

Kontextbasierte Unternehmenssuche

Studiengang: BSc in Informatik | Vertiefung: Digital Business Systems
Betreuer*innen: Ursula Deriu, Prof. Dr. Stephan Fischli
Experte: Prof. Dr. Torsten Braun

Kann bei der SBB eine kontextbasierte Suche mit Elasticsearch umgesetzt werden und wenn ja, wie?

Einleitung

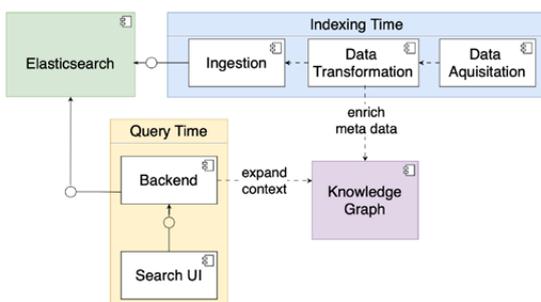
Die SBB strebt die Implementierung einer unternehmensweiten Suche an, die es den Mitarbeitenden ermöglicht, Daten aus verschiedenen Quellsystemen zu durchsuchen. Ein besonderer Fokus liegt dabei auf der kontextbasierten Suche: Die Suchergebnisse sollen nicht nur in Bezug auf die Suchfrage relevant sein, sondern auch in Bezug auf den Kontext der Benutzer:innen, insbesondere ihrer Rolle im Unternehmen und ihrer Organisationseinheit.

Das Ziel dieser Arbeit ist es, ein Proof-of-Concept (PoC) dafür zu entwickeln. Als zentrale Suchkomponente wird Elasticsearch eingesetzt. Es werden zwei Zielgruppen (Sicherungsanlagen und IT-Ausbildungsteam) ausgewählt sowie ein fixer Datensatz, der von diesen Gruppen verwendet wird.

Konzept

Um den Kontext der Mitarbeitenden besser zu verstehen, wird ein Knowledge Graph erstellt. Dieser Knowledge Graph verknüpft die Mitarbeitenden mit ihrer jeweiligen Rolle und Organisationseinheit. Während des Indexierungsprozesses werden die Metadaten der Dokumente mit Informationen aus dem Knowledge Graph erweitert.

Zur Abfragezeit werden die Daten der eingeloggten Benutzer:in aus dem Knowledge Graph ausgelesen, um damit bestimmte Suchtreffer kontextbezogen höher zu ranken.



Architekturdiagramm

Vorgehen

Die Umsetzung des PoC erfolgt inkrementell in fünf Versionen. Jede Version baut auf der vorherigen auf und enthält neue Features. Nach jeder Iteration wird die Qualität der Suche anhand von zielgruppenspezifischen Suchfragen und dazugehörigen erwarteten Resultaten (Goldstandard) gemessen. Um die Versionen vergleichen zu können, werden ausgewählte Metriken verwendet.

Features

- Version 1: Die Dokumente werden nach Elasticsearch Standardeinstellungen indiziert und abgefragt (Zerlegung der Dokumente in Wörter und Kleinschreibung).
- Version 2: Häufige, aber nicht informative Wörter werden entfernt. Fehlertoleranz wird eingefügt (trotz Tippfehler werden Resultate gefunden). Die Indexierung deutschsprachiger Dokumente wird durch die Zerlegung zusammengesetzter Wörter verbessert.
- Version 3: Autovervollständigung bei der Suche wird hinzugefügt.
- Version 4: Der Kontext beeinflusst das Ranking der Resultate. Der Einfluss des Kontexts kann durch Benutzereinstellungen gewichtet werden.
- Version 5: Die Resultate können nach Quellsystem gefiltert werden.

Schlussfolgerung

Durch die Analyse der Metriken lässt sich aufzeigen, dass die Suche in Bezug auf den Goldstandard von Version zu Version besser wird. Der Kontext kann sich sowohl positiv als auch negativ auf die Suchresultate einwirken, je nach Gewichtung, Suchfrage, Datenbestand und weiteren Faktoren. Wie genau diese Faktoren das Ranking beeinflussen, muss weiter analysiert werden. Damit eine qualitative Suche entstehen kann, muss die Suche mit mehr Daten, mehr Benutzergruppen und repräsentativerem Goldstandard getestet werden.



Ann-Sophie Junele



Nadia Salomé Suter