Prof. Dr. Dr. Lars Schmidt-Thieme, Stiftung Universität Hildesheim, Tagungsleiter GfKl, Informatikforscher, hat Musik und Mathematik studiert und sich in beiden Fächern eine Promotion erworben, in Mathematik mit einem Informatik-Thema.

„An einer Stelle ist der Mensch als Entscheider immer gefordert."

Interview mit Professor Lars Schmidt-Thieme

Seit vier Jahren ist Professor Dr. Dr. Lars Schmidt-Thieme Vorstandsmitglied der Deutschen Gesellschaft für Klassifikation (GfKl). Der promovierte Mathematiker und Musikwissenschaftler, der sich heute selbst als Informatiker bezeichnet, lehrt und erforscht an der Universität Hildesheim Wirtschaftsinformatik und Maschinelles Lernen. Er hat die GfKl 2012 gemeinsam mit seiner Kollegin Professorin Dr. Myra Spiliopoulou von der Otto von Guericke Universität Magdeburg organisiert.
Im Interview mit B.I.T.online erläutert der 41jährige Wissenschaftler das Fachgebiet und die Forschungsarbeit der Datenverwertungsexperten.

Herr Professor Schmidt-Thieme, die 36. Jahreskonferenz der GfKl ist so gut wie vorüber. Was ist Ihr erster Eindruck? Welche großen Trends sehen Sie?

Lars Schmidt-Thieme Also ein Trend in diesem Jahr war bestimmt Streaming Data. Datenströme auf der einen Seite, Datenanalyse in der verteilten Infrastruktur, in verteilten Systemen auf der anderen Seite. Das kann ein Compute Cluster sein, in dem sehr große Datenmengen (Big Data) gespeichert sind; verteilt, weil man sie auf einer Maschine nicht sinnvoll halten kann, z. B. in der Cloud. Aber das können auch ad hoc-Netzwerke sein, die z.B. durch verschiedene Autos dargestellt werden, die dann Informationen über Fahrverhalten, Fahrintention u.ä. austauschen.

Welche Vortragsthemen haben Sie persönlich am meisten interessiert?

Lars Schmidt-Thieme Das ist immer schwer zu sagen, wenn man als Tagungsleiter mit der ganzen Organisation beschäftigt ist. Für mich war natürlich meine eigene Sektion Data Analysis Machinery und Knowledge Discovery die, in der die für mich wichtigsten Vorträge stattfanden. Themen, die für meine Arbeitsgruppe interessant sind und die hier auf der Konferenz stark vertreten waren, sind einerseits die Empfehlungssysteme, die Recommender-Systeme, und andererseits Themen zu Datenströmen und verteiltem Data Mining. Davon habe ich jetzt am meisten für mich persönlich mitgenommen. Aber zu den Recommender-Systemen gab es beispielsweise

auch einen sehr interessanten Vortrag im Konferenzblock „Data Analysis and Classification in Marketing“. Da wäre ich jetzt auch als Informatiker trotz des wenigen Hintergrundes hingegangen, wenn es zeitlich geklappt hätte.

Welcher Grundgedanke steht hinter der Klassifikation? Sie belegen den Begriff in der Informatik anders als die Bibliothekswissenschaften.

Lars Schmidt-Thieme Klassifikation ist ein ganz abstraktes Setting und es gibt unendlich viele andere Anwendungen. In den Bibliothekswissenschaften beispielsweise hat man schrecklich viele Bücher, Artikel oder Dokumente, die man verschlagworten oder sogar in eine Taxonomie oder in eine Hierarchie einordnen möchte. Wenn man sagt, es gibt die und die Fächer, Unterfächer oder Aspekte, dann wird das Buch zu einer komplexen Entität. Diese kann man beschreiben durch Metadaten wie Autor, Verlag, Erscheinungsdatum und die kann man beschreiben durch ihren Inhalt, den ganzen Text, der da in dem Buch drinsteht. Man kann sogar noch viel mehr irgendwie dazu gehörende Informationen finden, etwa Reviews oder eine Biografie des Autors. Also es gibt für ein Buch oder auch einen wissenschaftlichen Aufsatz eine unglaublich reichhaltige Beschreibung. Solche observierten Informationen bezeichnen wir als Sensordaten oder Prädiktoren. Wenn Sie diese haben, brauchen sie einen Menschen, einen Fachreferenten auf sehr feingranularer Ebene, der hergeht und sagt, dieses Buch kann ich an dieser Stelle der Taxonomie einsortieren, und diese fünf Schlagworte beschreiben es ziemlich gut. Das nennen wir die Aktionen, diese Zuordnung zu bestimmten Kategorien.

Wenn wir die Aktionen hinreichend oft observiert haben, also gesehen haben, wie ein menschlicher Experte das macht, dann kann man zwei Dinge versuchen: Erstens, diese Aktionen maschinell zu lernen mit dem Ziel, den Menschen zu unterstützen. Anstatt dass der Fachreferent am Anfang vor einem weißen Blatt sitzt und sich fragt, welcher Kategorie kann ich dieses Buch zuordnen, sagt ihm die Maschine: diese drei Kategorien sind die wahrscheinlichsten. Dann muss er immer noch entscheiden, okay, das macht Sinn, es kommt hier hin. Es kann auch sein, dass er sich für eine vierte Variante entscheidet.

In Zusammenhängen, in denen es sehr wichtig ist, dass Entscheidungen auf keinen Fall ungeprüft automatisch vorgenommen werden, ist es sicherlich wichtig, dass der menschliche Supervisor unterstützt wird, es aber der Mensch ist, der am Schluss handelt. Wir nennen das Decision Support, Entscheidungsunterstützung.

Bei Anwendungen, die entweder weniger wichtig oder weniger kritisch sind – oder die man besser versteht – kann man die Zuordnung aber auch voll automatisieren. Das wäre der zweite Weg. In diesem Fall ist der Mensch als menschlicher Entscheider am Anfang wichtig, wenn das System aufgebaut wird. Er überwacht und trainiert es. Irgendwann, wenn der erste Vorschlag immer stimmt, dann kann man sagen: okay jetzt soll das System ganz übernehmen und das Labelling, die Zuordnung, automatisch vornehmen.

Was ist das wichtigste Ziel der Klassifikation?

Lars Schmidt-Thieme Abstrakt formuliert ist das wichtigste Ziel, den menschlichen Experten bzw. den tatsächlichen Zusammenhang so genau wie möglich zu erfassen und durch ein statistisches, probabilistisches Modell zu modellieren. Nicht den ganzen Menschen, nur die Entscheidung des menschlichen Experten in bestimmten Situationen, wie eben beschrieben. Den ganzen Menschen zu modellieren, da sind wir noch ganz weit weg. Es geht immer um einzelne Aspekte, um ganz konkrete, definierte Aufgaben. Für die Forschung und Entwicklung ist das sogenannte Lernproblem das zugrundeliegende Problem, also zum Beispiel das Zuordnen von Texten in eine Kategorie.

Die GfKl wurde 1977 gegründet. Da sah die Datenwelt noch ganz anders aus. Welche Bedeutung hat sie im Zeitalter der Volltexte? Haben sich die Aufgabenschwerpunkte verschoben?

Lars Schmidt-Thieme Das hat sich natürlich massiv verändert. In den Anfangszeiten war die GfKl ein Kristallisationspunkt für sehr viele verschiedene Fachgebiete, Wissenschaftler, die mit dem gemeinsamen Interesse an der Datenanalyse, aber möglicherweise sehr heterogenen Anwendungen zusammen gekommen sind. Das waren Probleme, die in ihren eigenen Communities vielleicht eher randständig waren. Heute ist das so, dass das Thema Datenanalyse in all diesen wissenschaftlichen Gemeinden im Herzen angekommen ist. Ich kann für meine eigene Community natürlich am besten sprechen: Wir haben eine große, sehr aktive internationale Community im Maschinellen Lernen – wir wür-

Konferenz im Theater: Die Hauptvorträge der GfKl wurden im Theatersaal der Hildesheimer Kulturpädagogik gehalten, der vollkommen schwarze Wände hat. Tagungsleiter Professor Dr. Dr. Lars Schmidt-Thieme begrüßt die Teilnehmer.

den es jetzt nicht Datenanalyse nennen – mit einer ganz eigenen Reihe von spezialisierten Konferenzen, zu denen dann nur Informatiker kommen und mit einer Reihe von eigenen Journalen. Es gibt aber in den einzelnen Fächern sehr viel mehr Betätigungsfelder und die Rolle der GfKl, denke ich, ist in der Zeit, wo es diese starke Verankerung in den Fächern gibt, die Interdisziplinarität. Bei uns kommen eben Wissenschaftler und Praktiker aus den verschiedensten Anwendungen zusammen. Das heißt, hier haben sie einmal die Möglichkeit und können als Biostatistiker einen Wirtschaftwissenschaftler hören und sehen, dass der eigentlich ein ähnliches Problem hat auf der abstrakten Ebene. Das ist glaube ich das, was diese Konferenz so fruchtbar macht.

Es ist vollbracht: Lars Schmidt-Thieme bedankte sich bei den Teilnehmerinnen und Teilnehmern für ihr Kommen mit einer freundlichen Geste und bei seiner Mitarbeiterin Ruth Janning, M.A. (links) für die große Unterstützung bei der Organisation mit einem Blumenstrauß.

Verstehen die Wissenschaftler aus den verschiedenen Bereichen sich eigentlich?

Lars Schmidt-Thieme Ja und nein. Es ist natürlich ein Lernprozess – leider oder Gott sei Dank – es ist natürlich so, dass man nicht einhundert prozentig die gleiche Sprache hat. Die gleichen Dinge heißen in den verschiedenen Fächern unterschiedlich und man muss erst einmal herausfinden, was ich soundso nenne, das nennen die soundso. Aber manchmal gehen die Unterschiede noch sehr viel tiefer. Da haben andere Fächerkulturen ganz andere Gütekriterien.

Man merkt auch an dieser Konferenz, dass die Teilnehmer eine sehr heterogene Gruppe sind. Aus welchen Fachgebieten kommen sie?

Lars Schmidt-Thieme Traditionell ist die GfKl ja in bestimmte Areas aufgeteilt. Da gibt es die abstrakten Grundlagengebiete Statistik und Mathematik auf der methodischen Seite, und die Informatik auf der technischen Seite und darüber dann die verschiedenen Anwendungsbereiche. Die Anwendungsbereiche, wie sie derzeit vertreten sind, aber auch in der Geschichte der GfKl vertreten sind, das sind die Wirtschaftswissenschaften, das quantitative Marketing, Bank- und Finanzwesen, das sind zwei ganz wesentliche Anwendungsgebiete, und die Biostatistik, d. h. die Anwendung in der Systembiologie, in der Medizin. Und daneben gibt es dann noch so einen bunten Strauß an kleinen und kleineren Gebieten.

Welchen Beitrag leistet die Klassifikation, wie Sie sie definieren, zur Informationssuche?

Lars Schmidt-Thieme Das ist mittlerweile ein ganz, ganz wesentlicher Beitrag. Wenn man sich anschaut, wie Volltextsuche angefangen hat, dann war das zunächst einmal die Verschlagwortung. Das heißt, man kann einfach durch die Verfügbarkeit des Volltextes alle Belegstellen finden. Das ist gewissermaßen die allereinfachste Art des Zugangs zu dieser Information. Das kann man schon sehr gut realisieren. Darüber gibt es aber ja noch viele weitere Schichten. Z.B. Schichten des personalisierten Zugangs. Es gibt ein berühmtes Beispiel aus dem Maschinellen Lernen. Wenn ich nach Michael Jordan suche, dann meine ich natürlich meinen Kollegen an der Universität in Berkley, USA, und nicht den berühmten amerikanischen Basketballspieler. So etwas kann ein System natürlich lernen: „Schmidt-Thieme ist Informatiker und arbeitet an überwachten Lernaufgaben für komplexe Daten; an Klassifikation und Regression." Wenn das System das weiß, kann es auch sehr personalisiert reagieren. Daraus kann man einen unglaublichen Nutzen schlagen. Natürlich mit dem Nachteil, dass andere Funktionen verlorengehen; etwa Funktionen, die man in der Frühzeit der Suchmaschinen hatte, Suchmaschinen als Wegweiser. Das geht verloren. Und da hatte man damals in der Übergangsphase in der Fachwelt diskutiert, ob das wohl gutgeht oder ob diese Funktion wichtig ist oder nicht, ob vielleicht Personalisierung anstelle einer offenen Suche gar nicht so schlau ist. Aber ich denke, das Thema hat sich erledigt.

Ist denn die Browsing-Funktion zugunsten der gezielten Suche verlorengegangen?

Lars Schmidt-Thieme Das hängt davon ab, wie man diese Werkzeuge benutzt. Ich glaube, das sind einfach zwei verschiedene Anwendungsfälle, verschiedene Szenarien. Es ist das Szenario, dass ich sehr fokussiert suche und das andere Szenario, dass ich Zeit habe und schaue, was gibt es denn da. Aber auch wenn ich rumschauen will, was es gibt, dann will ich in der Regel ja nicht so völlig breit schauen, beispielsweise, was jetzt ein Kollege aus der Netzwerktechnik macht, was ich gar nicht verstehe. Ich habe sehr wahrscheinlich einen gewissen Fokusbereich und da kann natürlich ein personalisierter Vorschlag interessant sein.

Der „Workshop on Classification and Subject Indexing in Library and Information Science" (LIS'2012) fand im Rahmen der GfKl-Konferenz statt. Woher kommt die Motivation von Datenex-

perten, einen Schwerpunkt auf Bibliotheks- und Informationswissenschaften zu legen und warum haben Sie das Thema nicht vollständig in die Konferenz integriert?

Lars Schmidt-Thieme Es gibt natürlich einen historischen Hintergrund. Als die GfKl geplant wurde, war Klassifikation als die Zuordnung von Entitäten zu bestimmten Klassen, bestimmten Gruppen innerhalb einer Taxonomie auch für den bibliothekarischen Bereich von Interesse. Von Anfang an bildeten Bibliothekare eine Fraktion innerhalb der GfKl. Warum sind sie heute noch da? Ich glaube, dass wir viele gemeinsame Probleme haben, dass es ein interessanter Anwendungsbereich ist. Um ein paar Probleme zu benennen: automatische Verschlagwortung, personalisierte Suche, automatische Kategorisierung, automatisches Labelling, Deduplizierung – das alles kann man maschinell lernen, das ist ein unglaublich wichtiges Problem im Bibliothekswesen; das Extrahieren von ganz einfachen Informationen wie: wer wird eigentlich zitiert in einem bestimmten Artikel. Für all diese Dinge braucht man unheimlich viel Datenanalyse und Maschinelles Lernen. Daher glaube ich, dass diese Kooperation mit den Bibliothekswissenschaften eine sehr fruchtbare sein kann.

Die andere Frage, warum sind sie nicht voll integriert? Das hat natürlich auch damit zu tun, dass sie eine sehr eigene Fachkultur haben. Und in den letzten Jahren versuchen wir im Vorstand der GfKl verstärkt den LIS-Workshop wieder stärker in die Konferenz zu integrieren. Auch der Leiter der Unibibliothek in Karlsruhe, Herr Scholz, der den Workshop mit organisiert hat, setzt sich sehr dafür ein. Ich persönlich glaube auch, das wird wieder mehr zusammenwachsen.

Die Themenblöcke der GfKl 2012

In diesem Jahr waren die Konferenzvorträge zum ersten Mal sogenannten Areas zugeteilt. Die Themenblöcke wurden von Fachleuten für den jeweiligen Bereich organisiert.

1. **Statistics and Data Analysis (SDA)**
2. **Machine Learning and Knowledge Discovery (MLKD)**
3. **Data Analysis and Classification in Marketing (DACMar)**
4. **Data-Analysis in Finance**
5. **Biostatics and Bio-informatics**
6. **Interdisciplinary Domains (InterDom)**
7. **Workshop Library and Information Science (LIS'2012), Workshop**

Haben Sie in Ihrer Arbeit einen Berührungspunkt mit Bibliotheken? Bibliothekarische Forschungsthemen?

Lars Schmidt-Thieme Wir haben eine Menge in diesem Bereich selber geforscht. Insbesondere habe ich ja von 2003 bis 2005 das io-Port-Projekt in Karlsruhe für Professor Rudi Studer am Institut AIFB der damaligen Universität Karlsruhe geleitet. Da gab es mit dem Fachinformationszentrum FIZ Karlsruhe eine enge Kooperation. Das Projekt ist dann ausgelaufen und im Moment verfolgen wir das Thema nicht intensiv. Aber die Methoden, die wir machen, sind in diesem Bereich sehr gut einsetzbar. Also mit dem entsprechenden Anwendungspartner wären wir da sofort wieder aktiv.

Das Wissen, das ich in meiner Gruppe aufbaue, sind Grundlagen. Wir verstehen uns als Querschnittgruppe und bauen in erster Linie methodisches Wissen auf und dann immer nur temporär ein gewisses Fachwissen, das wir für die Projekte brauchen, um mit den Anwendern reden zu können und ihre Anliegen zu verstehen. Wir haben grundsätzlich immer auch Disziplinfachleute in unseren Projekten. Wir machen das nie allein.

Herr Professor Schmidt-Thieme, wir danken Ihnen für das Gespräch.