

Semantische Suche mit Apache Solr

Projekt Big Data, Wintersemester 2017/18

Minh Hieu Nguyen, Eike Nils Knopp

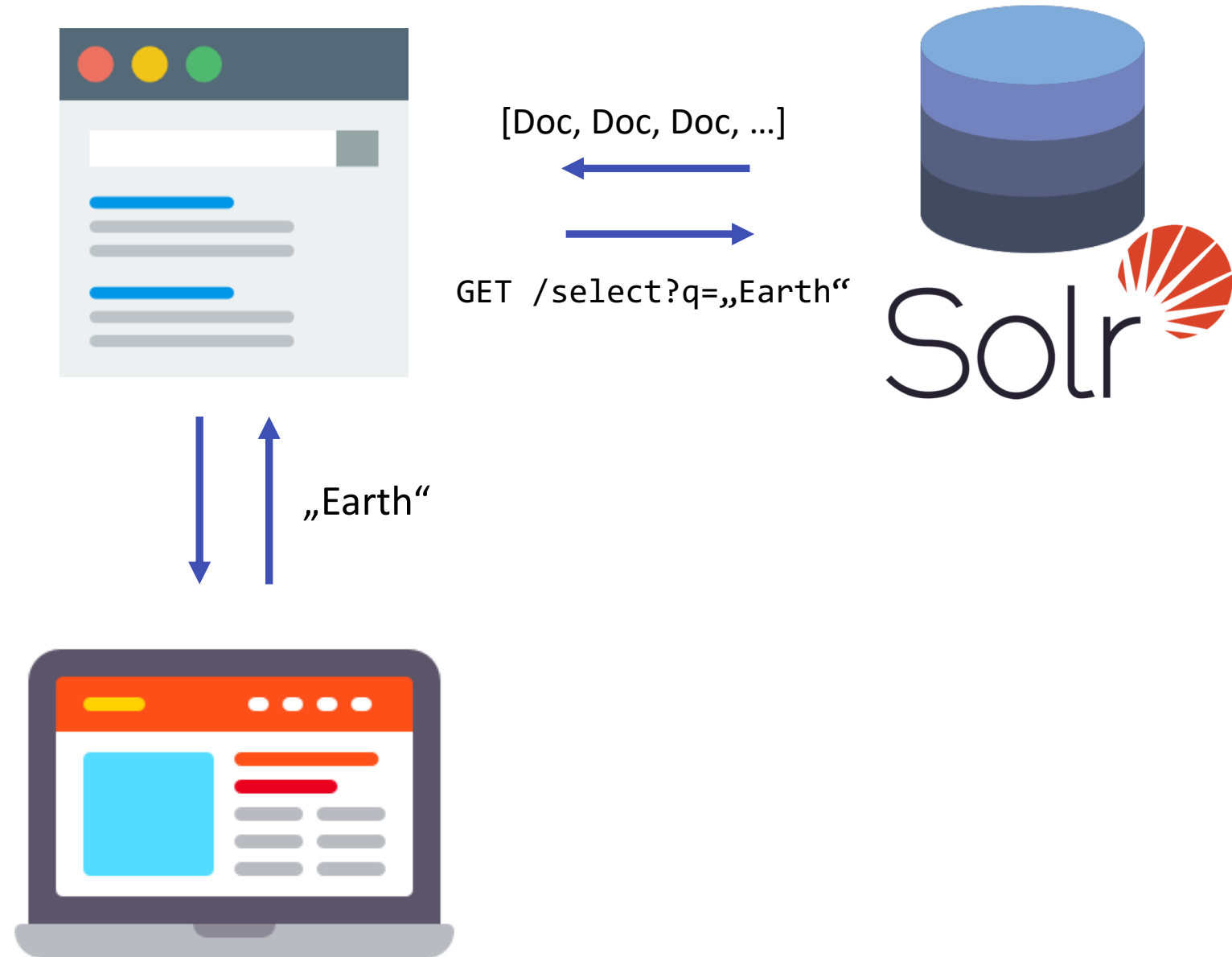
Aufgabenstellung

- Identifizierung von Strategien für semantische und effiziente Suche in wissenschaftlichen Daten
- Verbesserung der Erkundbarkeit & Durchsuchbarkeit der Daten

Vorgehensweise

- Erstellung eines Solr-Servers
- Entwicklung einer Website als Suchinterface für Solr
- Implementierung von passenden Lösung für semantische Suchen & Optimierung der Sucherfahrung

Architektur der Anwendung



Basis Funktionalitäten


Publication Time

From
July 2016

To
January 2018

APPLY

Author

Sautter 

Li (343)

Zhang (255)

Yang (242)

India

SEARCH

SEARCH RESULTS

RESULTS CLUSTERING

STATISTICS

Page 1 of 25111 results.

[Specimens from India at the University Museum at the University of Bergen](#)

by Hjelle, Boumans

Language: English | Group: Gbif | Tags: None | Published on 6/19/2017

This dataset includes specimens originating from India in the natural history collections at the University Museum of the the University of Bergen (UiB). **Animals**: The vertebrate collection contains 18 reptile specimens as well as a few bird eggs and fishes from India....

[Need For A National Resource Sharing Network in India: Proposed Model](#)

by Sahoo, Bibhuti Bhusan

Language: English | Group: Sdl | Tags: Article | Published on 7/1/2002

The paper deals with the need of Resource Sharing Network in India. The objective of this network is to develop resource sharing strategy for India and make cooperation among different types of LIC networks which include National Library of India. Internet technologies have brought a drastic chan...

...

< **1** 2 3 4 5 6 ... 2511 2512 2513 >

Synonyme

- Durchsuche die Daten nicht nur nach dem Suchbegriff, sondern auch nach Synonymen des Begriffs
- Mehr Ergebnisse für semantisch ähnliche Dokumente
- Nicht praktikabel für große Datenmengen

Vereinheitlichung des Sprach-Filters

- Sprachcodierungen als Synonymen betrachten:
z.B. en, Eng, en_us, english
- Probleme mit Multi-Wörter Synonymen wie z.B.:
„Modern greek“

Language

Filter 9-1

English (3574)

Dutch (611)

dutch (486)

english (427)

German (92)

Arabic (11)

Spanish (9)

Swedish (9)

ger (2)

swedish (2)

More

B2FIND EUDAT



Language

English (14107)

Spanish (5467)

Italian (1280)

German (754)

Portuguese (630)

Modern greek (1453-) (382)

Polish (329)

Turkish (326)

Catalan (246)

Serbian (207)

MORE

Word-Stemming

- Annahme: Worte mit gleichem Wortstamm sind semantisch ähnlich
- Rückführung eines Wortes auf seinen Wortstamm
- Suche nach dem Wortstamm des Suchbegriffs
- Alle Daten, die den Wortstamm enthalten, werden gefunden

Word-Stemming

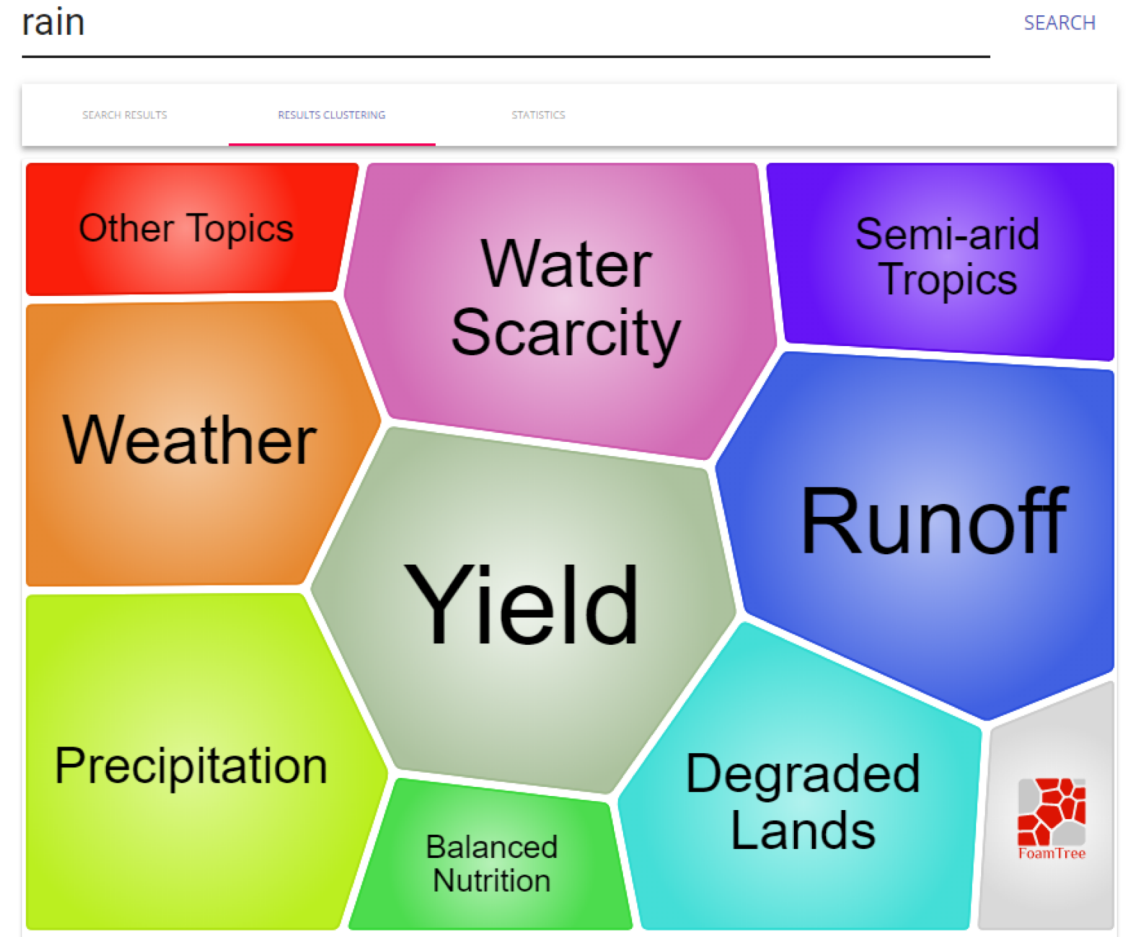
- Rain, raining, rained → rain
- Walker, walking, walked → walk

Auch wird nicht mehr zwischen Singular und Plural unterschieden:

- Cloud → cloud
- Clouds → cloud

Result-Clustering mit Carrot2

- Gruppierung der Ergebnisse in semantische Gebiete
- Generierung von Labels für Gruppen aus Inhalt der Ergebnisse



Clustering-Ergebnisse des Suchbegriffs "rain"

Result-Clustering mit Carrot2

- Qualität der Ergebnisse stark abhängig von den Daten
- Sehr wissenschaftliche Daten (Messwerte etc.) führen zu unverständlichen Labels

Ausblick

- Einfachere Methodik zur Durchsichtung bestimmter Felder