Die Forschungsgruppe Datenbanken und Informationssysteme an der Universität Innsbruck

Günther Specht¹

Online publiziert: 9. Februar 2018

© Der/die Autor(en) 2018. Dieser Artikel ist eine Open-Access-Publikation.

Zusammenfassung

In diesem Beitrag wird die Forschungsgruppe Datenbanken und Informationssysteme der Universität Innsbruck vorgestellt. Dabei wird ein Überblick über das Umfeld, die aktuellen Forschungsthemen und die Lehraktivitäten gegeben.

1 Einleitung

1.1 Die Forschungsgruppe DBIS

Die Forschungsgruppe (Lehrstuhl) Datenbanken und Informationssysteme (DBIS) an der Universität Innsbruck wurde 2006 von Günther Specht gegründet und wird seither von ihm geleitet. Sie besteht derzeit aus elf Mitarbeitern und drei assoziierten Mitarbeitern an der Sektion für Genetische Epidemiologie der Medizinischen Universität Innsbruck. Allein 2017 wurden vier Dissertationen eingereicht.

Unser Schwerpunkt liegt derzeit auf fünf Forschungsrichtungen:

- Hocheffiziente Hauptspeicherindexstrukturen
- Information Retrieval und Recommendersysteme
- Intrinsische Plagiatserkennung und Text Mining
- Isochrone und Geografische Informationssysteme
- Big Data in der Genetik (Bioinformatik)

Zusätzlich entwickeln wir noch eine Reihe viel genutzter Systeme, wie den BIM-Merkmalserver für die Bauindustrie, den Haplogrep-Server zur Bestimmung der Haplogruppen in der Genetik, den CloudGene-Server (Cloud Service für die Genetik, basierend auf Hadoop und MapReduce), einige Systeme für die Digital Humanities, sowie verschiedene Lern-Tools für Studierende, wie z.B. Relax¹, mit dem die relationale Algebra auf beliebigen Datensets interaktiv aus-

föhrbar ist und somit leichter verständlicher gemacht wird. Viele Systeme sind über unsere Homepage² zugreifbar.

Da das Umfeld, in dem wir arbeiten, nicht ganz so bekannt ist, möchten wir vorab kurz die Universität und das Institut für Informatik vorstellen.

1.2 Die Universität Innsbruck

Die Universität Innsbruck ist eine Volluniversität und die größte Universität Westösterreichs. Sie besteht aus 16 Fakultäten, das Spektrum reicht alphabetisch von Architektur bis Volkswirtschaft und chronologisch von der Philosophischen Fakultät (1669) bis zur School of Education (2012). Sie hat 28.300 Studierende, 126 Studienfächer, 4.600 MitarbeiterInnen, davon 3.200 wissenschaftliche MitarbeiterInnen. Unser Einzugsbereich umfasst Tirol, Vorarlberg, Südtirol (Italien), Liechtenstein und den Süden von Bayern. In den internationalen Rankings ist die Universität Innsbruck sowohl im Times Higher Education Ranking gelistet als auch im Shanghai Ranking. Dort steht sie 2016 in der Gruppe der 151–200 besten Universitäten weltweit und damit – gemeinsam mit der Universität Wien – an der Spitze der österreichischen Hochschullandschaft.

1.3 Das Institut für Informatik

Das Institut für Informatik wurde 2001 gegründet und besteht inzwischen aus 8 Forschungsgruppen (Lehrstühlen) und 11 ProfessorInnen (einschließlich assoziierter Professoren). Der Schwerpunkt des Instituts liegt auf der Praktischen und Angewandten Informatik einerseits und Computer Vision und der Robotik andererseits. Im Bachelorstudium haben wir jährlich ca. 130–160 Anfänger, insgesamt etwas über 850 Studierende und ca. 50 DoktorandInnen.

¹ <http://dbis-uibk.github.io/relax/calc.htm>.

✉ Günther Specht
guenther.specht@uibk.ac.at

¹ Institut für Informatik, FG Datenbanken und Informationssysteme, Universität Innsbruck, Technikerstr. 21a, 6020 Innsbruck, Österreich

² <https://dbis-informatik.uibk.ac.at/>.

Aktuell wurde unser Informatikstudium im Times Higher Education Computer Science Ranking 2017 in den Bereich 151–175 weltweit gelistet. Im U-Multirank Student Survey 2015, wurde das Institut mit A (dem höchsten Rang) in 13 Gebieten, einschließlich Forschung, Lehre und Internationalität bewertet. Insgesamt arbeiten am Institut 130 Mitarbeiter, die Hälfte davon auf Drittmittelstellen. Dazu wurden in den letzten fünf Jahren am Institut 34 Millionen Euro Drittmittel akquiriert. Das Institut für Informatik ist Teil der Fakultät für Mathematik, Informatik und Physik, der größten Fakultät der Universität.

2 Aktuelle Forschungsthemen

Im Folgenden präsentieren wir eine Übersicht über unsere Forschungsschwerpunkte.

2.1 Hocheffiziente Hauptspeicherindexstrukturen

Im Kerndatenbankbereich beschäftigt sich DBIS vor allem mit der Datenorganisation im Hauptspeicher. Im Speziellen wird in diesem Bereich an neuartigen Indexstrukturen für Hauptspeicherdatenbanken und der optimierten Speicherung von Graphdaten geforscht [1]. Im Bereich von Graphdatenbanken wurde der Ansatz eines hauptspeicherorientierten Datenbanksystems entwickelt, das Graphdaten nativ mittels intelligenter Zeigerstrukturen abbildet und zur Anfrageverarbeitung tiefensuchbasierte Kantentraversierungen verwendet. Ebenfalls wurde ein Algorithmus entwickelt, der Graphdaten in dreidimensionalen Speichertopologien für zukünftige hochparallele Rechner optimal verteilt [10]. Ein weiterer Schwerpunkt im Kerndatenbankbereich liegt auf der optimierten Indizierung im Umfeld von Hauptspeicherdatenbanken. In diesem Bereich wurde mit dem DCB-Tree (Delta Coded B-Tree) eine speichereffiziente Indizierung von kurzen Schlüsseln entwickelt [2]. Aktuell wird an einer trie-basierten Indexstruktur mit Namen HOT (Height Optimizing Trie) geforscht, die sowohl speichereffizienter als auch performanter als vergleichbare Alternativen ist. Im Gegensatz zu bekannten Trie-Strukturen ist der dabei entwickelte Ansatz robust gegenüber ungleich verteilten, langen Schlüsseln.

2.2 Music Information Retrieval

Ein wichtiger Punkt unserer Forschung umfasst den Bereich des Music Information Retrieval und der Recommender Systeme. Dabei liegt der Schwerpunkt auf Context-aware Recommender Systems – Empfehlungssysteme, die entsprechend dem Kontext des Benutzers personalisierte und angepasste Empfehlungen berechnen, da der wahrgenommene Nutzen einer Empfehlung stark vom Kontext ab-

hängt. Im Fall von Musikempfehlungen bedeutet Context Awareness, dass ein Benutzer im richtigen Moment (der richtigen Situation) das richtige Musikstück vorgeschlagen bekommt. Wir befassen uns daher mit zwei Bereichen: (i) Feature Extraktion und Engineering, um verbesserte Features zur Charakterisierung von Benutzern, Musikstücken und Kontext zu entwickeln und (ii) Recommender Systeme, die insbesondere das kontextspezifische Verhalten (und die entsprechenden Features) von Benutzern bestmöglich für die Personalisierung von Empfehlungen auszunutzen. Im Bereich des User Modelings arbeitet DBIS daran, verschiedene Arten von Kontext-Features zu extrahieren, um eine möglichst vollständige Charakterisierung des Benutzers und dessen Kontext zu erhalten [11]. Diese Features reichen von der aktuellen Situation des Benutzers (z. B. Fitnessstudio, Arbeit, Erholung) über die aktuelle Gemütslage des Benutzers bis hin zu kulturellen Aspekten. Als Datengrundlage dafür dienen Daten aus dem Social Web, wie sogenannte #nowplaying Tweets, Daten von Musikempfehlungsdiensten wie last.fm, aber auch Daten, die direkt von Plattformen für Music Streaming wie Spotify gecrawlt wurden. Daraus entstanden zwei Datensets: das #nowplaying-Datenset [25] und das Playlist-Datenset [11], die DBIS frei zur Verfügung stellt³. Diese neu entwickelten Features fließen dann in Empfehlungssysteme ein, die mit State-of-the-Art-Algorithmen versuchen, den Benutzer, dessen Kontext und sein kontext-spezifisches Musikverhalten zu modellieren [12, 13]. So bilden die bei DBIS entwickelten User-Modelle unter anderem ab, welche User entsprechend ihres kulturellen Kontextes und ihrer aktuellen Situation bestimmte Arten von Musik präferieren. Dabei legen wir großen Wert auf die Interpretierbarkeit der Modelle und forschen daher an Methoden basierend auf Regression, Matrix-Faktorisierung und Baum-basierter Klassifizierung.

2.3 Semistrukturierte Informationssysteme

Ein weiteres großes Standbein ist die Forschung im Bereich von semistrukturierten Informationssystemen. In kollaborativen Plattformen werden oftmals Tags (oder Key/Value-Paare) eingesetzt, um Ressourcen zu annotieren und zu kategorisieren. Diese Annotationen sind von den Benutzern frei wählbar, was dazu führt, dass sich die verwendeten Annotationen stark unterscheiden (z. B. durch die Verwendung von Synonymen oder durch die unterschiedliche Strukturierung der Informationen). Diese Heterogenität schränkt die Möglichkeit, diese Daten effizient abzufragen, stark ein. Daher verfolgt der von DBIS entwickelte Snoopy-Ansatz [4, 5] das Ziel, Empfehlungssysteme einzusetzen, um passende Elemente (Keys und Values) vorzuschlagen. So kann proaktiv der Heterogenität von Informationen in

³ <http://dbis-nowplaying.uibk.ac.at>.

semistrukturierten Informationssystemen entgegengewirkt werden, indem Benutzer direkt bei der Eingabe von neuen Informationen unterstützt werden. Dies führt einerseits zu homogenen Informationen und damit gesteigerter Abfrage-Performance und andererseits dazu, dass Benutzer mehr Informationen eingeben, da sich durch die Verfügbarkeit von Empfehlungen die Eingabe einfacher gestaltet [26].

2.4 Intrinsische Plagiatserkennung und Text Mining

Ein weiterer Fokus der DBIS-Gruppe in Innsbruck liegt auf der Analyse von Textdokumenten. Dabei widmet sich ein Teilbereich der intrinsischen Plagiatserkennung, in dem versucht wird, anhand von Stilbrüchen innerhalb eines Textdokuments mögliche Plagiate zu identifizieren. Hierfür werden die einzelnen Sätze des zu untersuchenden Dokuments in eine Grammatikbaum-Struktur übergeführt, deren Ähnlichkeit anschließend mit Hilfe von maschinellem Lernen gemessen werden kann. Werden zu große Unterschiede in der verwendeten Grammatik entdeckt, spricht das für einen möglichen Fall von Plagiarismus. Wir konnten zeigen, dass sich die von den verschiedenen Autoren verwendete Grammatik signifikant unterscheidet und sich die Analyse der Grammatik dadurch gut für diesen Zweck eignet [18, 19].

Ein weiteres Forschungsgebiet stellt die Autorenerkennung dar, in dem ein unbekanntes Dokument einem bestimmten Autor zugewiesen werden soll. Auch hier konnten durch die Analyse von verschiedenen grammatikalischen Merkmalen erhebliche Verbesserungen erzielt werden [20]. Die automatische Profilerstellung ist ein technisch verwandtes Themengebiet, bei dem versucht wird, Metainformationen wie etwa das Alter oder Geschlecht aufgrund des Schreibstils vorausszusagen [21]. Schließlich wird auch im Bereich der Multilingualität geforscht, in dem Algorithmen entwickelt wurden, um maschinell übersetzte Texte zu identifizieren und ihrer ursprünglichen Sprache mit hoher Genauigkeit zuzuordnen. Die Anwendungsgebiete der verschiedenen Arten der Textanalysen sind vielfältig: so kann etwa nach möglichen Plagiaten in Dokumenten gesucht werden, deren Quellen nicht digital verfügbar sind (z. B. ältere Bücher) und somit von Systemen nicht erkannt werden können, die auf externe Referenzdatenbanken zugreifen. Anwendungen der Autorennzuordnung von Texten reichen vom Erkennen von Ghostwriting über digitale Forensik bis hin zu Tantiemenstreitigkeiten bei Mehrautorenwerken. Seit 2015 ist DBIS im Organisationsteam des international angesehenen PAN-Workshops vertreten, der jährliche Tasks rund um das Thema Textanalyse veranstaltet [17, 22].

2.5 Isochrone und Geoinformationssysteme

In Geoinformationssystemen bezeichnen die Flächen innerhalb einer Isochrone die Gebiete, die von einem Ausgangspunkt aus innerhalb einer gegebenen Zeit erreichbar sind. Diese müssen nicht notwendigerweise zusammenhängend sein (siehe z. B. Zug-Haltestellen) und können Löcher enthalten (z. B. Berge). Mit Hilfe der entwickelten Algorithmen lassen sich viele interessante Fragestellungen beantworten, darunter z. B. wo Schulen oder Rettungstellen platziert werden müssen, um innerhalb einer gewissen Zeit von allen Punkten aus erreichbar zu sein. Wenn man die Gebiete betrachtet, von wo aus ein bestimmter Punkt innerhalb einer gegebenen Zeit erreichbar ist, spricht man von einer Incoming Isochrone, sonst von einer Outgoing Isochrone. Unser Ziel ist es, multimodale Isochrone, d. h. unter Benutzung beliebig kombinierbarer Verkehrsmittel, für große Netze effizient zu berechnen. Dabei werden sowohl kontinuierliche Verkehrsmittel (wie zu Fuß oder Auto) als auch diskrete (wie öffentliche Verkehrsmittel, die fahrplangebunden sind) berücksichtigt. Wir beziehen dabei auch die Höheninformation (z. B. wichtig beim Fahrrad) mit ein. Diese Forschung erfolgt in enger Zusammenarbeit mit den Datenbankgruppen der Freien Universität Bozen (Johann Gamper). Generell erweist sich der Dijkstra-Algorithmus zur Berechnung von Isochronen schnell als ineffizient. Unsere Ansätze mit den MineTX- und MineRX-Algorithmen [7] verwenden inkrementell adaptives sowie dynamisches Nachladen und Freigeben von Knoten. Damit werden auch große Netze wie Berlin, Washington DC oder die ganze Schweiz in IsoMAPs berechenbar. Auf unserer Webseite findet sich eine Web-Applikation⁴, mit deren Hilfe das Prinzip veranschaulicht und Isochrone mit verschiedenen Algorithmen berechnet werden können. Zur Erstellung geeigneter Datensets müssen Informationen aus unterschiedlichsten Quellen integriert sowie deren Datenqualität ermittelt und optimiert werden [15].

2.6 Der freeBIM-Merkmalserver

Unter dem Begriff BIM (Building Information Modeling) wird im Bauwesen nicht nur ein digitales Gebäudemodell verstanden, sondern alle im Zuge eines Bauvorhabens anfallenden Daten über den kompletten Lebenszyklus hinweg – von der Projektidee über die Planung, die Ausführung, den Betrieb bis hin zur Verwertung der Rohstoffe beim Abbruch. All diese Daten sollen normiert und integriert in einem Datenmodell gehalten werden, d. h. in einer Datenbank. Dabei sprechen wir bei einem Lebenszyklus von weit über 100 Jahren. Zur Zeit wird intensiv an der Normierung und Internationalisierung der dabei zur Verwendung

⁴ <https://dbis-isochrone.uibk.ac.at>.

kommenden Merkmale mit deren Bemessungen, Einheiten, Verantwortlichkeiten und Projektphasenabhängigkeiten gearbeitet. Im Rahmen zweier Kooperationsprojekte mit Tiroler Unternehmen und dem Arbeitsbereich für Baubetrieb, Bauwirtschaft und Baumanagement der Universität Innsbruck wurde von DBIS der erste öffentliche Merkmalserver (BIM Property Server) entwickelt [16], der inzwischen als ÖNORM A 6241-2 Teil des ÖNORM-Standards ist (das entspricht dem DIN-Standard). Die Daten werden durch das Normungsgremium des Austrian Standards Institute (ASI) (Pendant zur DIN) eingepflegt, online diskutiert, international abgeglichen (um Duplikate zu vermeiden), mit einem international eindeutigen „Global Unique Identifier“ (GUID) versehen, und als Norm für Österreich freigegeben. Intern kommt für den ganzen Prozess und für die Normdatenbank eine Graphdatenbank zum Einsatz, die sich durch ihre flexible Erweiterbarkeit in der heterogen zusammengesetzten Kooperation aus Architekten, Ingenieuren, und Informatikern bestens bewährt hat. Derzeit arbeiten wir an einer Internationalisierung, da auch andere Länder Interesse gezeigt haben, den BIM-Merkmalserver zu übernehmen.

2.7 Big Data in der Genetik

Schwerpunkt der bioinformatischen Forschungsgruppe an der Genetischen Epidemiologie sind die Methodenentwicklung, die Systementwicklung und die Bereitstellung von Services für genetische Anwendungen. Das hierfür entwickelte Framework Cloudgene⁵ wurde 2012 publiziert [14] und stellt die Grundlage für mehrere Systeme und Dienste dar. Cloudgene setzt dabei auf dem MapReduce-Programmiersmodell in Apache Hadoop auf.

Als eines der darauf aufbauenden Systeme bietet der Michigan Imputation Server⁶ einen kostenlosen Genotype Imputation Service. Hierbei werden unvollständige DNA-Informationen mit statistischen Ansätzen, basierend auf einem großen Referenzdatensatz, vervollständigt. Dieses System wurde in Zusammenarbeit mit der University of Michigan und dem Eurac Research Center (in Zusammenarbeit mit Goncalo Abecasis und Christian Fuchsberger) entwickelt. Mit bereits über 13 Millionen analysierten Genomen und über 2.500 aktiven Nutzern ist es eines der größten Cloud-Dienste für genetische Daten. Der zugrundeliegende Ansatz und die Methoden wurden in drei Nature Genetic Papers veröffentlicht [3, 8, 9].

Ein weiterer Service basierend auf Cloudgene ist der mtDNA-Server⁷ [23], ein kostenloser Dienst für die Analyse menschlicher mitochondrialer DNA-Daten. Hier wird die zuverlässige Identifizierung von geringsten Mutationen

mit hoher Genauigkeit erreicht. Zudem ermöglicht dieser Ansatz das Detektieren von Artefakten und Kontamination in Next-Generation-Sequenzierungsdaten. Dabei werden in massiven parallelen Sequenzierungsansätzen Daten im Hochdurchsatz generiert, wobei die Datenauswertung zum Bottleneck wird. Durch den skalierbaren MapReduce-Ansatz wird dem Rechnung getragen.

Die bereits erwähnten Mitochondrien, die in den Körperzellen für die Energieprozesse mitverantwortlich sind, besitzen eine eigene DNA (mitochondriale DNA, kurz mtDNA), die im Menschen ausschließlich mütterlicherseits vererbt wird. Dadurch ist es möglich, die Verbreitung des Menschen über tausende von Jahren zurückzuverfolgen. Dies wird durch die Klassifizierung von mtDNA-Sequenzen in sogenannte Haplogruppen ermöglicht. Diese Klassifizierung ist vor allem für evolutionäre, forensische und medizinische Genetik wichtig. Mit dem Tool HaploGrep⁸ [6] haben wir ein System entwickelt, das von vielen Gruppen weltweit eingesetzt wird. Dieses Tool wurde ebenso wie der bereits erwähnte mtDNA-Server in einem renommierten Journal⁹ veröffentlicht [24]. Die Services sind frei verfügbar und führten bereits zu zahlreichen Kooperationen weltweit.

3 Lehre

In der Lehre ist die Forschungsgruppe DBIS für die Datenbank- und Information-Retrieval-Ausbildung an der Universität Innsbruck verantwortlich. Diese wird für den Bachelor und Masterstudiengang Informatik, den Masterstudiengang Wirtschaftsinformatik und den Lehramtsstudiengängen angeboten. Im Bachelorstudium bieten wir neben der Grundvorlesung Datenbanksysteme und der Pflichtvorlesung Programmiermethodik in der Vertiefung die Lehrveranstaltung „Architektur und Implementation von Datenbanksystemen“ an, die abwechselnd als Inverted Classroom oder als Praktikum angeboten wird. Im Praktikum programmieren die Studierenden im Laufe eines Semesters den Kern eines Datenbanksystems selbst nach, um dann in Performanzwettbewerben gegeneinander antreten.

Alle zwei Jahre bieten wir ein sehr nachgefragtes Bachelorseminar zur Geschichte der Informatik an, das wir im Rahmen einer mehrtägigen Exkursion durchführen. Eine solche ging z. B. zum Heinz-Nixdorf-Museumsforum nach Paderborn, dem weltgrößten Computermuseum, weitere in die Informatikabteilung des Deutschen Museums in München. Unsere Erfahrung zeigt, dass es für die Studierenden etwas Besonderes ist, ihre Vorträge direkt vor den Originalen zu halten (ohne PowerPoint).

⁵ <http://cloudgene.uibk.ac.at/>.

⁶ <https://imputationserver.sph.umich.edu>.

⁷ <https://mtdna-server.uibk.ac.at/start.html>.

⁸ <http://haplogrep.uibk.ac.at/>.

⁹ Nucleic Acids Research, impact factor 10,1.

Auch die Masterseminare wurden neu konzipiert. In ihnen führen wir die Studierenden an die Arbeitsweise in wissenschaftlichen Konferenzen heran. So entsteht nach regelmäßigen Treffen und einem zweistufigen Review-Verfahren für die schriftliche Ausarbeitung, ein Peer-Review durch die Studierenden selbst und ein zweites durch uns, ein eigener Seminarband in Buchform, der zum Blockseminar in Konferenzform gedruckt erscheint. Inzwischen entstand eine auf 24 Bände angewachsene Seminarbandreihe.

Im Masterstudium, das in Innsbruck gänzlich in Englisch gehalten wird, bieten wir Vorlesungen mit Übungen an zu New Database Models, Information Retrieval und Data Warehouse Systems. Letzteres ist ein gemeinsam mit der Fakultät für Betriebswirtschaft neu konzipiertes Modul, in dem die Studierenden beider Fachrichtungen beide Sichtweisen kennenlernen, sowohl die informatische Sicht als Data-Warehouse-Systeme als auch die BWL-Sicht als Business Intelligence.

Inzwischen wurden bei DBIS 152 Master- und Bachelorarbeiten und 11 Dissertationen abgeschlossen. Dabei ist eine enge Betreuung und ein stetiger Dialog mit den Studierenden für uns wichtig und selbstverständlich.

Danksagung Ich möchte allen Mitarbeitern und Mitarbeiterinnen danken, die geholfen haben, DBIS aufzubauen und täglich mit großem Engagement daran weiterarbeiten: den Alumnis Michael Borovicka, Michael Felderer, Wolfgang Gassler, Günter Hackl, Christina Herzog, Gabriella Hirsch, Peter Kerschbaummayr, Michael Opitz, Dominic Pacher, Wolfgang Pausch, sowie den jetzigen Mitarbeitern Robert Binna, Rainer Breuss, Nikolaus Krismer, Benjamin Murauer, Martin Pichl, Doris Silbernagl, Sylvia Thaler, Michael Tschuggnall, Matthias Weiler, Eva Zangerle sowie Lukas Forer, Sebastian Schönherr und Hansi Weissensteiner von der Genetischen Epidemiologie.

Funding Open access funding provided by University of Innsbruck and Medical University of Innsbruck.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (<http://creativecommons.org/licenses/by/4.0/deed.de>) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Literatur

- Binna R, Gassler W, Zangerle E, Pacher D, Specht G (2010) Spiderstore: exploiting main memory for efficient RDF graph representation and fast querying. Proc. of the 1st Int. Workshop on Semantic Data Management (SemData) at the 36th Int. Conf. on Very Large Data Bases (VLDB 2010). Bd. 637. CEUR-WS, Singapore, S 1–6
- Binna R, Pacher D, Meindl T, Specht G (2015) The DCB-tree: a space-efficient delta coded cache conscious B-tree. In: In memory data management and analysis: first and second int. Workshops IMDM 2013, Riva del Garda, 26.8.2013 Springer, Cham, S 126–138 https://doi.org/10.1007/978-3-319-13960-9_10 (IMDM 2014, Hongzhou, China, September 1, 2014, Revised Selected Papers)
- Das S, Forer L, Schönherr S et al (2016) Next-generation genotype imputation service and methods. Nat Genet 48(10):1284–1287
- Gassler W, Zangerle E, Specht G (2011) The snoopy concept: fighting heterogeneity in Semistructured and collaborative information systems by using recommendations. In: Int. Conf. on collaboration technologies and systems (CTS 2011), S 61–68 <https://doi.org/10.1109/CTS.2011.5928666>
- Gassler W, Zangerle E, Specht G (2014) Guided curation of semistructured data in collaboratively-built knowledge bases. J Future Gener Comput Syst 31:111–119. <https://doi.org/10.1016/j.future.2013.05.008>
- Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2011) Haplogrep: a fast and reliable algorithm for automatic classification of mitochondrial dna haplogroups. Hum Mutat 32(1):25–32. <https://doi.org/10.1002/humu.21382>
- Krismer N, Silbernagl D, Specht G, Gamper J (2017) Computing isochrones in multimodal spatial networks using tile regions. Proc. of the 29th Int. Conf. on Scientific and Statistical Database Management (SSDM 2017), ACM, S 1–6
- Loh P, Danecek P, Palamara P et al (2016) Reference-based phasing using the haplotype reference consortium panel. Nat Genet 48(11):1443–1448
- McCarthy S, Das S, Kretschmar W et al (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48(10):1279–1283
- Pacher D, Binna R, Specht G (2014) Graph stores based on spatial computers. Proc. of the 7th Spatial Computing Workshop (SCW 2014), S 1–6
- Pichl M, Zangerle E, Specht G (2015) Towards a context-aware music recommendation approach: what is hidden in the playlist name? 15th IEEE Int. Conf. on Data Mining Workshops (ICDM 2015), IEEE, S 1360–1365
- Pichl M, Zangerle E, Specht G (2017a) Improving context-aware music recommender systems: beyond the pre-filtering approach. Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR 2017), ACM, S 201–208 <https://doi.org/10.1145/3078971.3078980>
- Pichl M, Zangerle E, Specht G, Schedl M (2017b) Mining culture-specific music listening behavior from social media data. In: IEEE Int. Symposium on Multimedia (ISM 2017) Taichung. IEEE Computer Society, S 208–215 <https://doi.org/10.1109/ISM.2017.35>
- Schönherr S, Forer L, Weissensteiner H, Kronenberg F, Specht G, Kloss-Brandstätter A (2012) Cloudgene: a graphical execution platform for mapreduce programs on private and public clouds. BMC Bioinformatics 13(1):200. <https://doi.org/10.1186/1471-2105-13-200>
- Silbernagl D, Krismer N, Augsten N, Specht G (2017) Recommending osm tags to improve metadata quality. In: Proc. of the 1st ACM SIGSPATIAL Workshop on Recommendations for Location-based Services and Social Networks (LocalRec 2017). ACM 6:1–6. <https://doi.org/10.1145/3148150.3148159>
- Specht G, Breuss R (2018) Der freeBIM Merkmalsserver. Beton Kal 1:363–366
- Stamatatos E, Tschuggnall M, Verhoeven B, Daelemans W, Specht G, Stein B, Potthast M (2016) Clustering by authorship within and across documents. In: Working notes papers of the CLEF 2016 evaluation labs. CEUR workshop proc, Bd. 1609, S 691–715
- Tschuggnall M, Specht G (2013a) Detecting plagiarism in text documents through grammar-analysis of authors. In: Proc. of the 15th Fachtagung des GI-Fachbereichs Datenbanksysteme für Business, Technologie und Web (BTW 2013), GI, LNI, S 241–259
- Tschuggnall M, Specht G (2013b) Using grammar-profiles to intrinsically expose plagiarism in text documents. In: Proc. of the 18th Int. Conf. on Applications of Natural Language to Informa-

- tion Systems (NLDB '13). LNCS 7934. Springer, Berlin, Heidelberg, S 297–302
20. Tschuggnall M, Specht G (2014a) Enhancing authorship attribution by utilizing syntax tree profiles. In: Proc. of the 14th Conf. of the European Chapter of the Association for Computational Linguistics (EACL '14), Bd. 2. Association for Computational Linguistics, S 195–199
 21. Tschuggnall M, Specht G (2014b) What grammar tells about gender and age of authors. In: Proc. of the 4th Int. Conf. on Advances in Information Mining and Management (IMMM '14), S 30–35
 22. Tschuggnall M, Stamatatos E, Verhoeven B, Daelemans W, Specht G, Stein B, Potthast M (2017) Overview of the author identification task at PAN-2017: style breach detection and author clustering. In: Working notes papers of the CLEF 2017 evaluation labs, CEUR workshop proc, Bd. 1866
 23. Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A, Specht G, Kronenberg F, Schönherr S (2016a) mtdna-server: next-generation sequencing data analysis of human mitochondrial dna in the cloud. *Nucleic Acids Res* 44(W1):W64–W69. <https://doi.org/10.1093/nar/gkw247>
 24. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schönherr S (2016b) Haplogrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res* 44(W1):W58–W63. <https://doi.org/10.1093/nar/gkw233>
 25. Zangerle E, Pichl M, Gassler W, Specht G (2014) #nowplaying music dataset: extracting listening behavior from twitter. In: Proc. of the 1st ACM Int. Workshop on Internet-Scale Multimedia Management (ISMM 2014), ACM, S 21–26
 26. Zangerle E, Gassler W, Pichl M, Steinhäuser S, Specht G (2016) An empirical evaluation of property recommender systems for wikidata and collaborative knowledge bases. In: Proc. of the 12th Int. Symposium on Open Collaboration (OpenSym 2016), ACM, Bd. 18, S 1–8 <https://doi.org/10.1145/2957792.2957804>