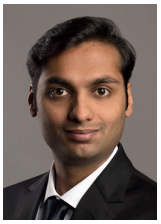
Vaibhav Kasturia

Personal Data

Place of Birth:	Delhi, India
Date of Birth:	18 November 1992
Nationality:	Indian (Permanent Resident of Germany)
Email:	vkasturia@deloitte.de, vbh18kas@gmail.com
LinkedIn:	vkasturia
Github:	vkasturia
Webpage:	www.vaibhavkasturia.com



Education

2015 - 18 Master of Science in Internet Technologies and Information Systems (ITIS)
Leibniz University Hannover, Hannover
Major: Data and Information
Research Project: "Building & Querying Semantic Layers for Web Archives"
Thesis: "Ranking Archived Documents for Structured Queries on Semantic Layers"
GPA: 1.1 (German Scale)

2011 - 15 Bachelor of Engineering (Hons.) in Computer Science
Birla Institute of Technology and Science - Pilani, Dubai
Thesis: "Software Development Practices at ESRI"
GPA: 1.1 (German Scale), 9.77 (Indian Scale)

Work Experience

May 2021 - Present — Consultant at Deloitte, Düsseldorf
Digital / AI Controls / Algorithms, Risk Advisory

Banking

- **Cloud-based Processing of Banking Documents:** Creation of a document pipeline for processing banking documents using Google DocumentAI Platform.

Life Sciences

- **Digitalization of Quality Management:** Transform Production and Quality Management system from heavily documentation-based, functionally segregated systems towards an intelligent and data-driven end-to-end solution.

Asset Management

- **ML Text Extraction of Securities Prospectuses:** Built a prototype that uses AI and rule-based solutions to recognize, extract, and display relevant data from the documents.
- **Explainable AI:** Developed a tool which uses Shapley values to explain the decision-making process of Deep Learning models like BERT, RoBERTa and DistilBERT.

Aug 2018 - Oct 2020 — Research Associate at University of Halle-Wittenberg, Halle (Saale)
Big Data Analytics, Webis Group

Query Understanding via Entity Linking

- **Goal:** Interpret ambiguous search engine queries to show more relevant results to the user, answer the query or help fill search engine's knowledge boxes.
- Designed and developed an automatic approach that uses query segmentation and entity linking to identify the most reasonable interpretations of a query based on the contained entities.
- Conducted an experimental comparison on a new corpus of 2,800 queries. It proves that my approach has better interpretation accuracy at a better run time than the previously best methods.

Total Recall in Systematic Reviews

- **Goal:** Find all relevant documents ("total recall") given a collection of potentially several thousands of documents somewhat related to a user-specified topic. A single systematic review may take up to 2 years without any machine-assistance.
- Built a system that reduces the review period by ordering these documents in descending relevance.
- Implemented several machine learning methods from an existing total recall approach (HiCAL) and tested these on botanical research datasets. The results show that machine learning reduces the human effort by almost 80 percent.

Argumentative Axiomatic Re-Ranking for Medical Search Queries

- People use search engines to seek health advice online.
- Using search engines to complete such decision making tasks, users are not able to discern authoritative from unreliable information.
- As part of a team, we developed an axiomatic approach to re-rank search results obtained by traditional search models, in order to promote more argumentative results for medical queries.

Activities

- Prepared and took exercises for undergraduate and graduate courses (Object Oriented Programming in Java, C Programming, Search Algorithms, Foundations of Computer Science and Concepts of Modelling).
- Supervised a team of 10 students in their software project internship.
  Topic: Develop a system to automatically migrate a company's old Excel-records in Excel to a database.
- Maintenance of the Big Data Analytics webpage.

May 2018 - Jul 2018 — Research Associate at Fraunhofer IAIS, Bonn

- Prototyped a Question Answering system for an accounting firm.
- As part of the team, contributed to the development of algorithms in the area of Deep Learning as well as Speech Processing for intelligent smart car systems.
- Small contributions to the project GEISER which dealt with the analysis of spatial data.

Oct 2016 - Jan 2018 — Student Assistant at L3S Research Center, Hannover

ALEXANDRIA Project

- Research on methods for the semantic and entity-based exploration of Web Archives.
- Aim of the project was to significantly advance semantic and time-based indexing for Web Archives, to efficiently index, retrieve and explore information about entities and events from the past.
- Built semantic profiles ("layers") that describe semantic information about the contents of Web Archives using Entity Linking Tools.
- Evaluated the semantic layers for complex information needs against keyword-based search systems like Google, Bing and HistDiv.
- Designed and evaluated statistical and advanced models (PageRank-like) to rank results returned by running queries on these layers.

Aug 2014 - Jan 2015 — Software Developer (Intern) at ESRI, Sharjah

- Handled Multidimensional Geo-data (GRIB, NetCDF, HDF, etc.)
- Analyzed Raster, Mosaic and Image Service Data Layers.
- Fixed bugs and changes requested for ArcGIS 10.
- Validated UI functioning of Raster and Geo-Processing Tools of ArcGIS Pro.
- Removed potential defects (by Coverity Analysis) in Raster Solutions of ArcGIS 10.

Technical Proficiency

Programming Languages:	Java, Python, C/C++
Data Science / Machine Learning:	Shap, Transformers, NumPy, pandas, scikit-learn, Matplotlib
Information Retrieval:	Apache Lucene
Database Systems:	Graph database (Virtuoso), NoSQL database (RocksDB), SQL databases (MySQL, PostgreSQL)
Natural Language Processing:	Entity Recognition, Entity Linking, Entity Disambiguation, Word Embeddings, Query Segmentation
Java Frameworks/Libraries:	Apache Maven, Apache Jena, Apache Tika, SparkJava, Standard Libraries (Apache Commons, Apache Lang, etc.)
Semantic Web Technologies:	RDF/RDFa, OWL, SPARQL, Turtle
Web Technologies:	HTML5, CSS, Materialize/Bootstrap, Flask, Streamlit
Geo-Information Systems:	ArcGIS
IDE Software:	Jupyter Notebook, IntelliJ, Eclipse, NetBeans, Visual Studio
Document Preparation:	MS Office, Apple Office Suite, LaTeX
Version-Control Software:	Gitlab/Github, GitKraken, SVN, CVS
Others:	Docker, Multithreaded Programming
Basic Knowledge:	Intel 8085 Programming, Wireshark, Scilab

Achievements and Awards

Oct 2021	Research Paper acceptance at WSDM 2022 (A* ranked conference in Computer Science)
Dec 2018	Best Master's degree certificate for 2017/18 by Leibniz University Hannover
Jun 2017	Best Research Paper Award Nomination at at JCDL 2017
2011 - 15	BITS Scholarship for Academic Excellence for the entire Bachelor's Degree

Publications

May 2021 **Query Interpretations from Entity-Linked Segmentations**
Vaibhav Kasturia, Marcel Gohsen and Matthias Hagen
The 15th ACM International Conference on Web Search and Data Mining (WSDM'22)
Phoenix (Arizona, USA)

Nov 2019 **Webis at TREC 2019: Decision Track**
A. Bondarenko, M. Fröbe, **V. Kasturia**, M. Völske, B. Stein and M. Hagen
28th International Text Retrieval Conference (TREC'19), Gaithersburg (Maryland, USA)

Jun 2018 **Ranking Archived Documents for Structured Queries on Semantic Layers**
Pavlos Fafalios, **Vaibhav Kasturia** and Wolfgang Nejdl
ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'18), Fort Worth (Texas, USA)

Nov 2017 **Building and Querying Semantic Layers for Web Archives (Extended Version)**
Pavlos Fafalios, Helge Holzmann, **Vaibhav Kasturia** and Wolfgang Nejdl
International Journal on Digital Libraries (IJDL)

Jun 2017 **Building and Querying Semantic Layers for Web Archives**
Pavlos Fafalios, Helge Holzmann, **Vaibhav Kasturia** and Wolfgang Nejdl
ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'17), Toronto (Ontario, Canada)

Jun 2017 **Towards a Ranking Model for Semantic Layers over Digital Archives**
Pavlos Fafalios, **Vaibhav Kasturia** and Wolfgang Nejdl
ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'17), Toronto (Ontario, Canada)

Technical Certifications

Aug 2021 AWS Certified Cloud Practitioner (Amazon Web Services)
Jun 2021 BSI IT-Grundschutz-Praktiker (Bitkom Akademie)
Dec 2020 Machine Learning, Stanford University (Coursera)
Mar 2020 Introduction to the Bash Shell on Mac OS and Linux (Pluralsight)
Oct 2019 Master Object Oriented Design in Java (Udemy)
Sep 2019 Complete Python Bootcamp (Udemy)
May 2019 Improving Deep Neural Networks (Coursera)
May 2019 Structuring Machine Learning Projects (Coursera)
May 2019 Neural Networks and Deep Learning (Coursera)

Extra-Curricular Activities

Oct 2019 — Sub-Reviewer
ECIR 2020 Conference, Lisbon and CHIIR 2020 Conference, Vancouver

Sep 2016 — Student Volunteer, Organizing Committee
TPDL 2016 Conference, Hannover

May 2016 — Student Volunteer, Organizing Committee
ACM WebSci'16 Conference, Hannover

Sep 2012 - Jun 2013 — General Secretary, Student Council
BITS Pilani, Dubai

Languages

English:	Native (C2)	IELTS General (Oct 2019, Overall Band: 8.0/9.0)
German:	Advanced (C1)	Goethe-Certificate C1 (Jul 2020, Overall Score: 74/100)
Hindi:	Mother tongue	

Interests and Activities

Sketching, Swimming, Photography, Traveling

Vaibhav Kasturia

Persönliche Angaben

Geburtsort:	Delhi, Indien
Geburtsdatum:	18. November 1992
Staatsangehörigkeit:	Indisch (Niederlassungserlaubnis für Deutschland)
Email:	vkasturia@deloitte.de, vbh18kas@gmail.com
LinkedIn:	vkasturia
Github:	vkasturia
Webpage:	www.vaibhavkasturia.com

Ausbildung

2015 - 18 Master of Science in Internet Technologien und Informations-Systeme(ITIS)
Leibniz Universität Hannover, Hannover
Schwerpunkt: Data and Information
Forschungsprojekt: "Building & Querying Semantic Layers for Web Archives"
Masterarbeit: "Ranking Archived Documents for Structured Queries on Semantic Layers"
GPA: 1,1 (Deutsche Notenskala)

2011 - 15 Bachelor of Engineering (Hons.) in Informatik
Birla Institute of Technology and Science - Pilani, Dubai
Bachelorarbeit: "Software Development Practices at ESRI"
GPA: 1,1 (Deutsche Notenskala), 9,77 (Indische Notenskala)

Berufliche Erfahrungen

Mai 2021 - Heute — Consultant bei Deloitte, Düsseldorf
Digital / AI Controls / Algorithms, Risk Advisory

Banking

- **Cloud-basierte Digitalisierung von Bankunterlagen:** Aufbau einer Pipeline zum prozessieren von Dokumente unter der Verwendung von Google DocumentAI Platform.

Life Sciences

- **Digitalisierung im Qualitätsmanagement:** Transformation des Produktions- und Qualitätsmanagementsystems von stark dokumentationsbasierten, funktional getrennten Systemen in eine intelligente und datengesteuerte End-to-End-Lösung.

Vermögensverwaltung

- **Maschinelle Extraktion Wertpapierprospekte:** Entwarf und Implementierte einen Prototyp der KI und regelbasierte Lösungen zum Erkennen und Extrahieren von relevanten Daten aus Dokumenten nutzt.
- **Erklärbare KI:** Entwickelte ein Tool welches mithilfe von Shapley-Werten den Entscheidungsprozess von Deep Learning Modellen wie BERT und RoBERTa erklärbar macht.

Aug 2018 - Okt 2020 — Wiss. Mitarbeiter an der Universität Halle-Wittenberg, Halle (Saale)
Big Data Analytics, Webis Group

Query Understanding mit Entity Linking

- **Ziel:** Interpretation von Mehrdeutigen Suchanfragen zur Anzeige relevanter Ergebnisse oder zum Ausfüllen von Knowledge Boxes der Suchmaschine.
- Entwarf und entwickelte einen automatischen Ansatz, der Query Segmentation und Entity Linking verwendete, um die sinnvollsten Interpretationen einer Suchanfrage auf der Grundlage der enthaltenen Entitäten zu ermitteln.
- Experimenteller Vergleich meines Ansatzes an einem neuen Korpus von 2.800 Anfragen mit aktuellen Methoden zeigt, dass mein Ansatz eine bessere Genauigkeit bei den Interpretationen und einer besseren Laufzeit liefert als die zu diesem Zeitpunkt besten Methoden.

Total Recall in Systematic Reviews

- **Ziel:** Finden aller relevanten Dokumente ("Total Recall") angesichts einer Sammlung von möglicherweise mehreren tausend Dokumenten, die in gewisser Weise mit einem benutzerspezifischen Thema zusammenhängen. Eine einzige Systematic Review kann ohne maschinelle Unterstützung bis zu 2 Jahren dauern.
- Es wurde ein System aufgebaut, das den Überprüfungszeitraum verkürzt, dabei werden die Dokumente in absteigender Reihenfolge ihrer Relevanz anordnet.
- Implementierte mehrere maschinelle Lernverfahren aus einem bestehenden Total-Recall-Ansatz (HiCAL) und testete diese an botanischen Forschungsdatensätzen. Die Ergebnisse zeigen, dass dieses maschinelle Lernen den menschlichen Aufwand um fast 80 Prozent reduziert.

Argumentatives Axiomatisches Re-Ranking für Medizinische Suchanfragen

- Häufig werden Suchmaschinen genutzt um Gesundheitsratschläge einzuholen.
- Bei der Verwendung von Suchmaschinen zur Durchführung solcher Entscheidungsaufgaben sind die Benutzer nicht in der Lage, maßgebliche von unzuverlässigen Informationen zu unterscheiden.
- Im Team entwickelten wir einen axiomatischen Ansatz zur besser ordnen von Suchergebnissen, welcher die mit traditionellen Suchmodellen erzielten Ergebnisse, passend zur medizinische Anfragen wichtet und somit übersichtlicher darstellt.

Aktivitäten

- Vorbereitung und Durchführung von Übungen für Bachelor- und Master-Studiengänge (Objektorientierte Programmierung in Java, C-Programmierung, Suchalgorithmen, Grundlagen der Informatik und Konzepte der Modellierung).
- Betreuung eines Teams von 10 Studenten bei ihrem Software-Projektpraktikum. **Thema:** Entwicklung eines Systems zur automatischen Migration älteren Excel-Datensätze in eine Datenbank.
- Pflege der Big Data Analytics Webseite.

Mai 2018 - Jul 2018 **Wiss. Mitarbeiter bei dem Fraunhofer IAIS, Sankt Augustin**

- Entwickelte einen Prototyp eines QA-Systems für eine Buchhaltungsfirma.
- Als Teil des Teams trug ich zur Entwicklung von Algorithmen im Bereich Deep sowie der Speech Processing für intelligente Fahrzeugsysteme bei.
- Kleine Beiträge zum Projekt GEISER, das sich mit der Analyse von Geodaten beschäftigte.

Okt 2016 - Jan 2018 **Studentische Hilfskraft im Forschungszentrum L3S, Hannover**

ALEXANDRIA Project

- Semantic und Entity-basierte Erforschung von Webarchiven.
- Ziel des Projekts war, die semantische und zeitbasierte Indizierung von Webarchiven um Informationen über Entitäten und Ereignisse aus der Vergangenheit effizient zu indizieren, abzurufen und zu erforschen.
- Erstellte Semantic Layers (Metadaten), die semantische Informationen über den Inhalt von Webarchiven mit Hilfe von Entity Linking Tools beschreiben.
- Evaluierte die Semantic Layers für komplexe Informationsbedürfnisse im Vergleich zu Schlüsselwort-basierten Suchsystemen wie Google, Bing und HistDiv.
- Entwickelte und evaluierte statistische und fortgeschrittene Modelle (PageRank-ähnlich), um die Ergebnisse zu ordnen, die durch das Ausführen von Abfragen auf diesen Layers zurückgegeben werden.

Aug 2014 - Jan 2015 **Softwareentwickler (Praktikant) bei ESRI, Sharjah**

- Verarbeitung Mehrdimensionale Daten (GRIB, NetCDF, HDF, usw.).
- Analysierte Raster-, Mosaik- und Image-Service Datenschichten.
- Behandlung von Fehlern und Änderungen von ArcGIS 10.
- Validierung von UI-Funktionalität der Raster- und Geo-Processing-Tools von ArcGIS Pro.
- Fehlerbeseitigung (durch Coverity-Analyse) in Raster Solutions von ArcGIS 10.

Technische Kenntnisse

Programmiersprachen:	Java, Python, C/C++
Data Science / Maschinelles Lernen:	Shap, Transformers, NumPy, pandas, scikit-learn, Matplotlib
Information Retrieval:	Apache Lucene
Datenbanken:	Graphdatenbank (Virtuoso), NoSQL Datenbank (RocksDB), SQL Datenbanken (MySQL, PostgreSQL)
Natural Language Processing:	Entity Recognition, Entity Linking, Entity Disambiguation, Word Embeddings, Query Segmentation
Java Frameworks/Bibliotheken:	Apache Maven, Apache Jena, Apache Tika, SparkJava, Standard Bibliotheken (Apache Commons, Apache Lang, usw.)
Semantic Web Technologien:	RDF/RDFa, OWL, SPARQL, Turtle
Web Technologien:	HTML, CSS, Materialize/Bootstrap, Flask, Streamlit
Geo-Information Systeme:	ArcGIS
Entwicklungsumgebungen:	Jupyter Notebook, IntelliJ, Eclipse, NetBeans, Visual Studio
Dokument Vorbereitung:	MS Office, Apple Office Suite, LaTeX
Versionsverwaltung:	Gitlab/Github, GitKraken, SVN, CVS
Weitere:	Docker, Multithreaded Programmierung
Grundkenntnisse:	Intel 8085 Programming, Wireshark, Scilab

Leistungen und Auszeichnungen

Okt 2021	Annahme von einem Research Paper auf der WSDM 2022 (A* Konferenz in Informatik).
Dez 2018	Zertifikat für den Besten Master-Abschluss im Studienjahr 2017/18 der Universität Hannover.
Jun 2017	Nominierung für *Best Paper Award* bei JCDL 2017 Konferenz.
2011 - 2015	BITS-Stipendium für Akademische Exzellenz für das gesamte Bachelors-Studium.

Veröffentlichungen

Mai 2021 **Query Interpretations from Entity-Linked Segmentations**
Vaibhav Kasturia, Marcel Gohsen und Matthias Hagen
The 15th ACM International Conference on Web Search and Data Mining (WSDM'22)
Phoenix (Arizona, USA)

Nov 2019 **Webis at TREC 2019: Decision Track**
A. Bondarenko, M. Fröbe, **V. Kasturia**, M. Völske, B. Stein und M. Hagen
28th International Text Retrieval Conference (TREC'19), Gaithersburg (Maryland, USA)

Jun 2018 **Ranking Archived Documents for Structured Queries on Semantic Layers**
Pavlos Fafalios, **Vaibhav Kasturia** und Wolfgang Nejdl
ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'18), Fort Worth (Texas, USA)

Nov 2017 **Building and Querying Semantic Layers for Web Archives (Extended Version)**
Pavlos Fafalios, Helge Holzmann, **Vaibhav Kasturia** und Wolfgang Nejdl
International Journal on Digital Libraries (IJDL)

Jun 2017 **Building and Querying Semantic Layers for Web Archives**
Pavlos Fafalios, Helge Holzmann, **Vaibhav Kasturia** und Wolfgang Nejdl
ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'17), Toronto (Ontario, Canada)

Jun 2017 **Towards a Ranking Model for Semantic Layers over Digital Archives**
Pavlos Fafalios, **Vaibhav Kasturia** und Wolfgang Nejdl
ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'17), Toronto (Ontario, Canada)

Technische Zertifizierungen

Aug 2021 AWS Certified Cloud Practitioner (Amazon Web Services)
Jun 2021 BSI IT-Grundschutz-Praktiker (Bitkom Akademie)
Dez 2020 Machine Learning, Stanford University (Coursera)
Apr 2020 Competitive Programming (Coding Ninjas)
Mär 2020 Introduction to the Bash Shell on Mac OS and Linux (Pluralsight)
Okt 2019 Master Object Oriented Design in Java (Udemy)
Sep 2019 Complete Python Bootcamp (Udemy)
Mai 2019 Improving Deep Neural Networks (Coursera)
Mai 2019 Structuring Machine Learning Projects (Coursera)
Mai 2019 Neural Networks and Deep Learning (Coursera)

Ehrenamtliche Aktivitäten

Okt 2019 — Sub-Reviewer
ECIR 2020 Konferenz, Lissabon und CHIIR 2020 Konferenz, Vancouver

Sep 2016 — Organisationskomitee
TPDL 2016 Konferenz, Hannover

Mai 2016 — Organisationskomitee
ACM WebSci'16 Konferenz, Hannover

Sep 2012 - Jun 2013 — Generalsekretär, Schülervertretung
BITS Pilani, Dubai

Sprachen

Englisch:	Muttersprache	IELTS General (Okt 2019, Gesamtpunktband: 8.0/9.0)
Deutsch:	Verhandlungssicher (C1)	Goethe-Zertifikat C1 (Jul 2020, Gesamtpunktzahl: 74/100)
Hindi:	Muttersprache	

Interesse und Aktivitäten

Zeichnen, Schwimmen, Fotografieren, Reisen