OST
Ostschweizer
Fachhochschule

Moderne Textsuche mit Datenbanken

Hybrid Text Search mit PostgreSQL

Präzisere Suchergebnisse durch die Fusion von klassischer Keyword-Suche und KI-basierter Semantiksuche.

Hybrid Text Search ist eine Kombination aus lexikalischer Suche (Keyword Search) und semantischer Suche (Vector Search), um relevante Informationen präziser zu finden. Dadurch werden nicht nur exakte Begriffe erkannt, sondern auch Zusammenhänge und Bedeutungen verstanden.

Für die Umsetzung wurde die Datenbank PostgreSQL mit den Erweiterungen pg_vector und pg_textsearch (BM25) eingesetzt. Die für die semantische Suche erforderlichen Embeddings wurden mit der Python-Bibliothek sentence_-transformers und dem LLM paraphrase-multilingual-MiniLM-L12-v2 berechnet.

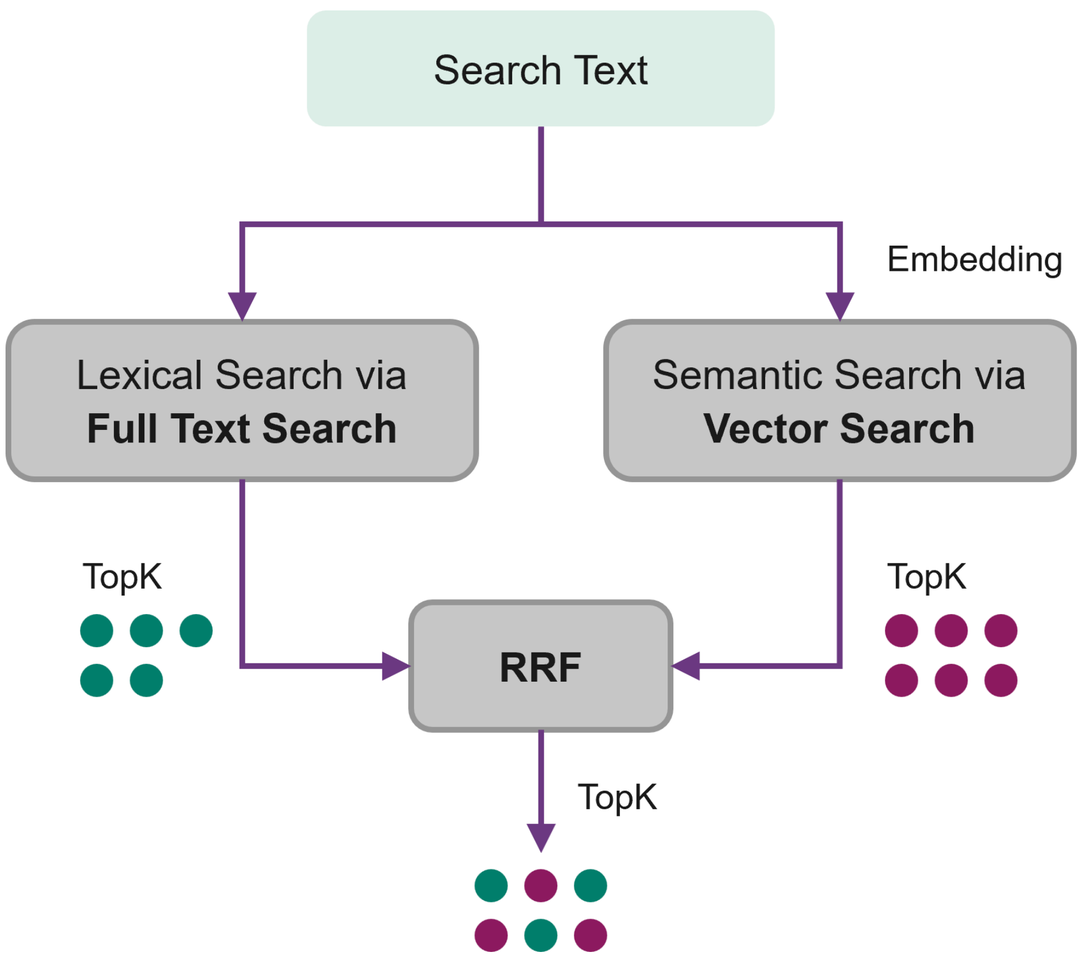

Abb. 1 Datenfluss von Hybrid Text Search

Bei einer Suchabfrage werden beide Algorithmen parallel ausgeführt. Die Suchergebnisse werden mittels Reciprocal Rank Fusion (RRF) fusioniert (zusammengeführt). RRF basiert auf den Rangpositionen der Treffer und erfordert keine Score-Normalisierung.

Die Tabelle unten zeigt einen Vergleich der Algorithmen. Als Datenbasis wurden zehn Dokumente zum Thema Ozon durchsucht (Quelle: Soekia.ch).

Search Text	Relevant Docs	Lexical Search		Semantic Search		Hybrid Search	
		Docs found	AP	Docs found	AP	Docs found	AP
Atembeschwerden Ozon	[9, 3, 4]	[9, 5, 3, 4]	0.94	[3, 7, 6, 9]	0.59	[3, 9, 5, 7]	0.72
Ozon-Alarm Schwimmbad	[4]	[4, 5, 3, 9]	1.0	[3, 4, 7, 6]	0.63	[4, 3, 9, 5]	1.0
Ausdehnung Ozonloch	[1, 6, 7, 10]	[1, 6, 10]	0.9	[3, 7, 6, 1]	0.59	[1, 6, 3, 7]	0.88
Geschichte Ozonloch Antarktis	[6, 1, 7, 10]	[10, 1, 6]	0.74	[6, 1, 10, 7]	1.0	[6, 10, 1, 7]	1.0
Ozonloch Ozon Loch	[1, 6, 7, 3]	[4, 10, 1, 8]	0.24	[3, 6, 7, 1]	0.86	[4, 6, 1, 3]	0.64
Ozonverlust Tonnen	[1, 6]	[1]	0.76	[3, 6, 7, 1]	0.57	[1, 3, 6, 7]	0.95
Ozonschicht UV - Bestrahlung	[7, 3, 6]	[7, 3, 10]	0.84	[7, 3, 10, 6]	0.98	[7, 3, 10, 6]	0.98
Ozon O3 Sauerstoff	[3]	[3, 5, 4, 9]	1.0	[3, 7, 6, 1]	1.0	[3, 9, 5, 7]	1.0
Ozonloch UV - Bestrahlung	[6, 10, 7, 1]	[10, 7, 3, 1]	0.58	[7, 3, 10, 6]	0.66	[7, 10, 3, 6]	0.7
Luftverschmutzung Städte	[9, 3]	[]	0.0	[9, 4, 3, 6]	0.95	[9, 4, 3, 6]	0.95
Gefahr für Badegäste im Sommer	[4, 9]	[4, 9]	1.0	[4, 9, 10, 3]	1.0	[4, 9, 10, 3]	1.0
Erhöhte Krebsgefahr durch Sonnenlicht	[7, 10]	[]	0.0	[10, 7, 9, 3]	1.0	[10, 7, 9, 3]	1.0
Abbau der Erdatmosphäre durch Industrieprodukte	[6, 9, 1]	[]	0.0	[2, 9, 7, 6]	0.48	[2, 9, 7, 6]	0.48
Mean Average Precision			**0.56**		**0.74**		**0.80**

Abb. 2 Vergleich der beiden Algorithmen

Um die Qualität der Suchergebnisse zu prüfen, wurde die Average Precision (AP) berechnet. Diese zeigt, ob sich die manuell als relevant eingestuften Dokumente (Relevant Docs) zu Beginn der Ergebnisliste befinden.

Resultat: Im Durchschnitt liefert Hybrid Text Search mit 0,80 die besseren Suchergebnisse als jeder Algorithmus einzeln.

Prof. Stefan Keller & IFS-Team
IFS Institut für Software, FH OST
stefan.keller@ost.ch

ost.ch/ifs