Vorläufiger Arbeitstitel

Betreuer:
Dipl.-Ing. Dr.techn. Roman Kern

Thomas Mauerhofer (1031957)

Graz, am 16. Mai 2017

Inhaltsverzeichnis

1 Einleitung

In der Gegenwart gehört die Suche nach Informationen via Internet und im Besonderen Google, zum Alltag der meisten Menschen. Hatte Google im Jahr 1998 noch 10.000 Suchanfragen am Tag, wurde dieselbe Zahl an Anfragen 2006 in einer Sekunde getätigt. (Quelle: http://www.internetlivestats.com/google-search-statistics/ –> Zugriff am 15.05.2017). Dadurch ist es nur naheliegend, dass auch die Anzahl an zugängichen Informationen stetig zunimmt. Täglich werden z.B. unzählige neuen Webseiten erstellt, Artikel geschrieben und wissenschaftliche Arbeiten veröffentlicht. Um diese Menge an Informationen zu managen und zwischen relevanten und nicht relevanten Quellen zu unterscheiden, verwendet man unterschiedliche Suchmaschinen. Dabei soll die Suche simpel aber dennoch genau sein.

Betrachtet man Suchmaschinen für wissenschaftliche Arbeit und Forschung existiert eine große Auswahl an Möglichkeiten. Eine der bekanntesten Suchmaschinen im wissenschaftlichen Bereich ist wohl Google Scholar, welches eine einfache Eingabe zur Suche verwendet und die gefundenen Arbeiten entsprechend ihrer Relevanz listet. Diese Listungen enthalten verschiedene Dateiformate aus unterschiedlichen Jahren zu diversen Themen, die nicht immer die Kriterien der Suchanfrage wiedergeben. Gerade für wissenschaftliche Arbeiten und Forschung ist es allerdings wichtig, ein präzises Suchergebnis zu erhalten.

Diese Arbeit befasst sich mit der Verbesserung von Suchanfragen und deren Ergebnissen für wissenschaftliche Arbeiten in PDF Format. Recherchiert man z.B. einen Autor, erhält man mit den gängigen Suchmaschinen nicht nur die veröffentlichten Artikel und Bücher, sondern häufig auch alle Quellen in welchen dieser zitiert wurde. Ziel dieser Arbeit ist es die Eingabe des Suchbegriffes robuster zu gestalten, die Usability der Suchmaschine zu erhöhen und so treffsicherere Ergebnisse zu liefern. Dies geschieht im Hilfe von einfachen search queries Strukuren, durch selbsterklärendes Front End, Spelling Checks und der Bearbeitung und Beurteilung im Back End.

2 Forschungsstand

Wie bereits in der Einleitung erwähnt befasst sich die geplante Masterarbeit mit der Verbesserung von Suchanfragen und deren Ergebnissen für wissenschaftliche Arbeiten in PDF Format. Laut [SK14] sind alle wissenschaftliche Arbeiten in ähnlichen Strukturen aufgebaut. Diese Strukturen unterteilen sich in capters, sections, subsections and so on, welche sich gut in dem [RBY99] beschriebene Structured Text Retrieval Model sehr gut anwenden lassen. Dieses Model beschreibt den Umgang von Suchanfragen und deren Ergebnissen für wissenschaftliche Arbeiten, sowie deren Bearbeitung und Beurteilung im Backend durch die Verwendung der Metainformation über die Strucktur des Dokuments. Um die Dokumente nach ihrer Priorität zu listen kommen verschiedene Ranking stradegies z.B. Jelinek-Mercer smoothing, zum Einsatz. Durch Erweiterungen wie contextualization or aggregation strategies werden die Ergebnisse des rankings noch verbessert.

3 Aufbau

was im Forschungsstand gefunden wurde –> was man weiter bearbeitet in der Arbeit, was man anders macht, welches Thema in der Literatur noch nicht ausreichend dargestellt wurde –> alles auf die eigene Arbeit beziehen

4 Gliederung

1. Introduction
2. Related Work
   a) Robuste Gestaltung des Suchbegriffes
   b) Search queries Strukturen
   c) Front End
   d) Spelling Checks
   e) Back End
3. Implementierung
   a) Basisstruktur
   b) Aufbau der Datenbank
   c) Userinterface
   d) Satisfaction of searchqueries
   e) Ranking system
   f) Spelling checks
4. Result
5. Conclusion

5 Auswahlbibliografie

Literatur

[CDM08] H. Schütze Ch. D. Manning, P. Raghavan. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[RBY99] B. Ribeiro-Neto. R. Baeza-Yates. *Modern Information Retrieval*. ACM Press, 1999.

[SC03] Y. Kanza Y. Sagiv S. Cohen, J. Mamou. Xsearch: A semantic search engine for xml. *Proceedings of the 29th international conference on Very large data bases*, 29:45–56, September 2003.

[SK14] K. Jack R. Kern St. Klampfl, M. Granitzer. Unsupervised document structure analysis of digital scientific articles. *International Journal on Digital Libraries*, 14:83–99, August 2014.

6 Zeitplan

Arbeitsschritt	**Aufgabe im Detail**	**Deadline**
Implementierungsphase	Erstellen einer Basisstruktur	
	Datenbank	
	Import von PDFs via pdf-extractor	
	Satisfaction of search queries and simple ranking system	
	Improve ranking system	
	Implement spelling checks	
	Improve Userinterface	
Schreibphase		
Abschlussarbeit	Korrektur	Deadline
	Druck und Abgabe	Deadline