

Suchmaschinentechnologie

Modul: Suchmaschinentechnologie	
Studiengang: Bibliotheksinformatik	Abschluss: Master of Science
Modulverantwortliche/r: Sascha Szott & Frank Seeliger	

Semester: 2	Dauer: 2	
Präsenzstunden: 50.0	davon V/Ü/L/P: 10.0/10.0/15.0/15.0	CP nach ECTS: 7.0
Art der Lehrveranstaltung: Pflicht	Sprache: Deutsch	Stand vom: 2018-12-06
Pflicht Voraussetzungen: Algorithmische Grundlagen (Suchen, Sortieren, Hashing), Elementare Datenstrukturen (Arrays, Listen, Sets, Maps, Bäume), Datenbanken und SQL, XML, Programmierung mit der Programmiersprache Java, mathematische Grundlagen der Informatik (Abiturniveau)		
Empfohlene Voraussetzungen: Datenmodellierung, Grundkenntnisse Linux (Kommandozeile), Internettechnologien inklusive Schnittstellenprogrammierung, Englisch (Lesen)		
Pauschale Anrechnung von:		
Besondere Regelungen: Die Vorlesung umfasst eine integrierte Übung. Die Studierenden müssen Übungsaufgaben (i.d.R. in Form von lauffähigen Java-Programmen) erstellen. Für eine Zulassung zur Modulprüfung müssen alle Übungsaufgaben erfolgreich bearbeitet werden.		

Aufschlüsselung des Workload	Stunden:
Präsenz:	50.0
Vor- und Nachbereitung:	150.0
Projektarbeit:	0.0
Prüfung:	1.5
Gesamt:	201,5

Suchmaschinentechnologie

Lernziele	Anteil
Fachkompetenzen	
<p>Kenntnisse/Wissen</p> <ul style="list-style-type: none"> • Die Studierenden kennen die Konzepte und Herausforderungen des Information Retrievals. Sie können die Architektur von Suchmaschinen beschreiben. • Die Studierenden kennen den Relevanzbegriff. Ihnen sind unterschiedliche Modelle des Information Retrieval (Boolesches Modell, Vektorraum-Modell, Probabilistisches Modell) bekannt. • Die Studierenden kennen wichtige Evaluierungsmaße für die Effektivität und Effizienz von Suchmaschinen. • Ihnen sind Algorithmen und Datenstrukturen für die effiziente Suche in Textkollektionen bekannt, die im Information Retrieval Anwendung finden. Es werden dabei verschiedene Suchanfragetypen betrachtet (Keyword-Suche mit einem oder mehreren Termen, Phrasensuche, Proximity-Suche, Wildcard-Suche). • Die Studierenden kennen Verfahren zur Unterstützung der Korrektur von Schreibfehlern in Anfragetermen (Spell-Checking). • Die Studierenden können einen Suchserver mit Apache Solr installieren und das Schema für einen Suchindex entwerfen. Sie kennen wichtige Konfigurationsparameter für den Suchserver Apache Solr. • Die Studierenden können textuelle Daten (Metadaten, Volltexte) aus unterschiedlichen Quellen in den Suchserver Apache Solr laden (Indexierung). • Sie können Suchanfragen in der Lucene/Solr-Anfragesprache formulieren. Ferner können Sie mittels eines Java-Programms und eines Solr-Clients auf den Suchserver Apache Solr zugreifen und Anfragen absetzen sowie die Antwort auswerten. • Die Studierenden kennen das Konzept des Relevance Feedback. • Die Studierenden kennen die Aufbau von Websuchmaschinen. Sie kennen die Funktionsweise von Crawlern und Algorithmen für die Link-Analyse. • optional: Sie kennen mindestens einen Klassifikationsalgorithmus und ein Cluster-Verfahren. 	50%

Suchmaschinentechnologie

<p>Fertigkeiten</p> <ul style="list-style-type: none"> • Die Studierenden kennen die Möglichkeiten und Grenzen von Suchmaschinen. Sie können Suchmaschinen hinsichtlich ihrer Effektivität und Effizienz bewerten. • Die Studierenden können eine Suchmaschine für die Suche in einer vorgegebenen Dokumentkollektion entwerfen, umsetzen und optimieren. • Die Studierenden sind in der Lage eine Suchmaschine mit Apache Solr aufzubauen. Dazu gehört die Installation, das Schema-Design, die Indexierung und die Formulierung von Suchanfragen. 	<p>40%</p>
<p>Personale Kompetenzen</p>	
<p>Soziale Kompetenz</p> <ul style="list-style-type: none"> • Kommunikations- und Präsentationsfähigkeiten gegenüber IT-ExpertInnen und weiteren stakeholdern 	<p>10%</p>
<p>Selbstständigkeit</p> <ul style="list-style-type: none"> • eigenständiges Lösen von Übungsaufgaben im Rahmen von Problemlösungsverhalten 	

<p>Inhalt:</p>
<ol style="list-style-type: none"> 1. Einführung in das Information Retrieval und Suchmaschinen 2. Textbasiertes Information Retrieval (Grundbegriffe: Dokument, Index, Relevanz, Anfrage, Term Frequency, Document Frequency) 3. Architektur von Suchmaschinen 4. Exkurs I: Grundlagen der Mengenlehre und Aussagenlogik 5. Retrievalmodelle I: Boolesches Modell (invertierter Index, Anfrageverarbeitung, Optimierungsmöglichkeiten wie Reorganisation der Anfrageausführung und Skip Pointer) 6. Exkurs II: Grundlagen der Linearen Algebra (Vektorraum, Vektor, Vektornorm, Skalarprodukt, Matrix, Matrixmultiplikation) 7. Retrievalmodelle II: Vektorraum-Modell (Top-k-Rankings, TF-IDF-Gewichtung, Gesetz von Zipf, Cosinus-Ähnlichkeit, Term-at-a-Time-Algorithmus, Prioritätswarteschlange, ungenaues Top-k-Retrieval, Document-at-a-Time-Algorithmus) 8. Exkurs III: Grundlagen der Wahrscheinlichkeitsrechnung (Wahrscheinlichkeitsraum, bedingte Wahrscheinlichkeit, Satz von Bayes, Gesetz der totalen Wahrscheinlichkeit, stochastische Unabhängigkeit, Zufallsvariable, Chance) 9. Retrievalmodelle III: Probabilistisches Modell (Probabilistic Ranking Principle, Binary Independence Retrieval-Modell, Okapi BM25)

Suchmaschinentechnologie

10. Indexstrukturen für die Unterstützung spezieller Suchanfragetypen (Phrasen-Suche, Proximity-Suche, Wildcard-Suche)
11. Algorithmen für die Korrektur von Schreibfehlern in Anfragetermen (Spell-Checking)
12. Evaluierung der Effizienz von Suchmaschinen (Laufzeit, Speicherplatz, Durchsatz, Latenz)
13. Evaluierung der Effektivität von Suchmaschinen (prinzipielles Vorgehen, Recall, Precision, F-Measure, alternative Bewertungsmaße für die Bewertung von Rankings)
14. Relevance Feedback (Prinzip, Vorstellung der unterschiedlichen RF-Varianten, Rocchio-Algorithmus und Erweiterungen)
15. Textverarbeitung (Dokumentvorverarbeitung, Indexaufbau, Komprimierungsalgorithmen)
16. Praktische Einführung in den Suchserver Apache Solr
17. Indexierung und Suche mit Apache Solr
18. Websuchmaschinen (Crawler, Algorithmen für die Link-Analyse (PageRank))
19. Klassifikationsverfahren (optional)
20. Clustering-Verfahren (optional)

Prüfungsform:

mündliche oder schriftliche Abschlussprüfung (100%)

Zusätzliche Regelungen:

Für eine Zulassung zur Modulprüfung müssen alle Übungsaufgaben erfolgreich bearbeitet werden.

Suchmaschinentechnologie

Pflichtliteratur:
Empfohlene Literatur:
<p>D. Manning, C. & Raghavan, P. & Schütze, H. (2008). <i>Introduction to Information Retrieval</i> by Manning, Christopher D., Raghavan, Prabhakar, Schütze, (2008) Hardcover. Cambridge University Press.</p> <p>Bruce Croft, W. & Metzler, D. & Strohman, T. (2010). <i>Search Engines: Information Retrieval in Practice</i>. Addison-Wesley.</p> <p>Klose, M. & Wrigley, D. (2014). <i>Einführung in Apache Solr</i>. O'Reilly Verlag GmbH & Co. KG.</p> <p>Baeza-Yates, R. & Ribeiro-Neto, B. (2010). <i>Modern Information Retrieval (ACM Press Books)</i>. Addison Wesley.</p>