DGI-Praxistage 2018

Hidir Aras*

Analyse von Patentdaten mittels Workflows und Big-Data-Technologien

<https://doi.org/10.1515/iwp-2019-2001>

Zusammenfassung: In diesem Beitrag geht es um das interdisziplinäre Erlernen von Data Science u. a. im Rahmen von Aus- und Weiterbildungsmaßnahmen mittels interaktiver Lernumgebungen am Beispiel der Analyse großer Datenmengen mit Patentinformationen für neue Nutzergruppen wie z. B. den Informationsspezialisten, welche in der Regel über wenig bis keine Kenntnisse z. B. über Verfahren des maschinellen Lernens verfügen. Mittels einer interaktiven Lernumgebung auf Grundlage von Scientific Workflows und Big-Data-Technologien können dabei neue Methoden des Text und Data Mining (TDM) effizient erlernt und im Rahmen praktischer Anwendungsfälle erprobt werden.

Deskriptoren: Patente, Weiterbildung, Textanalyse, Data Mining, Wissenschaftliches Arbeiten

Analysis of patent data using workflows and big data technologies

Abstract: This article deals with the interdisciplinary learning of data science, e.g. in the context of education and training measures by means of interactive learning environments for the analysis of large amounts of data with patent information for new user groups such as information specialists, who usually have little to no knowledge of e.g. machine learning. Using an interactive learning environment based on scientific workflows and big data technologies, new methods of text and data mining (TDM) can be efficiently learned and tested in practical applications.

Descriptors: Patents, Continuing education, Text analysis, Data mining, Scientific approach

Analyse des données de brevets à l'aide de workflows et de big data technologies

Resumé: Cet article traite de l'apprentissage interdisciplinaire de la science des données, notamment dans un cadre d'enseignement et de formation continue faisant appel à des environnements d'apprentissage interactifs pour l'analyse de grandes quantités de données comportant des informations sur les brevets à destination de groupes d'utilisateurs «non-techniciens» tels que les spécialistes en information. Grâce à un environnement d'apprentissage interactif basé sur des workflows scientifiques et des technologies big data, les nouvelles méthodes de text and data mining (TDM) s'apprennent efficacement et peuvent être évaluées dans le cadre d'applications pratiques.

Descripteurs: Brevets, Formation continue, Analyse de texte, Exploration de données, Travaux scientifiques

1 Einführung

Die immensen Fortschritte in der praktischen Anwendung der Künstlichen Intelligenz (KI) z. B. zur Gesichtserkennung, in der Dialog- und Spracherkennung, für das autonome Fahren, etc. ermöglichen bereits heute ungeahnte Möglichkeiten für neue Anwendungsgebiete für maschinelle Lernverfahren. Das Erlernen und Erproben von neuen Methoden des *Machine Learning* (ML) und der KI erfordert daher sowohl aus technologischer wie auch aus wirtschaftlicher Sicht¹ die Erforschung und Bereitstellung innovativer digitaler Lernumgebungen für *Data Science*. Eine wichtige Voraussetzung dafür ist die Schaffung geeigneter Rahmenbedingungen für das interdisziplinäre und praxisnahe Erlernen dieser Methoden, um auf zukünftige Entwicklungen im nationalen und internationalen Wirtschaftsraum vorbereitet zu sein.

Gerade Fachkräfte aus der Industrie wie z. B. Informationsspezialisten, die eine enorme Kompetenz über die Da-

*Kontaktperson: Dr. Hidir Aras, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, E-Mail: hidir.aras@fiz-karlsruhe.de

¹ <https://www.bigdata.fraunhofer.de/de/big-data/kuenstliche-intelligenz-und-maschinelles-lernen/ml-studie.html> [16.3.2019].

ten ihres Fachgebiets (z. B. Life Science, Ingenieurwesen, etc.) verfügen, jedoch wenig bis gar keine Kenntnisse neuer Technologien der Datenanalyse besitzen, könnten interaktive Lernumgebungen für Data Science bei der Umorientierung und Ausweitung wichtiger Kompetenzen helfen. Jedoch auch Anfänger oder Studenten würden von solch einer interaktiven Laborumgebung profitieren.

Die Möglichkeit zur interaktiven Durchführung von Experimenten zum Erlernen komplexer Analysemethoden – die z. B. anhand des in CRISP-DM² beschriebenen standardisierten Data-Mining-Prozesses realisiert werden – ist daher für das Erlernen von Data Mining immens wichtig. Dies erfordert primär die Analyse größerer Datenkorpora mittels neuerer Verfahren des *Text und Data Mining* (TDM) zur effizienten Erstellung komplexer Modelle und deren Integration in den Data-Mining-Prozess. Oftmals erfordert dies eine interdisziplinäre Herangehensweise und Wissen aus unterschiedlichen Bereichen, welche beispielsweise im Rahmen interaktiver Data-Science-Lernumgebungen zunächst im Rahmen einer Aus- und Weiterbildung erlernt und erprobt werden müssen, um anschließend auf Realweltprobleme anwendbar und damit für die Endnutzer gewinnbringend nutzbar zu sein.

Bezogen auf den Patentbereich ist ein wichtiger Fokus die technologische Unterstützung des Such- und Rechercheprozesses in unterschiedlichen Szenarien der Patentrecherche und die Analyse großer Datenmengen mit Patentinformationen für das Ermitteln unternehmenskritischer Erkenntnisse aus der Analyse bspw. im Rahmen der Technologieanalyse, welche als Konsequenz eine Investition in eine Produktidee oder weittragende Maßnahmen nach sich zieht. Daher ist die Erforschung und Entwicklung integrierter Lösungen, welche die Erfordernisse des Informationssuchenden und dessen unternehmensinterne Workflows berücksichtigt, von großer Bedeutung.

Zur systematischen Beschreibung wissenschaftlicher Methoden z. B. zur Datenanalyse können *Scientific Workflow* Systeme wie KNIME³ eingesetzt werden. Zur Realisierung komplexer Workflows der Suche, Analyse und Interaktion mit technisch-wissenschaftlicher Information wie Patentinformation ist darüber hinaus eine neue Art und Weise der Interaktion mit den Daten unabdingbar. Neue Ansätze sollen das Erlernen und die Validierung einzelner Schritte der Analyseprozesse im Rahmen der eingesetzten Datenpipeline interaktiv gestalten, d. h. einzelne oder partielle Schritte der Analysepipeline können re-iteriert, mo-

difiziert und erneut berechnet werden. Die dafür notwendige verteilte Systemarchitektur (Abb. 1) kann dabei aus einem *Hadoop/Spark*⁴ Cluster und weiteren Komponenten zur Realisierung der Suche, der Workflow-Prozessierung unter Einbeziehung unterschiedlicher Analyseservices und dem Datenmanagement aufgebaut werden.

Abbildung 1: Big Data Architektur zur Patentanalyse.

Workflows zur Suche, Analyse und Interaktion

Das Ziel einer neuen und effizienten Art und Weise mit technisch-wissenschaftlichen Informationen wie z. B. Patentdaten zu interagieren, beinhaltet eine Ausnutzung von Synergien zwischen Methoden des Text Mining, der Informationserschließung und der Visualisierung. Nutzer sollen in die Lage versetzt werden die zu analysierende Ergebnismenge oder den fokussierten Patentkorpus entlang diverser Parameter zu reorganisieren und „on-the-fly“ damit zu interagieren. Dies erfordert neben intelligenten Methoden der Daten- und Ergebnisintegration eine erweiterte Unterstützung zur Automatisierung sich wiederholender und wichtiger Teilaufgaben auf Grundlage von Datenpipelines, aber auch die Erweiterung der Möglichkeiten für

² https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining [16.3.2019].

³ <https://www.knime.com/> [16.3.2019].

⁴ <https://hadoop.apache.org/> [16.3.2019].

Nutzeranfragen. Die Interaktion mit den Ergebnissen einer umfangreichen Analyse erfordert weiterhin sowohl komplexe Visualisierungen und Interaktionsmöglichkeiten auf der Dokumentenebene als auch eine effiziente Visualisierung und Interaktion mit großen Informationsräumen, um in einer ausgewählten Domäne gezielt zu navigieren oder einen schnellen Überblick über relevante Aspekte zu erhalten, z.B. um einen Report zu erstellen, der wichtige Fakten zu einem technologischen Bereich zusammenfasst.

Abbildung 2: Die 3 Vs von Big Data⁵, Volume, Variety, Velocity.

TDM-Verfahren zur Textanalyse

Verfahren des Text und Data Mining (TDM) gewinnen im Rahmen der Analyse großer Datenmengen (*Big Data Analytics*) zunehmend an Bedeutung, weil nicht nur die Datenmengen (*Volume*) für die Analyse eine Herausforderung darstellen, sondern auch deren Heterogenität (*Variety*) sowie die Aktualität (*Velocity*) wie in Abbildung 2 veranschaulicht wird.

Die Entwicklung neuer Methoden des TDM zur Analyse komplexer technisch-wissenschaftlicher Informationen wie in Patenten erfordert zugleich die Erforschung und Verschmelzung unterschiedlicher Technologien wie z.B. der Suche, der semantischen Analyse auf der einen und effiziente Methoden für den Zugriff, die Extraktion und die Verlinkung von relevantem Wissen auf der anderen Seite.

Beispielsweise erlaubt die automatische Strukturierung der Patentvolltexte (Ansprüche, Beschreibung) und die Extraktion relevanter Informationen wie z. B. von wichtigen Schlüsselworten (Keyterms) nicht nur einen schnel-

leren Zugriff auf die einzelnen Textsektionen der Volltexte, sondern ermöglicht ebenso eine präzisere Suche. Eine semantische Strukturierung der Beschreibung und der Ansprüche kann darüber hinaus im Rahmen weiterer Analysen wie z.B. Term Extraction, Trendanalyse, etc. ausgenutzt werden.

2 Educational TDM-Workflows zur Patentanalyse

Wie zuvor erläutert können Scientific Workflows zur Analyse großer Datenmengen mit Patentinformationen mittels modernen Workflow-Systemen wie z. B. KNIME und einer Big-Data-Plattform umgesetzt werden. KNIME bietet dabei eine Art visuelle Programmierung zur Erstellung von Workflows, welche durch sequentielle oder komplexe Formen des Zusammenschaltens (*Plug-and-Play*) einzelner Module (*Nodes*) erzeugt werden können. Neben existierenden Funktionen zur Verarbeitung von Dateien, dem Zugriff auf Datenbanken, etc. können auch eigene spezialisierte Module entwickelt und in die Workflows integriert werden.

Der Prozess zur Recherche und Analyse von Patentinformation beginnt in der Regel mit einer Suchanfrage an ein kommerzielles oder öffentlich zugängliches Patentinformationssystem wie z. B. STN⁶ oder Espacenet⁷. Ist erst einmal die über eine Suchanfrage (z. B. Suchterme, Klassifikations-Codes, Erfindernamen, etc.) ermittelte Dokumentenmenge identifiziert, können zusätzliche Schritte zur Aufbereitung, Strukturierung oder diverse weitere Analyseschritte z. B. Topic Extraction ausgeführt werden.

Da die Analyse großer und komplexer Datenmengen mit Patentinformationen hohe Anforderungen für die involvierten Algorithmen und Systeme mit sich bringt, müssen hierfür skalierbare und verteilte Analyseprozesse zur Verarbeitung der Daten z. B. in der Hadoop/Spark-Plattform (Abb. 3) ausgeführt werden.

Scalable Big Data Analytics Workflows

Skalierbare Analyseprozesse für das Text und Data Mining mit Patentdaten erfordern die Abbildung und Integration unterschiedlicher Phasen im Analyseprozess, beginnend mit der Einschränkung der Daten auf die fokus-

⁵ <https://www.computerwoche.de/a/big-data-fordern-analysesysteme-heraus> [16.3.2019].

⁶ <http://www.stn-international.de> [16.3.2019].

⁷ <https://worldwide.espacenet.com/> [16.3.2019].

Abbildung 3: Big Data Patent Analytics Workflow.

Abbildung 4: Skalierbare Analyseprozess mit KNIME und Hadoop.

sierte Domäne, über die Indexierung und die Suche, bis hin zu Vor- und Nachbearbeitung und das Ausführen von skalierbaren Analyseprozessen in KNIME-Workflows.

Dazu können eigene KNIME-Knoten zur Abfrage des Suchindex und zur Interaktion mit den Such- und Analyseergebnissen entwickelt und in Referenzworkflows evaluiert werden. Die Auswertung von Analyseergebnissen erfordert geeignete Konzepte zur Aufbereitung, Visualisierung und Interaktion. Im Rahmen der TDM-Plattform wurden dabei bestehende Komponenten in KNIME erweitert und neue Visualisierungsmöglichkeiten für die Patentanalyse, z.B. die Darstellung des Zusammenhangs von Patentdokumenten über deren Keyterms mittels Network-Graphen, Trendverläufe und Häufigkeitsanalysen über die Zeit prototypisch umgesetzt und evaluiert.

Die ersten Schritte im Workflow dienen zur Strukturierung und Anreicherung des Textkorpus der Erfindungsbeschreibungen und der Patentansprüche mit Mehrwertdaten. Diese werden zunächst semantisch segmentiert und anschließend mit Mehrwerten wie Keyterms, numerischen oder chemischen Entitäten etc. angereichert.

Abbildung 4 zeigt einen skalierbaren Workflow zur Analyse mittels KNIME und Hadoop. Dazu muss ein entwickelter Dienst z.B. zu Extraktion von Nominalphrasen, über den Workflow in Form eines verteilt ausführbaren Dienstes „deployed“ und in einem Hadoop-Cluster ausgeführt werden. Die Ergebnisse der Analyse können anschließend direkt im KNIME-Workflow visualisiert werden.

Eine weitere Anwendung für skalierbare Annotatoren stellt bspw. die Verlinkung der Patentdokumente anhand der chemischen Annotationen im Patenttext mit externen

Wissensbasen aus dem Bereich Life Science, z.B. Pub-Med⁸, ChEMBL⁹, Drugbank¹⁰ dar.

3 Fallstudien aus der Patentanalyse

Der oben beschriebene Ansatz für Big Data Analytics mit Patentinformationen wird im Folgenden anhand zweier Analysefälle näher erläutert.

3.1 Erkennen technologischer Trends

Der in Abbildung 3 gezeigte Workflow zeigt die Ermittlung der für die Analyse relevanten Dokumente mittels einer Suchanfrage und den skalierbaren Analyseprozess zur Ermittlung technologischer Trends in der Big-Data-Plattform Hadoop/Spark. Dies geschieht mit einem speziellen Analyseservice, der in der verteilten Laufzeitumgebung auf eine große Menge an Dokumenten angewendet wird. Die Ergebnisse der Analyse werden anschließend an den KNIME-Workflow zurückgegeben und nach Aufbereitung für die Nutzerinteraktion z.B. in Form eines Trendverlaufs anhand der Häufigkeit der Patentanmeldungen pro Jahr visualisiert.

Zur Analyse der technologischen Trends werden zunächst aus jedem Patentdokument wichtige Sektionen wie

⁸ <https://www.ncbi.nlm.nih.gov/pubmed/> [16.3.2019].

⁹ <https://www.ebi.ac.uk/chembl/> [16.3.2019].

¹⁰ <https://www.drugbank.ca/> [16.3.2019].

Title, Abstract, Technical Field, Summary und die Independent Claims extrahiert und zur weiteren Analyse an die Spark NLP¹¹-Pipeline zur Extraktion wichtiger Substantive weitergeleitet. Für Phrasen, die aus mehr als zwei Wörtern bestehen werden N-Gramme berechnet.

Im gezeigten Big-Data-Analytics-Workflow wurde zur Extraktion der Topics der Latent Dirichlet Allocation (LDA)-Algorithmus Spark-MLlib¹²-Bibliothek verwendet, um die wichtigsten Themen aus dem Patentkorpus zu extrahieren. Für das Clustering der Patentdokumente anhand der Ähnlichkeit wird dabei das K-Means¹³-Verfahren eingesetzt. Zusammen mit den Themen können auch für die jeweiligen Topics signifikanten Schlüsselworte (Begriffe und Termini) extrahiert und bei der Analyse verwendet werden.

Für jedes Thema werden eine Reihe von Kandidatenkonzepten (Substantivbegriffe und Phrasen) ermittelt, welche das jeweilige Thema am besten repräsentieren. Zur Extraktion der aktivsten Konzepte wird folgendermaßen vorgegangen:

- Jedes Kandidatenkonzept erhält eine Bewertung in Abhängigkeit von seiner Häufigkeit und der durchschnittlichen Auftrittswahrscheinlichkeit im Patentkorpus.
- Anschließend werden die Kandidatenkonzepte aller Themen bewertet und die Topkonzepte mit der höchsten Punktzahl zurückgegeben.

Für unsere Analyse wurde das Thema „Information Security“ ausgewählt und alle Patente zu diesem Thema abgerufen, was 26.261 Dokumente ergab, die zwischen 1978 und 2015 veröffentlicht wurden.

Abbildung 5: Trendverlauf für „Information Security“.

Abbildung 5 zeigt das Diagramm des Publikationstrends für „Information Security“ mit einem linearen Anpassungsansatz zur Abschätzung des zukünftigen Trends.

3.2 Technologieanalyse und visuelle Interaktion

Patentklassifikationen wie die IPC¹⁴ werden von Experten verwendet, um Patentdokumente nach ihren Technologiegebieten zu klassifizieren. Studien haben gezeigt, dass die Ermittlung von Merkmalen oder Trends der Technologieentwicklung unter Verwendung der IPC-Codes inkonsistent ist und eine mangelnde Abdeckung insbesondere für neu entstehende Technologiebereiche besitzt. Analysen zeigen, dass die meisten aktuellen Gebiete im Bereich der Informatik wie Deep Learning, Big Data Cloud, Blockchain und Internet of Things (IoT) in den Beschreibungen des Patentklassifikationsschemas fehlen, sie aber eindeutig im Textsegment „Technical Field“ der Beschreibung eines Patentdokuments (Description) erscheinen.

Analyse der Technologiefelder in der Beschreibung der Erfindung

Zur Ermittlung impliziter bzw. neuer Technologiebereiche wurden die Patenttexte anhand der Informationen im Segment „Technical Field“ analysiert. Da zuvor die Patentbeschreibung mittels Methoden des TDM segmentiert wurde, kann im Workflow die Erfindung (Invention) und die damit verbundenen Technologiebereiche gezielt analysiert werden.

Dazu wurden 26.388 Patente, die sich auf den Bereich „Information Security“ beziehen, aus zwei großen Patentdatenbanken mit Patenten der EPO¹⁵ und der WIPO¹⁶ aus einem Patentarchiv extrahiert. Dabei enthielten 18.381 der extrahierten Patente eine „Detailed Description“. Nach der Segmentierung der Description mit einem speziellen Verfahren zur Strukturierung der textuellen Inhalte wurden insgesamt 12.476 Patentdokumente mit dem Segment „Technical Field“ ermittelt. Anschließend wurden signifikant relevante technologische Begriffe und Phrasen automatisch extrahiert und die Ergebnisse einer statistischen Analyse unterzogen und nachfolgenden Analyseschritten im Workflow weitergereicht.

11 NLP = Natural Language Processing.

12 <https://spark.apache.org/mllib/> [16.3.2019].

13 <https://de.wikipedia.org/wiki/K-Means-Algorithmus> [16.3.2019].

14 <https://www.wipo.int/classifications/ipc/> [16.3.2019].

15 <https://www.epo.org> [16.3.2019].

16 <https://www.wipo.int> [16.3.2019].

Abbildung 6: Visualisierung der „Technology Hotspots“ für das Fachgebiet Information Security.

Die Beziehungen zwischen den auftretenden Termini können dabei einfach auf deren gemeinsames Vorkommen im Textkorpus und deren Häufigkeiten zurückgeführt werden. Dieser Umstand kann indirekt zur Extraktion semantischer Relationen z. B. Wortzusammenhänge verwendet werden.

Eine *Co-occurrence Map* diene dazu, die Technologie-Hotspots des Bereichs „Information Security“ auf einer niedrigeren Ebene zu beschreiben. Basierend auf der Häufigkeit der Termini haben wir die 200 wichtigsten Termini gesammelt und über eine entsprechende Visualisierung im Workflow dargestellt.

Abbildung 6 zeigt die wichtigsten technologischen Hotspots im Bereich der Informationssicherheit wie *digital signature*, *public key cryptography*, *private key cryptography*, *access control*, *authentication systems*, *cryptographic algorithms*, and *key management*. Es werden auch detailliertere Bereiche wie *hash function*, *DRM systems* (Digital Rights Management) und *security with smart card* sichtbar.

4 Zusammenfassung

In Patenten steckt wertvolles Wissen, welches oftmals in keiner anderen Informationsquelle publiziert wurde. Aufgrund der immensen Bedeutung für die Innovations- und Technologieanalyse sind Patentdaten und die Nutzung technologisch kritischer Information auch außerhalb der

klassischen IP (*Intellectual Property*)-Welt von besonderer Bedeutung.

In diesem Beitrag wurde gezeigt, wie neue Methoden des Text und Data Mining bei der automatischen Erschließung und Analyse relevanter Information aus großen Patentkorpora eingesetzt werden können. Neben der Erforschung und Entwicklung neuartiger Analysemethoden mit Hilfe von Big-Data-Technologien und neuen Verfahren des Machine Learning wie z.B. *Deep Learning*, stellt die semantische Erschließung und die Verknüpfung mit externen Wissensquellen eine weitere wichtige Herausforderung für die Patentanalyse dar. Beispielsweise können zusätzliche Möglichkeiten einer effizienten Suche und die Exploration großer Datenmengen auch über einen wissensbasierten Ansatz z. B. unter Einbeziehung von domänenspezifischem Wissen aus der *Linked Open Data Cloud*¹⁷ erforscht und entwickelt werden.

Das Erlernen und Erproben neuartiger Methoden erfordert jedoch geeignete Data-Science-Lernumgebungen, welche auf einer Big-Data-Architektur bspw. mittels Hadoop/Spark aufgebaut werden können. Eine flexible und skalierbare Analyse großer Datenmengen (Big Data Analytics) mittels Scientific Workflows erlaubt weiterhin die transparente Einbeziehung von Ressourcen, Abläufen und Prozessen, welche durch die in den jeweiligen Unterneh-

¹⁷ <https://lod-cloud.net/> [16.3.2019].

men eingesetzten Informationssysteme und deren Anwendungskontexte bedingt sind. So lassen sich entwickelte Dienste mit einem API-Ansatz sowohl in bestehende Infrastrukturen und Dienste integrieren, als auch im Rahmen einer interaktiven Analyse und Visualisierung großer Datenmengen in einem Workflow, der unterschiedliche Etappen des Such- und Analyseprozesses abbildet, nutzen.

Literatur

- Aras, H.; Türker, R.; Geiss, D.; Milbradt, M.; Sack, H. *Get Your Hands Dirty: Evaluating Word2Vec Models for Patent Data*, In Proc. of the 14th Int. Conf. on Semantic Systems (SEMANTICS 2018), P&D Track, CEUR workshop proceedings vol. 2198.
- Aras, H.; Hackl-Sommer, R.; Schwantner, M.; Sofean, M. *Applications and Challenges of Text Mining with Patents*. IPaMin@KONVENS (2014).
- Hackl-Sommer, R.; Schwantner, M., *Patent Claim Structure Recognition* in: Archives of Data Science, Series A, 2017, v. 2(1), 15 S.
- Jie Tang, Bo Wang, Bo Gao, Minlie Huang, Peng Xu, Weichang Li, and Adam K. Usadi., *PatentMiner: Topic-driven Patent Analysis and Mining*. KDD'12, 2012.
- Sofean, M. *Automatic Segmentation of Big Data of Patent Texts* in: Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery. DaWaK 2017. Springer pp 343–351.
- Sofean, M.; Aras, H. & Alrifai, A.; Damaševicius, R. & Vasiljeviene, G. (Eds.). *A Workflow-Based Large-Scale Patent Mining and Analytics Framework*. Information and Software Technologies, Springer International Publishing, 2018, 210–223.
- Trappey, C. V. and H.Y. Wu, F. Taghaboni-Dutta, and A. J. C. Trappey., *Using patent data for technology forecasting: China RFID patent analysis*, Advanced Engineering Informatics, 2011.
- Trippe, A., *Guidelines for Preparing Patent Landscape Reports*. Patinformatics, LLC, With contributions from WIPO Secretariat. 2015.
- Tseng, Y., Lin, C., Lin, Y., *Text mining techniques for patent analysis*, Information Processing and Management, vol. 43, no. 5, pp. 1216–1247, 2007.

Dr. Hidir Aras

FIZ Karlsruhe – Leibniz-Institut für
Informationsinfrastruktur
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
hidir.aras@fiz-karlsruhe.de

Dr. Hidir Aras ist wissenschaftlicher Mitarbeiter und Projektleiter für Text- und Data-Mining bei FIZ Karlsruhe. Im Rahmen der angewandten Forschung beschäftigt er sich unter anderem mit Big Data Analytics, Text und Data Mining und der semantischen Analyse von Patentinformationen. Hidir Aras ist seit 2012 bei FIZ Karlsruhe beschäftigt und war vorher wissenschaftlicher Mitarbeiter und Doktorand an der Universität Bremen, wo er über das Thema „Semantic Interaction in Web-based Retrieval Systems“ promovierte. Zuvor arbeitete er, nach Abschluss seines Studiums der Wirtschaftsinformatik an der Universität Mannheim, mehrere Jahre bei der European Media Laboratory GmbH in Heidelberg in verschiedenen Forschungsprojekten, u. a. zu Mobilien Assistenzsystemen oder Semantic Web.