



TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Fakultät Elektrotechnik Feinwerktechnik Informationstechnik

Entwicklung eines Suchalgorithmusprototyps zur Bewertung von Suchergebnissen verschiedener Kategorien

Studienarbeit im Studiengang Software Engineering

vorgelegt von

Marc Jonas Roser

Matrikelnummer 364 7316

Betreuer:

Prof. Dr. Hans-Georg Hopf

Vorgelegt am 04.11.2022

© 2022

Dieses Werk einschließlich seiner Teile ist **urheberrechtlich geschützt**. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Autors unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen sowie die Einspeicherung und Verarbeitung in elektronischen Systemen.

Kurzdarstellung

Das Ziel der vorliegenden Studienarbeit ist es, eine bestehende Datenbank mit multimedialen Inhalten möglichst effizient nach unterschiedlichen Kriterien zu durchsuchen. Suchergebnisse sollen nach bestimmten Kriterien gewichtet, gefiltert und sortiert werden. Vorschläge für eine weiterführende Navigation auf der Suchergebnisseite sollen angeboten werden und Suchergebnisse sollen dazu nach Kontext und Wahrscheinlichkeiten gewichtet angezeigt werden. Das theoretische Fundament dieser Arbeit stellt die wissenschaftliche Betrachtung der Methoden zur Bewertung der Relevanz von Suchergebnissen dar. Die Arbeit untersucht die Möglichkeit, einen Suchbegriff so zu analysieren, dass ein Nutzer die bestmögliche Ergebnisliste bzw. zielgerichtete weiterführende Navigationsmöglichkeiten erhält. Die bestehende Anwendung „Crossload“ wird vorgestellt, um dem Leser einen Kontext zu bieten, in der sich die Entwicklung bewegt.

Abstract

The goal of this student research project is to search an existing database with multimedia content as efficiently as possible according to various criteria. Search results are to be weighted, filtered, and sorted according to certain criteria. Suggestions for further navigation on the search results page are to be offered, and search results are to be displayed weighted according to context and probabilities. The theoretical foundation of this work is the scientific consideration of methods for evaluating the relevance of search results. The work examines the possibility of analyzing a search term in such a way that a user receives the best possible list of results or targeted further navigation options. The existing application „Crossload“ is presented to provide the reader with a context in which the development takes place.

Eidesstattliche Erklärung

Hiermit versichere ich, Marc Jonas Roser, ehrenwörtlich, dass ich die vorliegende Studienarbeit mit dem Titel: „Entwicklung eines Suchalgorithmusprototyps zur Bewertung von Suchergebnissen verschiedener Kategorien“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Nürnberg, 04.11.2022

Marc Jonas Roser

Glossar

API

Application Programming Interface.

AWS

Amazon Web Services: Cloud Dienste gehostet von Amazon.

CD

Continuous Delivery / Continuous Deployment: Automatisches Ausrollen der neuen Funktionalität.

CI

Continuous Integration: Frühes Integrieren kleiner Änderungen in den Hauptzweig (Git Branches).

Crossload

Plattform zum Durchsuchen und Anhören einer umfassenden Predigt Datenbank.

JSON

JavaScript Object Notation: Dateiformat, das dem von Objekten in JavaScript gleicht.

Lucene

Open Source Suchbibliothek der Apache Foundation.

Matomo

Open Source Web Analyse Tool.

Mongo DB

Nicht relationale Datenbank.

Solr

Open Source Suchframework der Apache Foundation.

Suchmaschine

Eine Anwendung, die gezielt Ergebnisse aus dem Internet für den Nutzer aufbereitet und sortiert. Englisch: „Search Engine“.

Trial-and-Error

Wiederholtes Ausprobieren ohne Erfolgsgarantie mit Änderung der Startparameter um zum gewünschten Ziel zu gelangen.

UI

Benutzeroberfläche, der Teil der Anwendung, der für den Nutzer sichtbar und nutzbar ist. Englisch: „User Interface“.

Inhaltsverzeichnis

1. Einleitung	1
1.1. Relevanz des Themas	1
1.2. Ausgangssituation	1
1.3. Zielsetzung & Vorgehen	2
2. Theoretische Grundlagen	4
2.1. Relevanz	4
2.2. Methoden zur Bewertung von Relevanz	4
2.2.1. Textuelle Relevanz	5
2.2.2. Relevanz durch Attribute	5
2.2.3. Hyperlink Relevanz	6
2.2.4. Relevanz durch Nutzerverhalten	6
2.2.5. Performance	7
2.3. Auswertung der Relevanz von Suchergebnissen	8
3. Technische Grundlagen	10
3.1. Apache Lucene	10
3.2. Apache Solr	10
4. Crossload	12
4.1. Einführung	12
4.2. Technische Architektur	12
4.3. Suche bei Crossload	13
5. Anforderungen und Problemanalyse	15
5.1. Vorgehensweise	15
5.2. User Stories	15
6. Konzeption	17
6.1. Gemischte Ergebnisliste	17
6.2. Zuordnung des Suchbegriffes zu einer Kategorie	17
6.3. Vorschläge für weitere Navigation	18
7. Entwicklung des Prototyps	20
7.1. Gemischte Ergebnisliste	20
7.2. Zuordnung des Suchbegriffes zu einer Kategorie	21
7.3. Vorschläge für weitere Navigation	22
8. Auswertung	26

9. Fazit	27
A. Anhang	A
I. Bilder	A
I. Einleitung	A
II. Entwicklung	C
II. Source Code	D
Abbildungsverzeichnis	I
Tabellenverzeichnis	J
Literaturverzeichnis	K

1. Einleitung

1.1. Relevanz des Themas

Suchalgorithmen und relevante Suchergebnisse sind derzeit so relevant wie noch nie zuvor. Dabei wollen die Benutzer einer Suchmaschine in Sekundenbruchteilen Ergebnisse, die am besten zu ihrem Suchbegriff passen, ohne sich dabei viel Gedanken über die Formulierung eines solchen Begriffes zu machen. Ein Beispiel für einen solchen Algorithmus ist Google, welches seit den frühen 2000ern einen kometenhaften Aufstieg in der Welt der Suchmaschinen hinter sich hat, was anhand der erreichten Werbeeinnahmen sichtbar wird.¹

Google ist im Vergleich zu anderen Suchmaschinen so stark verbreitet², dass mittlerweile sogar der Duden das Verb „googeln“ als eigenen Begriff für die Recherche im Internet führt.³

Für die Entwicklung eigener Produkte stellt sich dabei die Frage, wie aus einem Suchbegriff, der meist nur aus wenigen Wörtern bis zu einem ganzen Satz besteht, relevante Suchergebnisse gefunden werden können. Dies würde in dem entwickelten Produkt zur Akzeptanz der Nutzer im Hinblick auf die entwickelte Funktionalität führen, da gewünschte Ergebnisse schneller und ohne großen Aufwand gefunden werden können, ähnlich wie es bei Google bereits der Fall ist.

1.2. Ausgangssituation

Derzeit besteht bei Crossload⁴ eine Plattform zur Durchsuchung einer umfangreichen Predigt Datenbank, welche mit einer Such API auf Basis von Spring Boot und Solr ausgestattet ist. Diese teilt auf der Suchergebnisseite die Ergebnisse nach Kategorien auf und somit können nur schwer übergreifende Suchanfragen getätigt werden. Zwar werden alle Treffer auf der gleichen Seite angezeigt, doch durch die Aufteilung nach Kategorien werden Ergebnisse gewisser Kategorien über anderen gezeigt, auch wenn niedrig positionierte Kategorien relevantere Ergebnisse enthalten.⁵

Durch eine Verbesserung der Relevanz, sowie einfacheres Suchen und weiterführende Vorschläge kann die Nutzerakzeptanz der Webseite weiter gefördert werden, da schneller bzw. überhaupt gesuchte Inhalte gefunden werden. Gefundene Inhalte werden direkt auf Crossload angehört, weswegen dadurch die mittlere Nutzungsdauer der Seite gesteigert wird.

¹Siehe A.1

²Siehe A.2

³Vgl. Duden [1]

⁴Siehe Crossload.org [2]

⁵Siehe A.3

1.3. Zielsetzung & Vorgehen

Das Ziel der vorliegenden Studienarbeit ist es, für die oben genannte Problemstellung einen Prototyp zur Erweiterung und Verbesserung des bisher genutzten Suchalgorithmus bei Crossload zu entwickeln.

Einleitend wird ein Einblick in die Grundlagen der Relevanz sowie mögliche Methoden und Funktionen zur Bewertung gegeben, sowie auf eine finale Auswertung der gesammelten Methoden eingegangen. Die hier erarbeiteten Grundlagen und Methoden werden im weiteren Verlauf mit in die Entwicklung einfließen. Anschließend folgt eine kurze Einleitung zu Solr, der genutzten Search Engine von Crossload, sowie zu Crossload selbst. Dieser theoretische Teil der Arbeit basiert größtenteils auf einer Literaturrecherche. Google Scholar, relevante Dokumentationen oder die einfache Google Suche stellen dazu die Grundlage dar. Die gefundenen Ergebnisse werden auf Qualität und Themenbezug geprüft.

Anhand der gegebenen Aufgabenstellung werden Anforderungen nach dem SMART Prinzip⁶ erarbeitet, die als Grundlage der darauffolgenden Konzeption und Entwicklung dienen sollen. Ebenso werden bereits etablierte Tools genutzt, um die Anforderungen weiter zu verfeinern.

Für die Entwicklung wird dabei das Wasserfallmodell genutzt. Obwohl es in seinen Nachteilen gegenüber z. B. agilen Methoden überwiegt, bietet es doch wenig Aufwand um die eigentliche Entwicklung herum und stellt eine klare Struktur bereit. Diese hilft, schnell ein Produkt oder wie in diesem Fall einen Prototyp, fertigzustellen. Das Modell ist auch vorteilhaft, weil die Anforderungen in ihrer Gesamtheit schon bekannt sind, bzw. vor der Entwicklung sein werden und keine weiteren Faktoren hinzukommen. Die dafür verwendeten Phasen, Anforderungserhebung, Entwicklung und Test der Anforderungen, werden in nachfolgender Abbildung verdeutlicht.

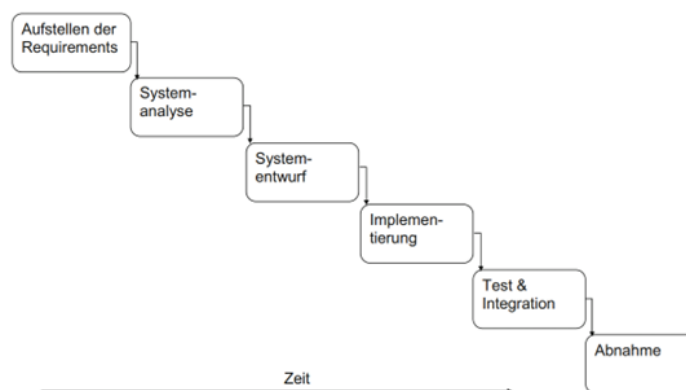


Abbildung (1.1) –Grundform eines Wasserfallmodells ohne Machbarkeitsanalyse. Nach Goll [4].

⁶Vgl. Witte 2019a, S. 67 [3]

Im Rahmen der Konzeption wird auf die bisherige Funktionalität eingegangen, um mögliche Probleme und Verbesserungspotenzial aufzudecken. Anschließend wird eine neue Berechnung des Relevanzscores mithilfe der erarbeiteten Methoden zur Berechnung der Relevanz geplant. Die Konzeption wird abgeschlossen mit der Planung der Entwicklung, mit welcher die gesammelten Anforderungen umgesetzt werden sollen.

Die Implementierung umfasst, die nach Kontext und Wahrscheinlichkeiten gewichtete und gefilterte Suche über eine Datenbank mit Datentypen verschiedener Kategorien bei der zusätzlich Vorschläge zur weiteren Navigation auf der Suchergebnisseite gegeben werden sollen.

Schlussendlich werden die Ergebnisse der Entwicklung zusammengefasst, die Anforderungen und die Implementation bewertet, sowie durch einen Ausblick, wie der Prototyp in einen produktiven Betrieb übergehen könnte, und mit einem Fazit abgerundet.

Im Lauf der Arbeit werden die folgenden Fragen beantwortet:

- Was ist Relevanz?
- Wie wird Relevanz in Suchmaschinen berechnet bzw. bewertet?
- Welche Anforderungen ergeben sich an ein solches Plugin?
- Mit welchen Methoden kann die aktuelle Implementierung verbessert werden?
- Welche Vorteile, Nachteile oder Hürden bringt die Implementation mit sich?

2. Theoretische Grundlagen

2.1. Relevanz

Relevanz ist allgemein beschrieben eine Beziehung zwischen einem Individuum, dem zeitlichen Rahmen, in welchem dieses eine Information benötigt und einer beliebigen Information.¹ Das bedeutet, dass Relevanz von Person zu Person unterschiedlich ist, da zum einen diverse Informationen nur zu einer bestimmten Zeit notwendig bzw. wichtig sind und zum anderen der Kontext der benötigten Information sich ständig ändert.

Um zu verstehen, woher die Relevanz stammt bzw. in der Informationstechnik verwendet wird, ist es wichtig, die Gewinnung von Informationen aus Objekten (*Information Retrieval*) zu verstehen. Dieser Teil der Wissenschaft beschäftigt sich mit dem präzisen Abruf von Informationen, um das Informationsbedürfnis (*Information need*) eines Nutzers zu stillen. Dieses Informationsbedürfnis wird von einem idealen Suchobjekt befriedigt, stellt also die Spezifikation eines idealen Objekts dar. Diese Spezifikation geht über den reinen textuellen Inhalt der Suche hinaus und umfasst auch kontextuelle, zeitliche und andere Aspekte. Die Bestimmung der Relevanz ist dabei die Aktivität bzw. Praxis um diesen idealen Inhalt zu finden.²

Im Falle von Crossload ist das ideale Objekt beispielsweise eine bestimmte Predigt, die zu einem bestimmten Zeitpunkt in einer bestimmten Kirche gehalten wurde oder ein Bild, welches zu einer gewissen Bibelstelle gehört. Im Beispiel Google, ist das Objekt eine bestimmte Webseite, die zu einem bestimmten Suchbegriff passt. Aufgabe von Suchmaschinen ist es, die Relevanz zwischen dem Suchbegriff und dem Objekt zu bestimmen und dem Nutzer die Objekte anzuzeigen, die dem Suchbegriff am nächsten kommen.

2.2. Methoden zur Bewertung von Relevanz

Eine Suchmaschine gibt nach einer Suchanfrage Ergebnisse standardmäßig sortiert nach der Relevanz der Ergebnisse abhängig zum gegebenen Suchbegriff des Nutzers zurück. Die Schwierigkeit dabei ist die Bestimmung der Relevanz für eine beliebige Website. Die dafür genutzten Funktionen und Methoden werden allerdings von den Unternehmen geheim gehalten, um einen Missbrauch ihrer Suchmaschine zu verhindern. Dennoch sind die am häufigsten genutzten Merkmale bekannt und in einigen wissenschaftlichen Arbeiten untersucht worden.³ Moderne Suchmaschinen nutzen dutzende oder gar hunderte verschiedener

¹Vgl. Bookstein S. 1 [5]

²Vgl. Manning, Raghavan, Schütze [6]

³Vgl. Zaragoza, Najork, S. 1 [7]

Methoden und Features um die Relevanz der verfügbaren Suchergebnisse zu bewerten, wird im Folgenden nur auf einige bekannte Methoden eingegangen.

Bei der Webapplikation Crossload handelt es sich um eine Internetsuche, welche beispielsweise mit Google, Bing, Ecosia oder anderen Suchmaschinen üblich ist. Deren grundlegende untersuchte Funktionalität, die Suche und Bestimmung der Relevanz, ist vergleichbar. Unterschiedlich sind sie in der Art der Ergebnisse (Google sucht nach Webseiten, Crossload nach Predigten, Bilder, etc.) und der Menge an untersuchten Ergebnissen. Aus diesem Grund werden nachgehend Methoden zur Bewertung der Relevanz von Informationen aus Webseiten betrachtet.

2.2.1. Textuelle Relevanz

Das einfachste Merkmal für die Bewertung der Relevanz ist den kompletten Inhalt nach der textuellen Relevanz zu bewerten. Da natürliche Sprache, die meist für Suchergebnisse genutzt wird, generell ungenau ist, wird mit sogenannten „Matching Functions“ versucht auch ungefähre Übereinstimmungen in einem Fließtext zu finden. Einige der verwendeten Funktionen um die textuelle Relevanz zu bewerten sind dabei:⁴

- Die Anzahl der Treffer für den Suchterm oder Abwandlungen
- Position des Suchterms (früheres Vorkommen, z. B. im Titel)
- Seitenstruktur (für Webseiten: Ist der Term eine Überschrift o. ä.)
- Grafisches Layout (für Webseiten: Ist der Term z. B. farblich markiert)
- Levenshtein Distanz⁵ (die minimale Anzahl an Operationen, um eine Zeichenkette in eine andere umzuwandeln)

2.2.2. Relevanz durch Attribute

Des Weiteren ist es auch möglich den durchsuchten Objekten Attribute zuzuweisen, um für Schlagwörter relevantere Ergebnisse zu erlangen. Diese können entweder von Nutzern selbst bestimmt werden, wie z. B. bei der Website Flickr⁶, um Bilder für bestimmte Themen höher werten zu lassen oder werden von Algorithmen aufgrund von Bilderkennung automatisch zugewiesen.

⁴Vgl. Zaragoza, Najork, S. 1 [7]

⁵Vgl. Levensthein [8]

⁶Vgl. Liu et.al., S. 1-3 [9]

Ein Beispiel hierfür ist Google, welches eine frei verwendbare Machine Learning API⁷ oder eine direkte Integration, in die Google Fotos App anbietet. Diese Anwendung teilt automatisch die gemachten Bilder in verschiedene Kategorien ein.⁸

Möglich gefundene Ergebnisse können auch durch Existenz oder Nichtvorhandensein eines Attributs höher gewichtet werden. Dadurch können zum Beispiel bereits aufbereitete Ergebnisse eine höhere Relevanz erhalten.⁹

2.2.3. Hyperlink Relevanz

Für Suchergebnisse im Internet oder andere miteinander verlinkte Seiten, wie z. B. in internen Dokumentationsseiten, Wikis o. ä., können auch die Hyperlinks genutzt werden, um die Relevanz eines Ergebnisses zu bestimmen. Ein Hyperlink besteht hierbei aus dem angezeigten Text auf der Quellseite und einem Link zur Zielseite oder auf einen bestimmten Abschnitt dergleichen. Hyperlinks werden aber nicht von einer Maschine in eine Seite eingefügt, sondern jeder Link wird von Menschen gesetzt. Aus diesem Grund kann man hier von „menschlicher Intelligenz“ sprechen.¹⁰

Um einen Treffer höher zu gewichten, ist eine Option die Anzahl an Verlinkungen auf eine Seite zu zählen und absteigend zu sortieren.¹¹ Alternativ kann der angezeigte Linktext noch zusätzlich als eine Art Attribut (2.2.2) oder erweiterte textuelle Referenz (2.2.1) gesehen werden, der dann bei der Auswertung einer Suche mitverwendet wird.¹²

2.2.4. Relevanz durch Nutzerverhalten

Um unabhängiger von manuellem Verlinken zwischen Seiten zu werden, haben bekannte Suchmaschinen auch Möglichkeiten entwickelt, die Anzahl der „erfolgreich“ gefundenen Treffer zu zählen und als relevanter zu gewichten. Im Umfeld einer Internetsuche wäre der „erfolgreich“ gefundene Treffer ein Klick auf die entsprechende Website. Diese können entweder live oder durch Auswertung von Log Dateien analysiert werden. Andere Wege um die Anzahl an Besuchen auf einer Website zu messen, umfassen Tracking Methoden, Toolbars oder Werbung. Die Bewertung der Relevanz durch das untersuchte Nutzerverhalten ist überaus erfolgreich, da hier von einer Art Schwarmintelligenz ausgegangen wird, die Nutzern für die gleiche Suche Ergebnisse anzeigt, die schon viele Benutzer davor angeklickt haben.¹³

⁷Vgl. Google ML Dokumentation [10]

⁸Vgl. Google Fotos [11]

⁹Siehe Crossload Search API [12]

¹⁰Vgl. Zaragoza, Najork, S. 2 [7]

¹¹Vgl. Marchiori [13]

¹²Vgl. Page, Brin, Motwani und Winograd [14]

¹³Vgl. Joachims, Radlinski, S. 1 [15]

2.2.5. Performance

Da Suchmaschinen dem Nutzer eine bestmögliche Benutzererfahrung, auch bekannt als User Experience, ermöglichen wollen, sollen die gefundenen Webseiten dies bieten. Eine Möglichkeit dies zu messen, ist die Performance einer Website. Die Performance einer Webseite oder eines Suchergebnisses umfasst die Ladegeschwindigkeit, Speicherverbrauch und benötigte Leistung um das Ergebnis komplett anzuzeigen. Da dies nicht für x-Millionen Treffer bei jeder Suchanfrage getestet werden kann, werden mögliche Suchtreffer vorher indiziert und nach der Performance untersucht. Dadurch entsteht ein Performance-Score, welcher dann für die Relevanz verwendet werden kann.¹⁴

Im aktuellen Beispiel von Crossload, ist die Performance eines Suchergebnisses durch die Anzahl an gefundenen Treffern und die Anzahl an gefundenen Treffern mit einem bestimmten Attribut, wie z. B. „Performance: 5“ oder „Performance: 4“, definiert.

¹⁴Vgl. Manning, Raghavan, Schütze [6]

2.3. Auswertung der Relevanz von Suchergebnissen

Um letztendlich Ergebnisse mit der höchsten Relevanz zu erhalten, wird meist eine Kombination aus mehreren der o.g. Methoden benutzt, um die komplette Relevanz für einen Treffer zu bewerten. Die Herausforderung dabei ist die genaue Gewichtung der einzelnen Methoden, um die Relevanz eines Treffers optimal zu bewerten. Für jeden Treffer wird dann ein Relevanzscore berechnet, der sich aus den einzelnen Methoden zusammensetzt. Nach diesem Score wird in einer Liste absteigend sortiert, um das relevanteste Ergebnis als erstes Element zu erhalten.

Sollte sich der Score eines Treffers in der Relation zu anderen Wertungen weit absetzen, kann dieser Treffer dem Nutzer auch direkt vorgeschlagen werden.¹⁵ Dieser Vorschlag kann anschließend hervorgehoben werden, um dem Nutzer die Entscheidung zu erleichtern. Damit der Vorschlag für den Nutzer sichtbar ist, wird der Vorschlag in der Liste der Suchergebnisse grafisch hervorgehoben, oder es wird ein separater Abschnitt für den Vorschlag angezeigt. Dies wird von Suchmaschinen wie Google, Bing oder Ecosia bereits verwendet. Diese Suchmaschinen präsentieren dem Nutzer ein ideales Ergebnis meist an der Seite der restlichen gefundenen Treffer.¹⁶

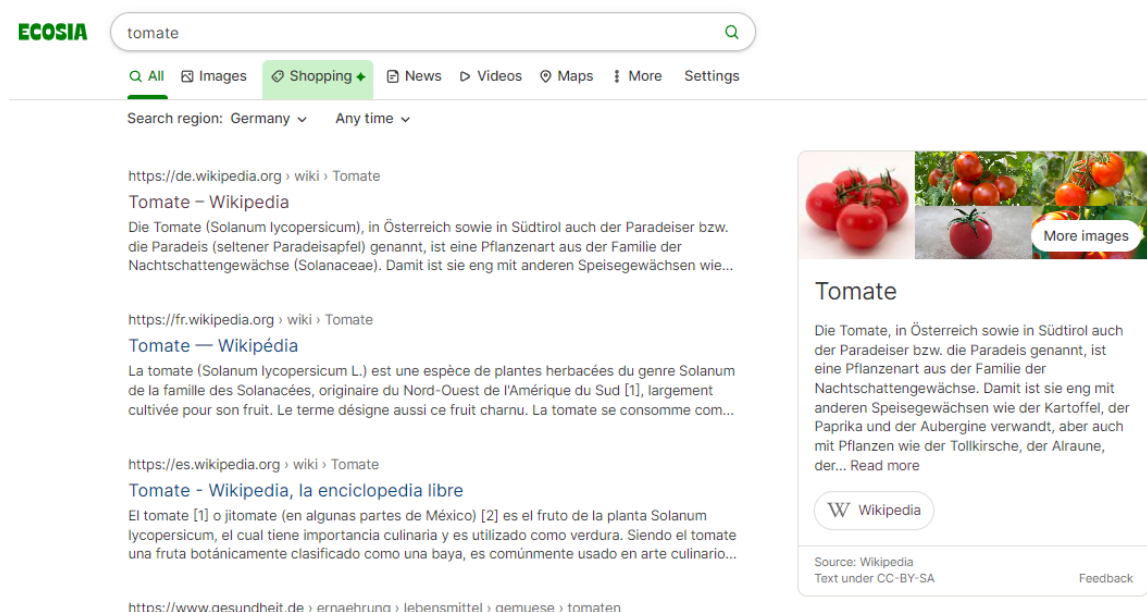


Abbildung (2.1) –Vorschlag eines Wikipedia Artikels zum Thema Tomate [19].

¹⁵Vgl. Turnbull, Berryman, S. 225-228 [16]

¹⁶Vgl. Google [17], Bing [18], Ecosia [19]

Dieses Vorschlagen von Ergebnissen kann auch verwendet werden, um bereits bei der Eingabe einer Abfrage durch Autovervollständigung (eng. „Search Completion“) geschehen.¹⁷ Der Unterschied ist hierbei der angezeigte Ort, an dem die Vorschläge erfolgen. Für den Suchvorschlag wird er direkt in der Suchleiste angezeigt, während der Vorschlag eines Treffers in der Ergebnisliste erfolgt.



Abbildung (2.2) –Vorschlag von Suchbegriffen zum Thema Tomate [19].

Für kleinere Anwendungen ist hierbei meist ein manuelles Einstellen nach einem Trial-and-Error Verfahren notwendig, bei denen einige wenige Methoden unterschiedlich gewichtet werden. Dies wird dann von Zeit zu Zeit wiederholt, wenn neue Erkenntnisse aus Tests oder dem produktiven Betrieb zurückkommen.¹⁸ Große Suchmaschinen nutzen hierfür allerdings, wie bereits erwähnt, hunderte Methoden und evaluieren deren Erfolg im produktiven Betrieb durch proprietäre statistische Methoden.¹⁹

¹⁷Vgl. Turnbull, Berryman, S. 206-218 [16]

¹⁸Vgl. Zaragoza, Najork, S. 3 [7]

¹⁹Vgl. Taylor, Zaragoza, Craswell, Robertson, Burges [20]

3. Technische Grundlagen

3.1. Apache Lucene

Apache Lucene, oder auch kurz Lucene genannt, ist eine mächtige Suchbibliothek, die plattformunabhängig von verschiedenen bekannten Apps, wie z. B. Netflix eingesetzt wird. Sie nutzt für die Indizierung zu durchsuchender Dokumente Textfelder, wie z. B. „title“ für den Titel eines Dokumentes, um sowohl den Inhalt als Volltext, als auch die Attribute durchsuchen zu können.¹ Lucene stellt dabei folgende Funktionen bereit:²

- **Sortierte Suche:** die besten Ergebnisse werden zuerst gezeigt
- **Leistungsstarke Queries:** Phrasenabfragen, Platzhalterabfragen („wildcards“), Abfragen zur Nähe bzw. Ähnlichkeit eines Suchergebnisses („proximity search“), exakte Phrasenabfragen, Bereichsabfragen für Datum/Uhrzeit und Zahlen
- **Feldbasierte Suche:** Alle oder bestimmte Felder durchsuchen
- **Boolesche Operatoren:** Beliebige Kombinationen zwischen Suchbegriffen (AND, OR, NOT) um einzelne Abfragen zu kombinieren.
- **Sortieren nach einem beliebigen Feld**

3.2. Apache Solr

Solr ist von Crossload verwendete Such Engine, die als Web Schnittstelle dient, um auf einer Datenmenge Suchanfragen mit Apache Lucene auszuwerten.³

Solr benutzt hier die Indexing Funktionen von Lucene, um in Echtzeit alle verfügbaren Dokumente zu indizieren, um bei einer Suche nur den Index durchsuchen zu müssen. Mit Apache Zookeeper wird dann eine API zur Verfügung gestellt, welche Synchronisierung, Namensregister und die Verteilung der Konfiguration bereitstellt. Inhalte werden anhand von Boostingmechanismen höher oder schlechter bewertet. Als Entwickler gibt man hierfür mögliche Textfelder an, auf denen Solr automatisch ein Textmatching anwendet.⁴ Die wichtigsten Features von Solr sind:⁵

- Volltextsuche
- Facettensuche (generierte Attribute)

¹Siehe Apache Lucene [21]

²Vgl. Schindler, Bräuer, Diepenbroek, S.4 [22]

³Siehe Apache Solr [23]

⁴Vgl. 2.2.1

⁵Siehe Apache Solr [23]

- Suchvorschläge und Autokorrektur
- Echtzeit Indizierung des Suchergebnisindizes
- Arbeiten mit Word und PDF Dokumenten
- Erstellung eigener Boostingmechanismen (z. B. mit Java)

4. Crossload

4.1. Einführung

Crossload ist eine deutsche Predigt Datenbank, deren Ziel es ist, mit modernen Technologien und einem ansprechendem User Interface (UI) den Zugang zu Predigten und anderem christlichen Material zu vereinfachen. Hierzu werden teils Predigten aus anderen Systemen importiert, teils Autoren angefragt, welche dann regelmäßig ihre eigenen Predigten selbstständig hochladen. Dadurch sind sowohl ältere Predigten, etwa von Martin Luther, als auch Predigten zu aktuellen Themen und Weltgeschehen verfügbar. Zudem gibt es Schnittstellen zu christlichen Verlagen oder Webseiten wie CLV¹ oder Evangelium 21^{2,3}. Auf Crossload gibt es derzeit Predigten mit und ohne Video, Bücher, Bilder, Musik, Hörbücher und andere bzw. noch nicht kategorisierte Inhalte.

4.2. Technische Architektur

Technisch ist Crossload wie folgt aufgestellt:

- **Frontend:** UI entwickelt mit Angular zum Durchsuchen der Datenbank und direktem Streaming der Predigten. Für die Analyse und Statistiken wird Matomo verwendet. Dieser Dienst verfolgt besuchte Seiten, angehörte Inhalte und abgegebene Suchanfragen. Das Open-Source Pendant zu Google Analytics enthält auch Statistiken zur durchschnittlichen Dauer eines Besuches.⁴
- **Backend/Suche:** Eine auf Solr basierte REST API mit allen veröffentlichten Inhalten und anderen Metadaten. Sie wird verwendet, um performante Suchanfragen des Frontends bearbeiten zu können und relevante Ergebnisse zu liefern.
- **Redaktion:** Aufbereitung und Anlegen von Inhalten verschiedenster Kategorien und anderer Metadaten.
 - **Angular UI:** Redaktionelles Backend zum Pflegen aller Daten von Crossload.
 - **Node.js RESTFUL API:** Schnittstelle zwischen der UI, der Datenbank und AWS.
 - **AWS:** Speicherung von Dateien (Audio, Video, Bilder).
 - **Mongo DB:** Datenbank zur Verwaltung und Speicherung aller Daten.

¹Siehe CLV [24]

²Siehe Evangelium 21 [25]

³Vgl. Pfeiderer, Crossload [2]

⁴Siehe Matomo [26]

4.3. Suche bei Crossload

Bei Crossload werden verschiedene Typen bzw. Kategorien von Inhalten in der von Solr indizierten Datenbank über eine Spring Boot Anwendung an das Webfrontend zur Verfügung gestellt. Diese Inhalte werden anhand von eingegebenen Suchbegriffen, aktuellen Ereignissen oder Attributen wie Themen, Autoren oder Bibelstellen gefiltert.

Falls der Nutzer keine Suchanfrage eingegeben hat, werden die neuesten bzw. die aktuell relevantesten Inhalte angezeigt. Diese Suche kann direkt über die Startseite erreicht werden, wenn man auf den Suchknopf (eine Lupe) klickt. Sollte der Nutzer in das Suchfeld etwas eingegeben haben, wird die Suche direkt nach dem Klick auf die Lupe gestartet und er erhält Ergebnisse abhängig von seiner Eingabe.

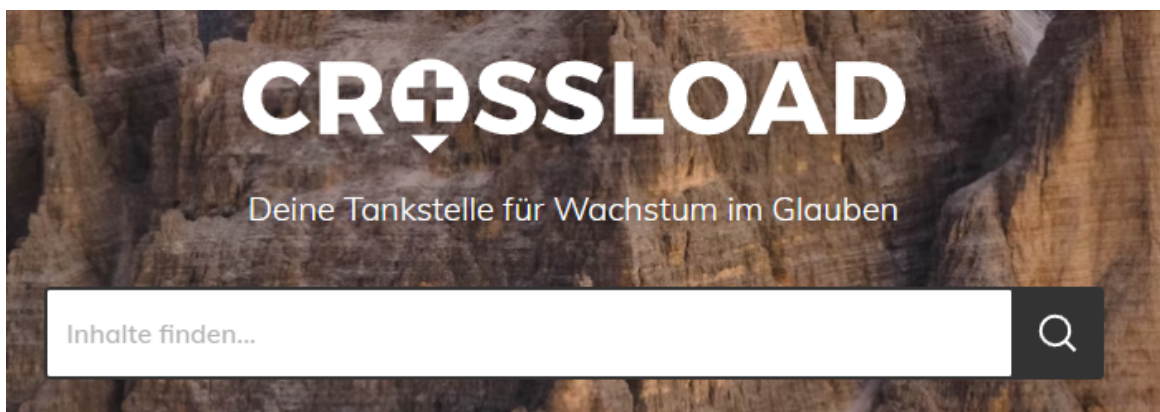


Abbildung (4.1) –Startseite der Suche bei Crossload [2].

Der initiale und derzeit implementierte Gedanke ist, die Inhalte auch in diesen Kategorien zu übertragen und in fester Reihenfolge anzuzeigen. Diese Suchergebnisseite enthält pro Kategorie 5 Inhalte und einen Link, um alle Ergebnisse dieser Kategorie anzuzeigen. Dadurch konnte bisher dem Nutzer ein schneller Einblick in die Inhalte gegeben werden, ohne dass er sich durch viele Ergebnisse klicken muss. Die Kategorien sind hierbei nach der erwarteten Häufigkeit sortiert, in welcher die Entwicklung das Verhalten des Nutzers erwartet. Ein Auszug dieser Suche ist im Anhang (A.3) sichtbar.

Diese Vorgehensweise hat jedoch einige Nachteile:

- **Relevanz:** Der möglicherweise relevanteste Inhalt wird nicht als Erstes angezeigt, da dessen Kategorie relativ weit unten angezeigt wird.
- **Übersicht:** Es ist schwer für den Nutzer eine Übersicht über alle gefundenen Inhalte zu erlangen.
- **User Experience:** Höchstwahrscheinliche Treffer (90-100 % Trefferwahrscheinlichkeit) werden nicht direkt vorgeschlagen. Ein Treffer ist hierbei sehr wahrscheinlich relevant, wenn er ein Treffer mit einem sehr hohen errechneten Relevanzscore ist.

Diese Nachteile sollen im Verlaufe der Entwicklung behoben werden. Ebenso sollen auch die bisher genutzten Methoden zur Berechnung der Relevanz verbessert werden. Diese umfassen derzeit:

- Text Matching auf verschiedene Textteile und Attribute. Hier werden verschiedene Attribute in 3 Kategorien (hoch, mittel, niedrig) wie folgend bewertet:
 - **Hoch:** Titel, Serie, Thema, Autor
 - **Mittel:** Untertitel, Schlagwörter, Kategorie, Thema
 - **Niedrig:** Verlag, Standort, Dateiname, Speech to Text, Mitschrift, Suchsnippet
- Matching des Suchterms zu einem Bibelvers.
- Inhalte mit Video oder Bild werden höher bewertet, da diese qualitativ hochwertiger sind.
- Filter für mitgegebene Query Parameter: Kategorie, Serie, Event, Thema, Jahreszahl oder Dauer. Inhalte, die nicht zu diesem Filter passen, werden komplett aussortiert.

5. Anforderungen und Problemanalyse

Im folgenden Kapitel sollen alle wesentlichen Funktionen des geplanten Prototyps dargestellt werden. Da es sich um ein kleines Projekt mit einem abgestecktem Rahmen handelt und es kein Team gibt, welches die Entwicklung durchführt, wird das Wasserfallmodell für diese Entwicklung verwendet.¹

5.1. Vorgehensweise

Zur Erfassung der Anforderungen bzw. Requirements Engineering werden User Stories benutzt. Diese sind bekannt aus agilen Softwareentwicklungsmodellen, wie z. B. Scrum, entstanden aber durch praktische Erfahrungen in der Softwareentwicklung. Konzeptioniert wurden sie von Dr. Ivar Jacobsen² und Ron Jeffries³. Mithilfe einfacher Sprache wird aus der Sicht des Stakeholders das Ziel einer Story in einem kurzen Satz zusammengefasst. Anschließend wird dieses Ziel begründet, um die Wichtigkeit und Existenzberechtigung der Story zu begründen. User Stories sind dabei auch Anforderungen nach dem SMART Prinzip⁴, da diese nur einen sehr kleinen abgesteckten Teilbereich einer Funktionalität enthalten. Dadurch sind sie einfacher schätzbar, umsetzbar und testbar. Anhand der ermittelten User Stories werden nach der Entwicklung Akzeptanztests durchgeführt, um den Erfolg des Endproduktes objektiv zu bewerten.

5.2. User Stories

Die Anforderungen umfassen alle Aktionen, welche der Nutzer in der Anwendung durchführen will. Ziel aller Anforderungen ist die übergreifende Suche über mehrere Kategorien effizient anhand mehrerer Kriterien zu durchsuchen und zu bewerten. Anhand dieses Ziels werden User Stories entwickelt, in denen ein Nutzer und andere Personen ihre Anforderungen an das zu entwickelnde Produkt stellen.

Nichtfunktionale Anforderungen werden bei dieser Anforderungserhebung nicht beachtet, da es sich hierbei um die Erweiterung einer bestehenden API handelt und Aspekte wie Benutzerfreundlichkeit und User Experience hierbei wenig relevant sind, bzw. das Entwicklungsumfeld durch die bereits bestehende Anwendung vorgegeben ist. Die Priorität der einzelnen User Stories ergibt sich aus der unten gegebenen Reihenfolge.

¹Vgl. 1.3

²Vgl. Jacobson, Spence, Kerr 2016 [27]

³Vgl. Ron Jeffries [28]

⁴Vgl. Witte 2019a, S. 67 [3]

Ich, als Benutzer, will ...

1. ... für ein gegebenes Suchkriterium relevante Suchergebnisse über mehrere Kategorien hinweg erhalten, damit mit einer einzelnen Suche nur eine geringe Teilmenge der Datenbank angezeigt wird.
2. ... die erhaltenen Suchbegriffe nach Kontext und Wahrscheinlichkeit gewichtet erhalten, damit diese im späteren Verlauf sortiert werden können. Der Kontext ergibt sich aus möglichen Schlagworten, die im Suchbegriff verwendet worden, womit z. B. eine Kategorie, ein Attribut eines Ergebnisses oder höher gewertet wird. Beispiele wären:
 - a) ... der Titel eines Buches wird „relativ“ genau als Suchbegriff eingegeben, folglich wird dieses Buch stärker gewichtet.
 - b) ... der Suchbegriff enthält den Term „Video“, folglich werden alle Videos priorisiert.
3. ... die erhaltenen Suchbegriffe anhand des errechneten Gewichts absteigend sortiert zurückgegeben bekommen, damit das relevanteste Suchergebnis auf der Suchergebnisseite ganz oben steht.
4. ... ein ideales Suchergebnis als grafisch hervorgehobenen Vorschlag angezeigt bekommen, damit auf der Suchergebnisseite eine schnelle Navigation zu diesem Ergebnis möglich ist. Dabei soll ein Inhalt nur vorgeschlagen werden, wenn dessen Relevanz um einiges höher ist, als das der anderen gefundenen Ergebnisse. Dies verhindert, dass von ähnlich relevanten Suchergebnissen eines ohne Berechtigung hervorgehoben wird, auch wenn es das Relevanteste in dieser Liste ist.⁵

Die User Stories 2.a und 3 sind in der Anwendung bereits vorhanden und sind in der aufgeführten Liste nur zur besseren Übersichtlichkeit aufgeführt. Sie sollen nach der Entwicklung der neuen API gleich bleiben und nicht verändert werden.

⁵Siehe 2.1

6. Konzeption

Bevor mit der Entwicklung des Prototyps gestartet werden kann, geht die Planung und Konzeption der Erweiterung voraus. Der Entwurf einer Software ist die Basis für jede Softwareentwicklung und sollte vor der Implementierung erfolgen. Anfangs wird die momentane Anwendung auf bereits implementierte Funktionalität überprüft und schließlich mithilfe der erarbeiteten Methoden zur Bewertung der Relevanz auf Grundlage der gesammelten Anforderungen verbessert.

Grundsätzlich findet die Anwendung bereits passende bzw. relevante Inhalte durch das Matching der verschiedenen Attribute¹ und Textteile.² Ebenso das Matching bezüglich des Bibelverses oder des initialen Boosting über ein vorhandenes Video oder Bild führt bereits zum gewünschten Ergebnis und eine Änderung würde hier keinen nennenswerten Mehrwert bieten.

Dennoch gibt es einige Ideen, relevantere Inhalte für den Nutzer herauszugeben: gemischte Ergebnislisten, der Zuordnung des Suchbegriffes zu einer Kategorie auf die Kategorien und Vorschläge zur weiteren Navigation.

6.1. Gemischte Ergebnisliste

Als oberstes Ziel wird die Liste der gefundenen Inhalte, momentan aufgespalten in die verschiedenen Kategorien wie z. B. Bild, Video, Predigt, Buch, etc., in eine große Liste überführt. Dadurch können relevante Inhalte, die bisher durch die vordefinierte Sortierung der Kategorien auf der Suchseite nicht als Erste aufgelistet wurden, an der Stelle angezeigt werden, an welcher der Nutzer sie erwartet.

Damit der Nutzer dennoch die jeweilige Kategorie der Suchergebnisse sieht, wird anschließend zu der ausführlichen Version des Ergebnisses ein kleines Symbol mit dessen Kategorie hinzugefügt. So geht die bisherige Funktionalität nicht komplett verloren und der Nutzer erhält die relevantesten Inhalte direkt an erster Stelle und sieht sofort dessen Kategorie.

6.2. Zuordnung des Suchbegriffes zu einer Kategorie

Eine Möglichkeit, die Relevanz der Suchergebnisse im Sinne der Aufgabenstellung zu verbessern, wäre es, anhand des Suchbegriffes herauszufiltern, ob z. B. ein Schlagwort wie „Video“ oder „Bild“ verwendet wurde und somit Inhalte dieser Kategorie relevanter sind.

¹Siehe 2.2.2

²Siehe 2.2.1

Dafür müssten relevante Schlagwörter ermittelt werden und auch in allen möglichen Varianten untersucht werden, um ein hilfreiches Matching zu erhalten, welches dann anhand dem Attribut „Hauptkategorie“ nachvollzogen werden kann. Eine gewisse Menge an Varianten kann vordefiniert werden, um einen Großteil der Anfragen korrekt abzufangen. Um letztendlich aber eine mehr und mehr vollständige Menge an Varianten und Suchbegriffen zu erhalten müssen die abgegebenen Suchabfragen untersucht werden. Diese können aber mit Matomo untersucht werden und mit der Zeit angepasst werden.³

Für den Start wären folgende Schlagwörter für die vorhandenen Kategorien denkbar:⁴

- **Predigten (mit Video):** Video, Film, Stream, Live.
- **Predigten (mit und ohne Video):** Predigt, Vortrag, Mahnwort.
- **Bücher:** Buch, Bücher, Taschenbuch, Sammelband, Reader, Druck, Bestseller.
- **Bilder:** Bild, Darstellung, Zeichnung, Aufnahme, Foto, Fotografie.
- **Musik:** Song, Melodie, Hymne, Stück, Gesang, Klavier, Musik, Orchester.
- **Hörbücher:** Hörbuch, Hörbücher, Audiobook.
- **Sonstige:** Sonstige, andere.

6.3. Vorschläge für weitere Navigation

Optimalerweise gibt der Nutzer eine Suchanfrage ein, zu der ein Inhalt eine sehr hohe Relevanz hat und alle anderen Inhalte eine recht niedrige. Sollte dies der Fall sein, so könnte dieser Inhalt in einer Vorschlagsbox über der Ergebnisliste angezeigt werden, damit der Nutzer visuell sieht, dass dies der Inhalt ist, den er höchstwahrscheinlich sucht. Eine Berechnung hierfür ist nicht klar definiert, als ersten Versuch wird überprüft, ob der berechnete Score des relevantesten Inhalts mindestens doppelt so groß ist, wie der des nächsten Inhalts. Dieses Vorgehen, die Annahme, dass relevante Inhalte mindestens den doppelten Relevanzscore haben um vorgeschlagen zu werden, muss aber in der Entwicklung und im produktiven Betrieb weiter geprüft und verfeinert werden.

Eine mögliche Schwachstelle hierbei könnten sehr relevante Inhalte direkt am Anfang sein, bei denen die darauffolgenden Inhalte im Vergleich irrelevant sind. Somit könnten auch beide Inhalte vorgeschlagen werden, was aber aus Gründen der Nutzerfreundlichkeit nur auf Einen minimiert wird. Dafür müsste die ganze Liste oder ein Teil, z. B. die Top 10, auf

³Siehe 4.1 [26]

⁴Siehe Duden [29]

die durchschnittliche Relevanz geprüft werden und Inhalte, die sich stark (ebenfalls doppelt so groß) nach oben von diesem Durchschnitt unterscheiden, als Vorschläge genommen werden.

Beispielsweise könnte eine Liste mit den folgenden Scores vorliegen:

- **Treffer 1:** Score 92.
- **Treffer 2:** Score 91.
- **Treffer 3:** Score ≤ 10 .
- **Treffer 4:** Score ≤ 10 .
- **Treffer 6:** Score ≤ 10 .
- **Treffer 5:** Score ≤ 10 .
- **Treffer 7:** Score ≤ 10 .
- **Treffer 8:** Score ≤ 10 .
- **Treffer 9:** Score ≤ 10 .

Wenn nur Treffer 1 und 2 verglichen werden, könnte man zu dem Schluss kommen, dass keiner der gefundenen Inhalte vorgeschlagen wird, da der Score von Treffer 1 nicht doppelt so groß ist, wie der von Treffer 2. Doch wenn man sich den Durchschnittswert der restlichen Inhalte anschaut und der Score von Treffer 1 mindestens doppelt so groß ist, wie dieser Durchschnittswert, wird Treffer 1 als Vorschlag angezeigt.

Für die Implementierung wird hierbei zuerst geprüft, ob ein Inhalt auf Basis seiner direkten Nachfolger vorgeschlagen werden kann. Falls dies nicht zutrifft, wird anschließend mit der Prüfung des Durchschnitts fortgefahren.

7. Entwicklung des Prototyps

Die Umsetzung des Prototyps erfolgt in mehreren Schritten. Zu Beginn wird wie in 6.1 beschrieben, die Liste aller Ergebnisse zusammengeführt und absteigend nach der Relevanz sortiert. Anschließend wird die Zuordnung des Suchbegriffes zu einer Kategorie¹ implementiert, indem die Liste der möglichen Synonyme mit dem Suchterm abgeglichen wird und die Ergebnisse geboostet werden, wenn der Suchterm ein Synonym enthält. Zuletzt werden die resultierenden Inhalte nach einem möglichen Vorschlag, beschrieben in 6.3, dem Ergebnis hinzugefügt.

7.1. Gemischte Ergebnisliste

Für die Entwicklung der gemischten Ergebnisliste im Crossload Frontend ist wichtig, dass die schon implementierte Funktionalität der Suche nach Kategorien nach der Entwicklung immer noch möglich sein muss. Dazu finden sich auf der Suchergebnisseite mehrere Tabs, bei denen auch eine Suche innerhalb von Kategorien möglich ist. Sie wird von der gemischten Suchergebnisseite angeführt, auf der wir momentan die Aufteilung in verschiedene Kategorien sehen.²

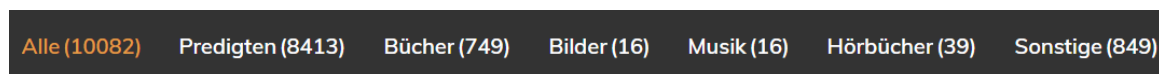


Abbildung (7.1) –Leiste der Suchergebnisse in den verschiedenen Kategorien [2].

Ziel ist es, die gemischte Suchleiste so zu ändern, dass eine große Liste mit allen Kategorien gemischt angezeigt wird. Anstatt nach Ergebnissen pro Kategorie abzufragen und diese in den einzelnen Sektionen darzustellen, muss eine gesammelte Anfrage an die Such API gesendet werden. Diese wird nach dem gegebenen Sortierkriterium, in diesem Fall absteigend nach Relevanzscore sortiert von der Such API zurückgegeben. Dafür wird zu den vorhandenen Kategorien eine „gemischte“ Kategorie hinzugefügt, bei welcher dann kein Suchfilter mitgegeben wird. Danach sind nur noch kleine Änderungen notwendig, um das für den Nutzer schön dargestellt zu bekommen.

¹Siehe 6.2

²Siehe Anhang A.3

```
1 // Create a new category for the unified / mixes list
2 export const MIXED_RESULTS_OPTION = {
3   category: "mixed",
4   path: RESULTS_BASE_URL + "/gemischt",
5   lastPathSegment: "gemischt",
6   text: "Alle",
7   longText: "Alle",
8 } as const
```

Erstellen der gemischten Kategorie [30]

```
1 // Remove category from params, when search category is mixed
2 if (category === "mixed") {
3   delete url.category
4 }
```

Löschen der Kategorie aus den API Parametern [30]

Das Ergebnis zeigt eine zusammengeführte Liste, die komplett nach Relevanz sortiert ist.³ Mithilfe der Symbole, die bereits bei den Ergebnissen die Kategorie bereits kennzeichneten, kann der Nutzer einfach erkennen, welcher Inhalt von welcher Kategorie ist.

7.2. Zuordnung des Suchbegriffes zu einer Kategorie

Für die Zuordnung des Suchbegriffes zu einer Kategorie werden die Synonyme in Java konfiguriert. Dazu wird eine Enumeration erstellt, die eine Liste der Synonyme enthält.⁴ Durch diese Enumeration können Änderungen leicht eingebaut werden und durch die vorhandene CI/CD Pipeline direkt in die produktive Umgebung eingespielt werden. Somit wäre der Nachteil der fest definierten Synonyme ausgeglichen, da Korrekturen schnell eingespielt werden können.

Eine Alternative zur Konfiguration in Java wäre eine ausgelagerte JSON Datei. Da diese aber genau wie die Java Enumeration, durch die CI/CD Pipeline, in die produktive Umgebung gelangen würde, wäre kein großer Nutzen dabei, eine JSON Datei der Enumeration vorzuziehen. Der Unterschied dabei ist aber die Komplexität beim Einlesen der Datei und die fehlende Typisierung in der JSON Datei. Aus diesem Grund wurde sich für die Enumeration entschieden.

Der nächste Schritt ist der Vergleich des Suchterms mit den verfügbaren Synonymen. Dazu wird überprüft, ob dieser ein Synonym vollständig enthält. Damit sind auch Sonderfälle wie der Plural oder Beugungen enthalten, da für mögliche Spezialfälle bereits Synonyme in die Listen eingefügt wurden.

³Siehe Anhang A.4

⁴Siehe II

Für die Überprüfung werden Java Streams benutzt. Diese machen den Source Code lesbarer und kleiner, da viele Kontrollstrukturen wegfallen. Mithilfe der vorhandenen Klasse CrossloadCriteriaBuilder werden alle Inhalte der gefundenen Kategorie geboostet. Für Predigten wird hierbei noch für den Spezialfall der Predigt mit Video unterschieden, bei nur Predigten mit Video geboostet werden und andere Predigten keine besondere Behandlung erfahren.

```

1  private void addCategoryBoost(String searchTerm, SolrQuery query) {
2      // Check if search term is a sermon category
3      SermonCategory.getAllCategories()
4          .stream()
5          // Check if search term contains a category
6          .filter(category -> category.hasCategory(searchTerm))
7          .forEach(category -> {
8              CrossloadCriteriaBuilder instance = CrossloadCriteriaBuilder.
9              getInstance();
10             if(SermonCategory.VIDEO.equals(category)) {
11                 // Boost for Sermons with video
12                 instance.addCriteria(SchemaField.HAS_PRIMARY_VIDEO, "true", 75);
13             }
14             else {
15                 // Boost of found category
16                 instance.addCriteria(SchemaField.MAIN_CAT, category.getId(), 75);
17             }
18             // Add boost to query
19             query.addCriteria(instance);
20         });
21     }

```

Code für Überprüfung und Boosten der Kategorien. [\[31\]](#)

Durch das hier verfügbare Skript werden nun die Kategorien anhand des gegebenen Suchterms geboostet.

7.3. Vorschläge für weitere Navigation

Für die Vorschläge sind Anpassungen sowohl in Suche, um den Vorschlag herauszuarbeiten, als auch im Frontend, um den Vorschlag anzuzeigen, notwendig.

In der Such API wird hierfür zunächst die zurückgegebene Antwort um ein Vorschlagsobjekt (ßuggestion") erweitert.⁵ In diesem Objekt wird für das Frontend entweder die vorgeschlagene Predigt stehen oder kein Wert, wenn es keinen Vorschlag gibt.

Um einen Vorschlag zu finden, müssen die relevantesten Inhalte nach den bereits erwähnten Kriterien untersucht werden.⁶ Im derzeitigen Code wird immer nur eine Suchseite mit bis zu

⁵Siehe II

⁶Siehe 6.3

20 Ergebnissen zurückgegeben. Wenn diese benutzt werden, um das relevanteste Ergebnis zu finden, würde man maximal einen Vorschlag pro Seite bekommen. Dieser wäre von Seite pro Seite unterschiedlich.

Mit Solr ist es allerdings möglich, eine Query mit einem Ergebnis und den gleichen Suchparametern abzugeben, die nur die ersten 10 Elemente einer nach dem Relevanzscore absteigend sortierten Liste zurückgibt.

```

1  final JsonRequest suggestionQuery = new JsonRequest(params)
2      // Set query
3      .withParam("q", solrQuery.getQuery())
4      // Set filter queries
5      .withParam("fl", "*", [child limit=100])
6      // Get only 10 values
7      .withParam("rows", 10)
8      // Start at the beginning, always
9      .withParam("start", 0)
10     // Sort by relevance score
11     .withParam("sort", "score desc,main_id desc");

```

Solr Query für die Top 10 Ergebnisse, absteigend sortiert nach Relevanz [31]

Diese gefundenen Ergebnisse werden dann untersucht, um einen Vorschlag zu finden. Das Ergebnis (ein Inhalt oder *null*) wird dann dem zurückgegebenen Ergebnis mitgegeben.

```

1  public Sermon parseSuggestion(List<Sermon> mostRelevant) {
2      // Case 0: If there is only one content in the list, return that as
3      // suggestion
4      boolean onlyOneElement = mostRelevant.size() == 1;
5
6      // Case 1: Score of the first element is at least twice as much of the
7      // second score
8      boolean biggerThanDoubleTheScore = mostRelevant.size() >= 2 &&
9      mostRelevant.get(0).getScore() >= 2 * mostRelevant.get(1).getScore();
10
11     // return the first element if one of the cases is true
12     if(onlyOneElement || biggerThanDoubleTheScore) {
13         return mostRelevant.get(0);
14     }
15
16     // Case 2: Get the average of the most relevant sermons and check if the
17     // first has at least twice as much of that
18     if(!mostRelevant.isEmpty()) {
19         // Get average of the most relevant sermons
20         Float average = (float) mostRelevant
21             .stream()
22             .mapToDouble(sermon -> sermon.getScore())
23             .summaryStatistics()
24             .getAverage();
25     }
26 }

```

```
22     // return the first element if the first element has at least twice as
    much of the average
23     if(mostRelevant.get(0).getScore() >= 2 * average) {
24         return mostRelevant.get(0);
25     }
26 }
27
28 return null;
29 }
```

Relevanteste Inhalte nach einem Vorschlag durchsuchen [\[31\]](#)

Im Frontend müssen nun ebenso die Änderungen in den verschiedenen Typen gemacht werden, damit die Vorschläge auch ausgelesen werden können. Für die Anzeige wird dann die bereits verfügbare Liste der gefundenen Inhalte verwendet, um doppelten Code zu minimieren und den Vorschlag direkt anzuzeigen.

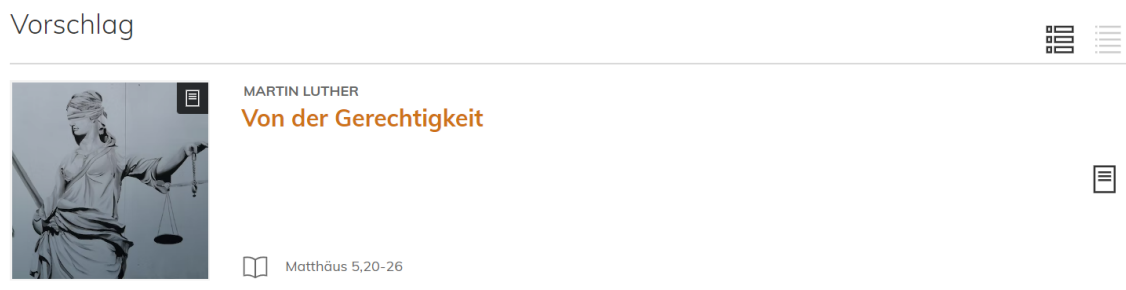


Abbildung (7.2) –Anzeige des Vorschlags [2].

```

1 // If there is a suggestion, show it
2 <div role="rowgroup" attr.data-cy="category_suggestion" *ngIf="response.
  suggestion">
3   <app-content-list-header text="Vorschlag"></app-content-list-header>
4
5   <app-content-list
6     [contents]="[response.suggestion]"
7     [maxVisible]="maxCategoryResults"
8   ></app-content-list>
9 </div>

```

Code für Anzeige des Vorschlage [30]

8. Auswertung

Ziel des Prototyps war es, für die Website Crossload relevantere Inhalte in der Suche herauszufiltern und anzuzeigen. Im Verlauf dieses Kapitels werden die konzeptionierten und entwickelten Ergebnisse anhand dieses Ziels genauer untersucht.

Anhand des Prototyps und den erhobenen funktionalen Anforderungen kann festgestellt werden, inwiefern diese Funktionalitäten für den Nutzer möglich sind. Für den Abgleich dieser Funktionalitäten wird für jede Anforderung ein kurzer Titel gegeben, der Status, ob diese erfüllt wurde oder nicht, sowie eine Begründung oder Zwischenstand der Bearbeitung.

Anforderung	Status	Begründung
1. Gemischte Suchergebnisse über alle Kategorien	Erledigt	Suchergebnisse werden zusammengeführt auf der Suchergebnisseite angezeigt.
2.a Relevanz für ähnliche Schlagwörter	Gegeben	Ähnliche Suchtitel werden bereits durch *-Zeichen in der gegebenen Suche gefunden (Vgl. 7.3).
2.b Relevanz anhand von Kategorien	Erledigt	Für eingegebene Kategorien werden ähnliche Inhalte um ein vielfaches geboostet.
3. Nach Relevanz absteigende Sortierung	Gegeben	Sortierung und Richtung bereits auf Suchergebnisseite gegeben.
4. Vorschläge	Erledigt	Der relevanteste Vorschlag wird von der Suche herausgefiltert und falls gefunden, auf der Suchergebnisseite angezeigt.

Tabelle (8.1) –Anforderungsanalyse

Der endgültige Stand aller funktionalen Anforderung kann in folgenden Kennzahlen zusammengefasst werden:

- 3 von 5 Anforderungen wurden erfüllt.
- 2 von 5 Anforderungen waren bereits gegeben und wurden nicht verändert.

9. Fazit

Zusammenfassend lässt sich sagen, dass die Entwicklung des Prototyps erfolgreich war und nach einem Review durch andere Entwickler auch auf Crossload live geschaltet werden kann. Alle gestellten Anforderungen wurden umgesetzt und sind erfolgreich lokal getestet worden. Durch Betrachtung der theoretischen Grundlagen der Relevanz und der vorhandenen Dokumentation von z. B. Solr konnten bereits vorhandene Erkenntnisse in die Entwicklung einfließen.

Durch diesen Prototyp konnte bereits existierende Funktionalität, die Suche nach Inhalten erweitert werden. Außerdem wird dem Nutzer ein Mehrwert in Form von relevanteren Inhalten auf der Suchergebnisseite geboten. Durch eine Wiederverwendung von Komponenten im Frontend wurde die Übersichtlichkeit und Wartbarkeit der Anwendung nicht verletzt.

Die größten Schwierigkeiten bzw. Aufwände dieser Entwicklung lagen in der theoretischen Ausarbeitung der Grundlagen der Relevanz, um geeignete Methoden zu finden, mit welchen die existierende Relevanz verbessert werden konnte. Außerdem mussten geeignete Punkte in der Implementation der Suche und im Frontend gefunden werden, um Änderungen vorzunehmen, ohne bereits funktionsfähige Programmteile einzuschränken.

Der entwickelte Prototyp könnte noch durch automatisierte Testfälle verbessert werden, welche aber derzeit in der Such API und im Webfrontend nur minimal vorliegen. Der Grund hierfür ist die begrenzte Zeit, die die ehrenamtlichen Entwickler in das Projekt einbringen können. Statt automatisierte Testfälle auszuarbeiten, werden derzeit neue Funktionen höher priorisiert.

Dieser Prototyp ist hierbei in der Suche von Crossload nur ein kleiner Teil einer ganzen Kette von Verbesserungen und Änderungen, die vorgenommen werden, um den Nutzern relevantere Inhalte zu bieten.

A. Anhang

I. Bilder

I. Einleitung

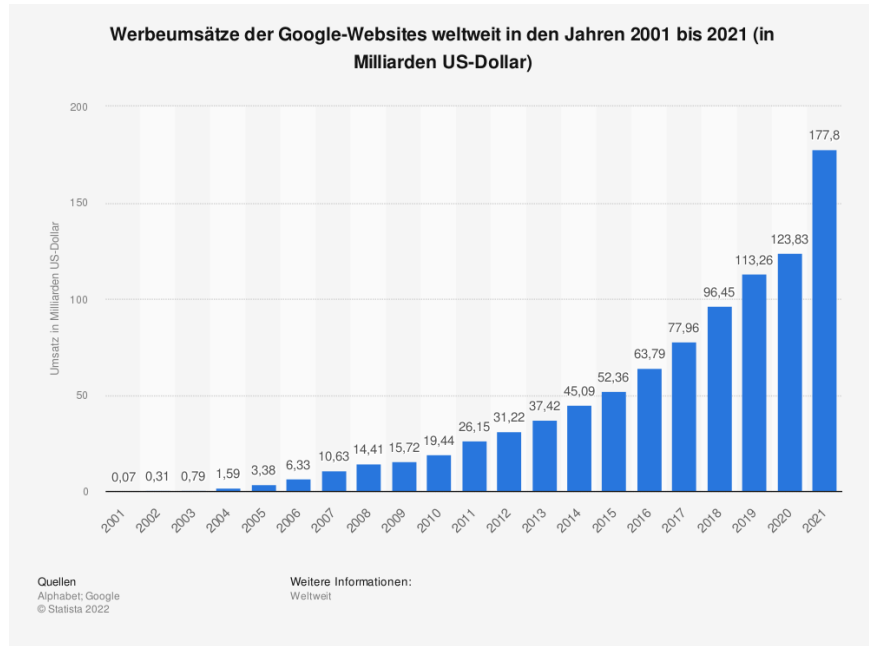


Abbildung (A.1) –Werbeumsätze Google Websites [32]

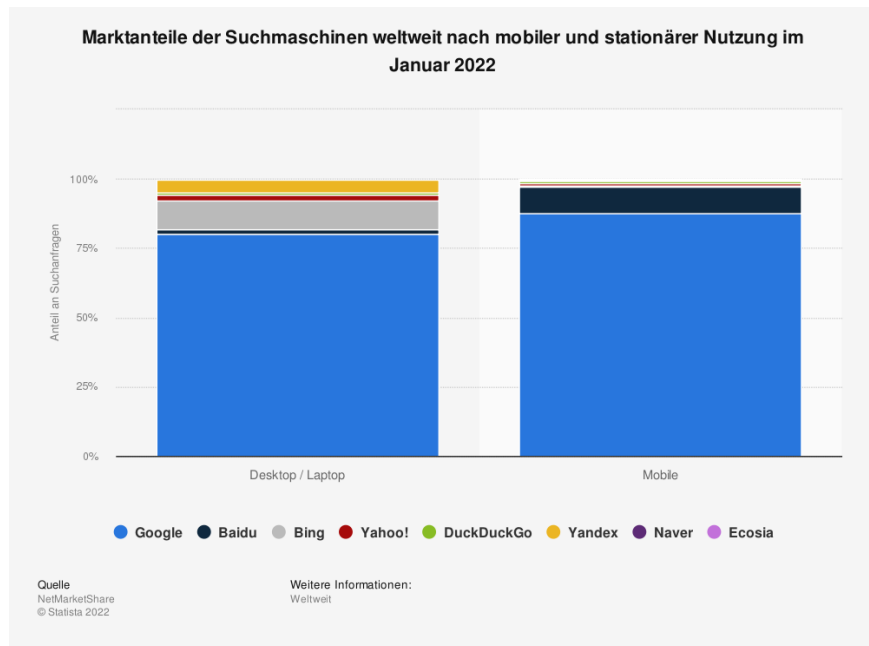


Abbildung (A.2) –Marktanteil Google [33]

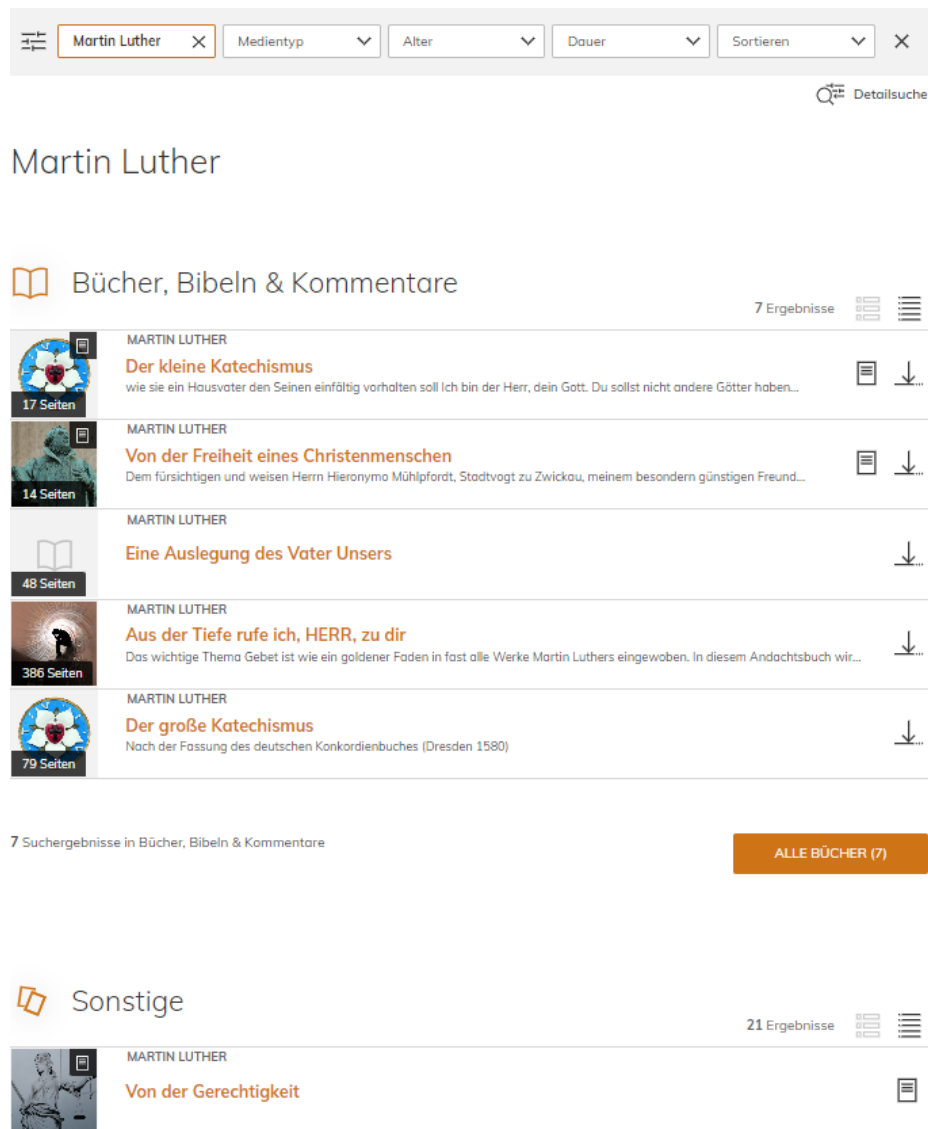


Abbildung (A.3) – Aktuelle Crossload Suche [2]

II. Entwicklung

Martin Luther

Q Alle

28 Ergebnisse














	<p>MARTIN LUTHER Von der Gerechtigkeit</p> <p>Matthäus 5,20-26</p>	
	<p>MARTIN LUTHER Eine einfältige Weise zu beten, für einen guten Freund</p> <p>Lukas 11,9-13 01.01.1535</p>	
	<p>MARTIN LUTHER Von den guten Werken</p> <p>Matthäus 19,17 29.03.1520</p>	
	<p>MARTIN LUTHER Von weltlicher Obrigkeit und Wieweit man ihr Gehorsam schuldig sei Luthers Zwei-Reiche-Lehre</p> <p>Matthäus 5,30-39 01.01.1523</p>	
	<p>MARTIN LUTHER Ein kleiner Unterricht, was man in den Evangelien suchen und erwarten solle</p> <p>Römer 1,1-4 01.01.1522</p>	
	<p>MARTIN LUTHER Vom ehelichen Leben</p> <p>1. Mose 1,27-31 01.01.1522</p>	

Abbildung (A.4) –Überarbeitete Crossload Suche [2]

II. Source Code

```

1  package org.biblepool.search.dto.sermon;
2
3  import java.util.ArrayList;
4  import java.util.Arrays;
5  import java.util.List;
6  import java.util.Objects;
7  import java.util.regex.Pattern;
8  import java.util.stream.Collectors;
9
10 public enum SermonCategory {
11     VIDEO("sermon", Arrays.asList("Video", "Film", "Stream", "Live"), true),
12     SERMON("sermon", Arrays.asList("Predigt", "Vortrag", "Mahnwort"), false)
13     ,
14     BOOK("book", Arrays.asList("Buch", "Bücher", "Taschenbuch", "Sammelband"
15     , "Reader", "Druck", "Bestseller"), false),
16     PICTURE("picture", Arrays.asList("Bild", "Darstellung", "Zeichnung", "
17     Aufnahme", "Foto", "Fotografie"), false),
18     MUSIC("music", Arrays.asList("Song", "Melodie", "Hymne", "Stück", "
19     Gesang", "Klavier", "Musik", "Orchester"), false),
20     AUDIOBOOK("audio", Arrays.asList("Hörbuch", "Hörbücher", "Audiobook"),
21     false),
22     OTHER("other", Arrays.asList("Sonstige", "Andere"), false);
23
24     private String id;
25     private Boolean hasVideo;
26     private List<String> synonyms = new ArrayList<>();
27     private Pattern pattern;
28
29     public static List<SermonCategory> getAllCategories() {
30         return Arrays.asList(SermonCategory.values());
31     }
32
33     private SermonCategory(String id, List<String> synonyms, Boolean
34     hasVideo) {
35         setId(id);
36         setSynonyms(synonyms);
37         setHasVideo(hasVideo);
38     }
39
40     public List<String> getSynonyms() {
41         return synonyms;
42     }
43
44     public void setSynonyms(List<String> synonyms) {
45         if(Objects.isNull(getSynonyms())) {
46             this.synonyms = new ArrayList<>();
47         }
48         else {

```

```
43     this.synonyms = synonyms;
44 }
45 }
46
47 public String getId() {
48     return id;
49 }
50
51 public void setId(String id) {
52     this.id = id;
53 }
54
55 public Boolean hasVideo() {
56     return hasVideo;
57 }
58
59 public void setHasVideo(Boolean hasVideo) {
60     this.hasVideo = hasVideo;
61 }
62
63 public void setPattern() {
64     String regex = synonyms.stream().collect(Collectors.joining("|"));
65     pattern = Pattern.compile("(" + regex + ")", Pattern.CASE_INSENSITIVE)
66 ;
67 }
68
69 public boolean hasCategory(String term) {
70     return pattern.matcher(term).matches();
71 }
```

Enumeration der verschiedenen Kategorien [30]

```

1  package org.biblepool.search.business.paging;
2
3  import org.biblepool.schema.api.sermon.dto.Sermon;
4
5  import java.util.List;
6  public class Page {
7
8      private Meta meta;
9      private List<Sermon> content;
10     private Sermon suggestion;
11
12     public Sermon parseSuggestion(List<Sermon> mostRelevant) {
13         // Case 0: If there is only one content in the list, return that as
14         suggestion
15         boolean onlyOneElement = mostRelevant.size() == 1;
16
17         // Case 1: Score of the first element is at least twice as much of the
18         second score
19         boolean biggerThanDoubleTheScore = mostRelevant.size() > 2 &&
20         mostRelevant.get(0).getScore() >= 2 * mostRelevant.get(1).getScore();
21
22         if(onlyOneElement || biggerThanDoubleTheScore) {
23             return mostRelevant.get(0);
24         }
25
26         // Case 2: Get the average of the most relevant sermons and check if
27         the first has at least twice as much of that
28         if(!mostRelevant.isEmpty()) {
29             Float average = (float) mostRelevant
30                 .stream()
31                 .mapToDouble(sermon -> sermon.getScore())
32                 .summaryStatistics()
33                 .getAverage();
34
35             if(mostRelevant.get(0).getScore() >= 2 * average) {
36                 return mostRelevant.get(0);
37             }
38         }
39
40         return null;
41     }
42
43     public Meta getMeta() {
44         return meta;
45     }
46
47     public void setMeta(Meta meta) {
48         this.meta = meta;
49     }
50 }

```



```
46
47     public List<Sermon> getContent() {
48         return content;
49     }
50
51     public void setContent(List<Sermon> content) {
52         this.content = content;
53     }
54
55     public Sermon getSuggestion() {
56         return suggestion;
57     }
58
59     public void setSuggestion(Sermon suggestion) {
60         this.suggestion = suggestion;
61     }
62 }
```

Vorschlag in der zurückgegebenen Antwort [\[31\]](#)

Abbildungsverzeichnis

1.1. Grundform eines Wasserfallmodells ohne Machbarkeitsanalyse. Nach Goll [4].	2
2.1. Vorschlag eines Wikipedia Artikels zum Thema Tomate [19].	8
2.2. Vorschlag von Suchbegriffen zum Thema Tomate [19].	9
4.1. Startseite der Suche bei Crossload [2].	13
7.1. Leiste der Suchergebnisse in den verschiedenen Kategorien [2].	20
7.2. Anzeige des Vorschlags [2].	25
A.1. Werbeumsätze Google Websites [32]	A
A.2. Marktanteil Google [33]	A
A.3. Aktuelle Crossload Suche [2]	B
A.4. Überarbeitete Crossload Suche [2]	C

Tabellenverzeichnis

8.1. Anforderungsanalyse 26

Literaturverzeichnis

- [1] Duden, “Duden | Wie schreibt man „googeln“? | Rechtschreibung.” <https://www.duden.de/rechtschreibung/googeln>, 2022.
- [2] D. Pfeiderer, “CROSSLOAD.” <https://crossload.org/info/ueber>, Sept. 2022.
- [3] F. Witte, *Testmanagement und Softwaretest*. Wiesbaden: Springer Fachmedien, 2016.
- [4] J. Goll, *Methoden und Architekturen der Softwaretechnik*. Wiesbaden: Vieweg+Teubner Verlag, 2011.
- [5] A. Bookstein, “Relevance,” *Journal of the American Society for Information Science*, vol. 30, pp. 269–273, Sept. 2007.
- [6] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.
- [7] H. Zaragoza and M. Najork, “Web Search Relevance Ranking,” in *Encyclopedia of Database Systems* (L. Liu and M. T. Özsu, eds.), pp. 4650–4655, New York, NY: Springer New York, 2018.
- [8] V. I. Levenshtein *et al.*, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet Physics Doklady*, vol. 10, pp. 707–710, Soviet Union, 1966.
- [9] D. Liu, X.-s. Hua, M. Wang, H.-j. Zhang, and L. Yang, “WWW 2009 MADRID! Track: Rich Media / Session: Tagging and Clustering Tag Ranking *,” 2009.
- [10] G. Developers, “Image labeling | ML Kit.” <https://developers.google.com/ml-kit/vision/image-labeling>, 2022.
- [11] G. Photos, “Google Photos.” <https://www.google.com/photos/about/>, 2022.
- [12] Crossload, “CROSSLOAD / Backend / solr-schema · GitLab.” <https://gitlab.crossload.org/crossload/backend/solr-schema>, 2022.
- [13] M. Marchiori, “The quest for correct information on the Web: Hyper search engines,” *Computer Networks and ISDN Systems*, vol. 29, pp. 1225–1235, Sept. 1997.
- [14] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the web.,” Technical Report 1999-66, Stanford InfoLab / Stanford InfoLab, Nov. 1999.
- [15] T. Joachims and F. Radlinski, “Search Engines that Learn from Implicit Feedback,” *Computer*, vol. 40, pp. 34–40, Aug. 2007.

- [16] D. Turnbull and J. Berryman, *Relevant Search: With Applications for Solr and Elasticsearch*. Manning, 2016.
- [17] Google, “Tomate - Google Suche.” <https://www.google.de/search?q=tomate&addon=opensearch>, Jan. 2022.
- [18] Bing, “Tomate - Search.” <https://www.bing.com/search?q=tomate&form=QBLH&sp=1&pq=t&sc=10-1&qs=n&sk=&cvid=EEDC767904AD4E1E85359FC4B02026FD&ghsh=0&ghacc=0&ghp>, Jan. 2022.
- [19] Ecosia, “Tomate - Ecosia - Web.” <https://www.ecosia.org/search?q=tomate&addon=opensearch>, Jan. 2022.
- [20] M. Taylor, H. Zaragoza, N. Craswell, S. Robertson, and C. Burges, “Optimisation methods for ranking functions with multiple parameters,” in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management - CIKM '06*, (Arlington, Virginia, USA), p. 585, ACM Press, 2006.
- [21] A. Lucene, “Welcome to Apache Lucene.” <https://lucene.apache.org/index.html>, 2022.
- [22] U. Schindler, B. Bräuer, and M. Diepenbroek, “Data Information Service based on Open Archives Initiative Protocols and Apache Lucene,” in *German E-Science Conference*, May 2007.
- [23] Solr, “Apache Solr Reference Guide :: Apache Solr Reference Guide.” <https://solr.apache.org/guide/solr/latest/>, 2022.
- [24] CLV, “CLV | Bücher, die weiterhelfen.” <https://clv.de/>, 2022.
- [25] Evangelium21 e.V., “Startseite | Evangelium21.” <https://www.evangelium21.net/>, 2022.
- [26] Matomo, “Matomo - The Google Analytics alternative that protects your data.” <https://matomo.org/>, 2022.
- [27] I. Jacobson, I. Spence, and B. Kerr, “Use-Case 2.0: The Hub of Software Development,” *Queue*, vol. 14, pp. 94–123, Jan. 2016.
- [28] R. Jeffries, “Essential XP: Card, Conversation, Confirmation.” <https://ronjeffries.com/xprog/articles/expcardconversationconfirmation/>, 2022.
- [29] D. Synonyme, “Synonyme online finden | Synonymwörterbuch | Duden.” <https://www.duden.de/synonyme>, 2022.
- [30] C. G. C. frontend, “Integration...studienarbeit · CROSSLOAD / Frontend / Frontend · GitLab.” <https://gitlab.crossload.org/crossload/frontend/frontend/-/compare/integration...studienarbeit>, 2022.

- [31] C. G. C. solr-search, “Master...studienarbeit · CROSSLOAD / Backend / solr-search · GitLab.” <https://gitlab.crossload.org/crossload/backend/solr-search/-/compare/master...studienarbeit>, 2022.
- [32] Alphabet, “Werbeumsätze der Google-Websites weltweit in den Jahren 2001 bis 2021.” <https://de.statista.com/statistik/daten/studie/75181/umfrage/werbeumsatz-der-google-websites-seit-2001/>, Feb. 2022.
- [33] NetMarketShare, “Marktanteile der Suchmaschinen - Mobil und stationär 2022.” <https://de.statista.com/statistik/daten/studie/222849/umfrage/marktanteile-der-suchmaschinen-weltweit/>, Feb. 2022.