ontologieaufbau" eine erste schnelle Bewertung der Idee von Ledl ab. FIZ Karlsruhe hat im Projekt NanOn Theorien und Methoden zur automatischen Erstellung von Ontologien praktisch ausprobiert und untersucht, inwieweit sich Text Mining Methoden für den Aufbau einer Ontologie und auch für die automati-

Professor Magnus Pfeffer, Hochschule der Medien, Stuttgart, hat ein Verfahren entwickelt, das Sacherschließungsinformationen von Monografien zwischen Datenbanken abgleichen und fehlende Angaben automatisch ergänzen kann. Es wurde bereits auf die Verbundkataloge SWB, Hebis, B3Kat und HBZ angewandt.

Programmleitung und Mitglieder des Organisationsteams der GfKl 2012 verabschieden die Konferenzteilnehmer.

sche Annotation wissenschaftlicher Artikel eignen. Gemeinsam mit dem Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB) des KIT und dem INM Leibniz-Institut für Neue Materialien wurde drei Jahre lang eine Ontologie für die chemische Nanotechnologie aufgebaut. „Ursprünglich wollten wir die gesamte Nanotechnologie abdecken, aber eine erste frühe Erkenntnis war, dass sie insgesamt viel zu groß ist, um dafür eine Gesamtontologie in einem Projekt aufbauen zu können. Deshalb haben wir uns dann eingeschränkt", berichtete Schwantner. Er hob hervor, dass es wichtig und notwendig sei, sich von vornherein auf ein bestimmtes Fachgebiet festzulegen. Außerdem müssten gleich zu Beginn „Competency Questions" definiert werden, d.h. „es muss klar formuliert werden, welche Fragen zum Schluss mit Hilfe der zu erstellenden Ontologie beantwortet werden sollen". Die Karlsruher Anwendungsforscher legten dafür z.B. die Fragen „Which metals show surface plasmon resonance?" oder „Which surfactants are used in surface modification" fest. Die Comptetency Questions dienten als Richtschnur und für Benchmarking während des Erstellens.

Als Datengrundlage wurde ein Textkorpus von rund 47.000 Dokumenten (Patente und Volltexte) aus dem Gebiet der chemischen Nanotechnologie ausgewertet. Die Analyse ergab einhundert Millionen laufende Worte, aus denen 6.600 Terme extrahiert wurden. Davon waren 27% sehr relevant, 39% relevant und 34% nicht relevant (bewertet durch menschliche Experten). Die relevanten Begriffe wurden in die Ontologie aufgenommen und in semantische Relationen gesetzt. Dafür wurden bestehende Ontologien (rund 1.800 Klassen) genutzt, ergänzt um gedankliches Explorieren (1.200) und weitere manuelle Annotation (350). Durch Text Mining wurden dann noch einmal fast 3.900 weitere Klassen gefunden. Am Schluss hatte die Ontologie rund 7.400 Klassen. Klassen wird hier in der Bedeutung der Informatik, also zum Bezeichnen eines Begriffes, seiner Derivate (Singular, Plural u.a.) und Synonyme (Abkürzungen, Trivialnamen etc.) benutzt. Schwantner wies übrigens in seinem Vortrag auch mehrfach auf die unterschiedliche Bedeutung gleicher Ausdrücke aus dem Fachwortschatz von Informatikern und Bibliotheks- und Informationswissenschaftlern hin.

Bei der Verarbeitung der in NanOn erstellten Ontologie zeigten sich dann sowohl Probleme mit dem eingesetzten graphbasierten Rechenverfahren, das mit der Aufgabe überfordert war, als auch bei der Analyse der Relationen, die Ergebnisse zwischen 0% und 100% Bezug brachten. Bei der Lösung komplexer Fragestellungen allerdings, dargestellt durch die Competency Questions, konnten die Wissenschaftler nachweisen, dass „man mit einer Ontologie tatsächlich weiter kommt".

Das Fazit der Projektpartner: Trotz aller Unterstützung durch die Maschinen sei der intellektuelle Aufwand sehr hoch. Gleichwohl sorgten automatische Verfahren bei der Erstellung von Ontologien für größere Vollständigkeit. Die Qualität der automatischen Annotation hänge stark davon ab, wie vollständig die Ontologie bezüglich der Synonyme, der verschiedenen Begriffe und der Schreibweisen sei. Der Forschungsverbund Nanotechnologie der Leibniz-Gemeinschaft hat angekündigt, an der Ontologie weiter zu arbeiten.

RVK-Sacherschließungsinformationen auf vier Verbundkataloge übertragen

Einen Meilenstein in der sinnvollen automatisierten Nutzung und Weiterverwertung vorhandener Daten und Informationen hat Magnus Pfeffer mit einem vom ihm entwickelten Verfahren zur Übernahme von Sacherschließungsinformationen aus Verbundkatalogen gelegt. Der junge Professor ist ein Vertreter der nachwachsenden Generation bereits grundlegend „digital" ausgebildeter Informationswissenschaftler. Er hat in Kaiserslautern Informatik studiert, an der Humboldt Universität in Berlin den Masterstudiengang Library Information Science absolviert und war danach an der Universitätsbibliothek Mannheim als Fachreferent und stellvertretender Leiter der Abteilung Digitale Bibliotheksdienste tätig. Seit November 2011 ist er an der Hochschule der Medien (HdM) in Stuttgart Professor für Bibliotheks- und Informationsmanagement mit den Lehrgebieten Informationsmanagement und Spezialbibliotheken.

Das Verfahren von Pfeffer überträgt Sacherschließungsinformationen von erschlossenen Titeln (RVK/RSWK) auf nicht erschlossene Titel in anderen Datenbanken, um sukzessive vorhandene Klassifikationen gleichermaßen in alle Verbunddatenbanken zu bringen. Die Grundidee ist der Abgleich einer Kombination von Autoren/Urheberangaben und dem vollständigen Titel einer Monografie. Das Verfahren wurde zunächst auf die Datenbank des Südwestverbundes (SWB) und auf das Hessische Bibliotheks- und Informationssystem (Hebis) angewendet. „Als die Zahlen auf dem Tisch lagen, war klar: das lohnt sich", berichtete Pfeffer in Hildesheim. Vor dem Abgleich waren im SWB von 12.777.191 Monografien 3.979.796 mit SWD-Schlagwörtern und 3.235.958 mit RVK-Notationen erschlossen. Nach der Anwendung des Verfahrens (auf einen Datenabzug des Katalogs) konnten zusätzlich 636.462 Monografien mit SWD und 959.419 Monografien mit RVK Sacherschließungsinformationen ergänzt werden. Der Zuwachs beim Hebis-Bestand mit 8.844.188 Nachweisen lag bei knapp 1,1 Mio. RVK und 1,3 Mio. RSWK. Pfeffer berichtete,

Ausgezeichnete Forschungsarbeiten

Florent Domenach und Sarah Frost erhielten für ihre hervorragenden Forschungsarbeiten den Best Paper Award der GfKl 2011. Übergeben wurden die Auszeichnungen bei der Eröffnung der GfKl 2012. Die Preisträger hatten dabei die Gelegenheit, ihre ausgezeichneten Arbeiten zu präsentieren. Domenach hat an der University of Nicosia, Zypern, gemeinsam mit Ali Tayari „Implications of Axiomatic consensus properties" untersucht. Sarah Frost arbeitet mit Professor Daniel Baier, Universität Cottbus, an der Verbesserung der Nutzung von Bilddaten im Web, deren bisherige Ergebnisse sie im Paper „Performance of the Earth Mover's Distance on image clustering" vorstellte.

dass Sacherschließungsexperten der beiden Einrichtungen die Einträge in zufälligen und systematischen Stichproben überprüft und den Ergebnissen eine hohe Qualität bescheinigt haben. Sie empfahlen die Übernahme in die Produktivdatenbanken, was mittlerweile geschehen ist. Als nächstes wurde das Verfahren auf die Katalogdatenbanken des Bibliotheksverbundes Bayern (BVB B3Kat) und des Hochschulbibliothekszentrums Nordrhein-Westfalen (HBZ) angewandt. HBZ (Bestand 13.271.840) gewann fast 2,3 Mio. RVK und etwas über 1 Mio. RSWK-Erschließungen hinzu. Beim B3Kat, der fast 23 Mio. Titel nachweist, kamen rund 1,1 Mio. RVK und 1, 3 Mio. RSWK-Erschließungsinformationen hinzu.

Pfeffer kündigte an, weitere Titel aus Deutschen, Schweizerischen und Österreichischen Katalogen und Open Data aus anderen (Verbund-)Katalogen in sein Verfahren zu übernehmen. Außerdem plant er, das Verfahren in die quelloffene und dokumentierte Software „culturegraph" zu überführen, die auch die Grundlage für den gleichnamigen Dienst der DNB und des HBZ ist. Damit stünden Verfahren und Ergebnisse allen Interessierten zur Nutzung und Weiterentwicklung in einer modernen Entwicklungsumgebung zur Verfügung.
http://culturegraph.sourceforge.net/

Der Mensch wird zum Supervisor für die Maschinen

Das intelligente Zusammenwirken von Mensch und Maschine, bei dem die Fähigkeiten von Maschinen zur schnellen Bewältigung umfassender Aufgaben genutzt werden und der menschliche Experte als Supervisor oder Kontrolleur der Maschinenarbeit fungiert, kristallisierte sich im Verlauf der GfKl 2012 wie auch im LIS'2012 Workshop als derzeit bevorzugter Weg zur Lösung der aktuellen Probleme in der Wissensorganisation und Datenverwertung heraus. Mensch und Maschine arbeiten dabei auf zwei verschiedenen Wegen miteinander. 1. Die Maschine observiert und analysiert Daten auf hoch abstrakter (mathematisch/statistischer) Ebene, verarbeitet sie und liefert die Ergebnisse dem menschlichen Experten zur Bewertung und Korrektur. 2. Der menschliche Experte trainiert die Maschine, indem er Hand in Hand mit ihr arbeitet und ihr auf diese Weise Stück für Stück Wissen zur Einordnung von Fragestellungen in einen Kontext vermittelt. Beide Ansätze werden bereits in Softwaresystemen im Tagesgeschäft in der Wirtschaft und in vielen anderen Bereichen der Gesellschaft eingesetzt.

Die Mathematiker, Physiker, Informatiker, Biostatistiker, Bibliotheks- und Informationswissenschaftler, die auf GfKl durch deren besonderen interdisziplinären Ansatz zusammentreffen, arbeiten daran, die diesen Systemen zugrundeliegenden Methoden und Lösungen weiter zu entwickeln und durch effiziente Datenverwertung neue Wege für eine moderne Wissensorganisation zu finden. ❙

Vera Münch
Freie Journalistin und
PR-Beraterin/PR+Texte
vera-muench@kabelmail.de