Erschließung, Analyse und Durchsuchbarkeit von Schulwebseiten

Um thematisch eingeschränkte Bereiche des WWW[1] für eine detailierte inhaltliche Analyse oder auch eine sogenannte fokussierte Suche bereitzustellen, wird zunächst ein Crawl der Webseiten benötigt, der dann einerseits z. B. mit Verfahren der automatischen Sprachverarbeitung (engl. natural language processing, NLP) analysiert werden kann und andererseits auch als Basis einer speziellen Suchmaschine dienen kann.

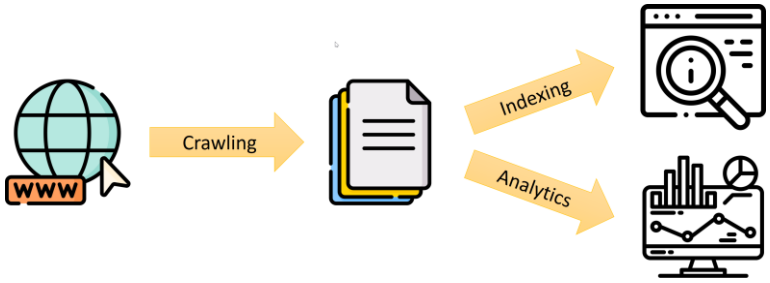

Abbildung 1 Bausteine der Verarbeitungspipeline (Bildquelle: Flaticon.com)

Crawling

Das Crawling – also das Herunterladen eines kleinen Teilausschnitts des WWW unter Ausnutzung und Verfolgung der Linkstrukturen – beginnt üblicherweise mit dem Festlegen der sogenannten Seed-URLs. Der Begriff „Seed“ , zu deutsch Samen, bezeichnet im Falle eines Crawls die Adressen der initialien Webseiten, welche von dem Crawler für die Erstellung der Dokumentenkollektion (in der Mitte in Abbildung 1) zunächst heruntergeladen werden sollen.

Bei der Beschaffung und Bereitstellung der Seed-URLs entsteht bereits das erste „Vereinheitlichungsproblem“. Die verfügbaren Listen mit den URLs zu einzelnen Schulen liegen in unterschiedlichen Formaten mit unterschiedlichen Informatsionsinhalten vor. Deshalb müssen in einem ersten Schritt die verschiedenen Formate vereinheitlicht werden. Dieser Schritt wird weithin als Information Extraction (IE) bezeichnet. Er zielt in unserem Fall darauf ab, aus vielen, heterogenen (semi-)strukturierten Datensätzen einen einheitlichen Datensatz mit den benötigten Informationen zu generieren. Im Fall der Schulwebseiten werden deshalb aus den bereitgestellten Daten bzw. Listen die Schulnamen, die URLs, Postleitzahl, Ort und Addresse extrahiert.

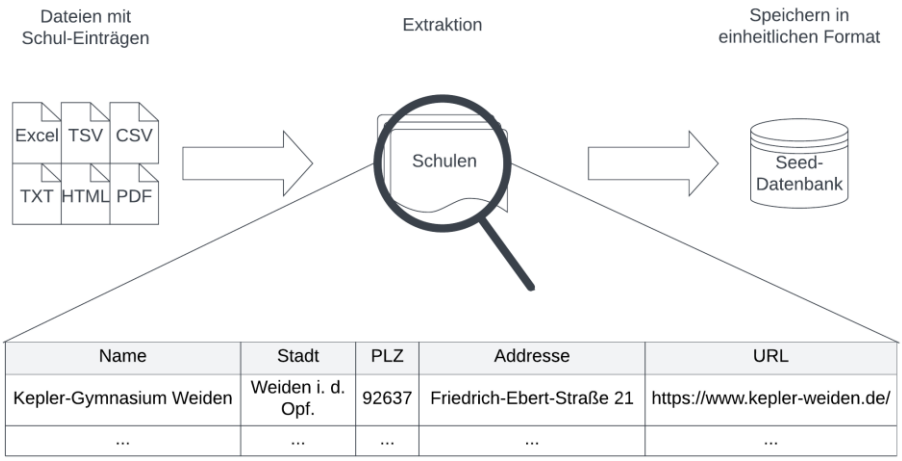

Abbildung 2 Erstellung möglichst eihentlicher Ausgangsdaten mit Seed-URLs

[1] Im Folgenden werden *World Wide Web*, *WWW* und *Web* synonym verwendet.

Die Qualität der Extraktion ist dabei kritisch, da sie die Möglichkeiten der örtlichen Zuordnung ebenso beeinflusst, wie z. B. die Erkennung von Duplikaten.

Die auf diese Weise bereitgestellten Seed-URLs werden dem Crawler der Vertikalen Suche übergeben. Dieser beginnt nun die Webseiten der Schulen herunterzuladen, weiterführende URLs zu extrahieren, sowie erste Analysen durchzuführen und Metadaten zu generieren. Dieser Schritt ist laufzeittechnisch anspruchsvall, da der in eingesetzte Crawler[2] nicht nur eine grobe Datenextraktion und Bereinigung durchführt, sondern auch einen Netzwerkgraphen sowie ein Recovery-System (für Systemausfälle) verwendet. Da der Crawler als „Distributed Real Time Computation System" konzipiert ist, ist es jedoch möglich die Laufzeit mithilfe von Rechenclustern zu reduzieren.

Die vom Crawler heruntergeladenen Webseiten, sowie die generieten Metadaten, werden in einer Datenbank abgelegt und für die Indexierung der hauseigenen vertikalen Search Engine bereitgestellt.

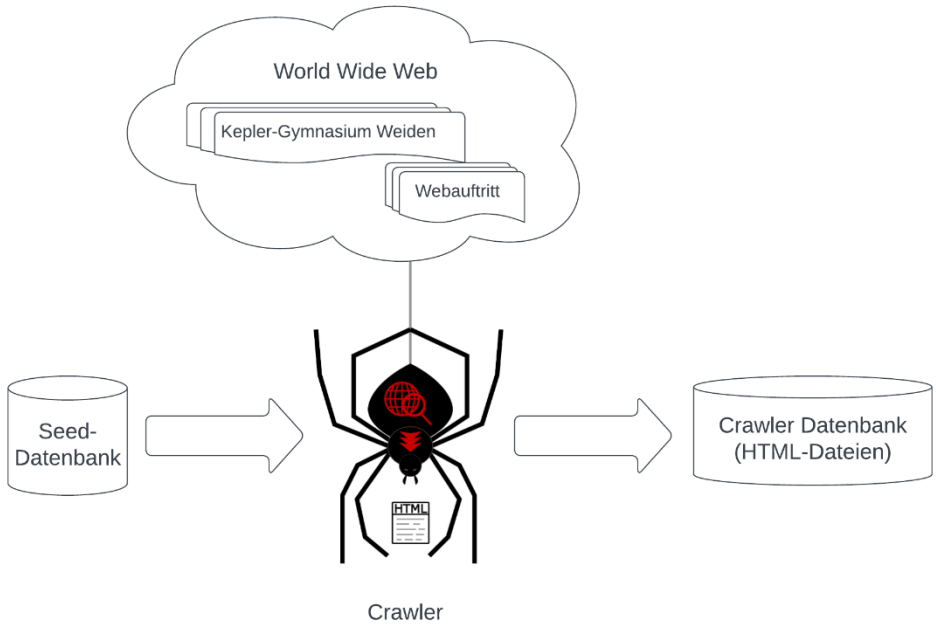

Abbildung 3 Erstellung der Crawler Datenbank auf Basis der Seed-URLs

Beim Crawling müssen sehr viele Parameter optimiert werden. Zunächst ist hier die Politeness Policy zu nennen, die insbesondere verhindern soll, dass Webseiten von Schulen durch den Crawler zu stark belastet werden. Noch wichtiger für das Ergebnis sind aber die Parameter zum Crawl selbst:

- Soll beim Crawlen einer Webseite eine Breitensuche oder eine Tiefensuche durchgeführt werden? (Beeinflusst die Reihenfolge der Betrachtung der Webseiten und damit in Verbindung mit den anderen Parametern auch, welche Seiten eines Web-Auftritts gecrawlt werden)
- Wie viele Seiten sollen pro Web-Auftritt maximal betrachtet werden?
- Welches Datenvolumen soll pro Web-Auftritt maximal betrachtet werden?
- Welche Datentypen sollen in den Crawl einbezogen werden (HTML, PDF, Bilder, …)?
- Sollen auch Web-Seiten außerhalb der eigentlichen Seed-URL gecrawlt werden? (Seiten, die von einem Web-Auftritt verlinkt werden, geben zum Teil auch Aufschluss über Aktivitäten, Themen, …)

Von großer Bedeutung ist auch die Behandlung der einzelnen zugegriffenen Seiten. Zum einen muss ggf. auch JavaScript ausgeführt werden, da aktuelle Webseiten Inhalte häufig erst asynchron nachladen. Zum anderen sind Inhalte ggf. auch bereits zu bereinigen und zu analysieren. Z. B. Hinweise und

[2] StormCrawler: http://stormcrawler.net/

Links zum Datenschutz sollten ggf. nicht weiterverfolgt werden. Hier stellt sich auch die Frage, welche Aspekte beim Crawling und welche bei der Analyse behandelt werden sollten.

Analytics

Der eine Pfad der Auswertung der gecrawlten Dokumente besteht dann in einer übergreifenden Analyse, die zunächst weniger dazu gedacht ist, konkrete Fragen zu beantworten, sondern einen Überblick zu gewinnen. Wenn man es zugespitzt formulieren will, dann geht es hier nicht um das überprüfen von Hypothesen, sondern um das generieren von Hypothesen.

Dazu sind verschiedene Verfahren denkbar.

- Topic-Modeling zu thematischen Übersicht
- räumliche Analysen
- Named-Entity-Recognition
- einfache Wortstatistiken
- Sentiment Analysis
- Trend Mining (sofern man den Datenbestand über die Zeit analysieret)
- …

Für all diese Verfahren ist die Qualität der Datenbasis eine wichtige Voraussetzung:

- Der Datenbestand sollte balanciert sein. Einzelne Schulen mit umfangreichen Webauftritten sollten den Datenbestand nicht dominieren.
- Sogenannte „Boilerplate“-Inhalte sollten nicht betrachtet werden. Beispiele sind Navigationstexte, Texte zur DSGVO …).
- Probleme der Mehrsprachigkeit sind zu lösen.
- …

Die Sicherstellung dieser „Datenqualität“ stellt vor dem Hintergrund der technischen und inhaltlichen Heterogenität der Web-Auftritte der Schulen eine Herausforderung dar.

Indexing

Um konkrete Informationsbedarfe erfüllen zu können, werden die Dokumente in eine sogenannte vertikale Suchlösung eingespeist. Technisch kann man hier z. B. auf Solr[3] oder ElasticSearch[4] setzen.

Die Aufgaben im Rahmen einer vertikalen Suchlösung teilen sich in zwei „Schritte“ auf: Das Indexieren zur „offline-Zeit“ und die eigentliche Suche zur „online-Zeit“.

Während der Indexierung werden die Daten aus der Crawler-Datenbank ausgelesen und für die Suche vorverarbeitet. Dabei werden nicht nur die einzelnen Dokumente nach bestimmten Kriterien ausgewählt und eingeordnet. Ferner werden sowohl die Schulen als auch die Dokumente mit weiteren Informationen angereichert und erweitert (z. B. Koordinaten der Schulgebäude). Zusätzlich werden in diesem Fall statistische Sprachmodelle (z. B. LDA oder Word Embeddings) sowie spezielle Such-Indices (z. B. Schule ↔ Dokument) angelegt.

[3] https://solr.apache.org/
[4] https://www.elastic.co/de/elasticsearch/

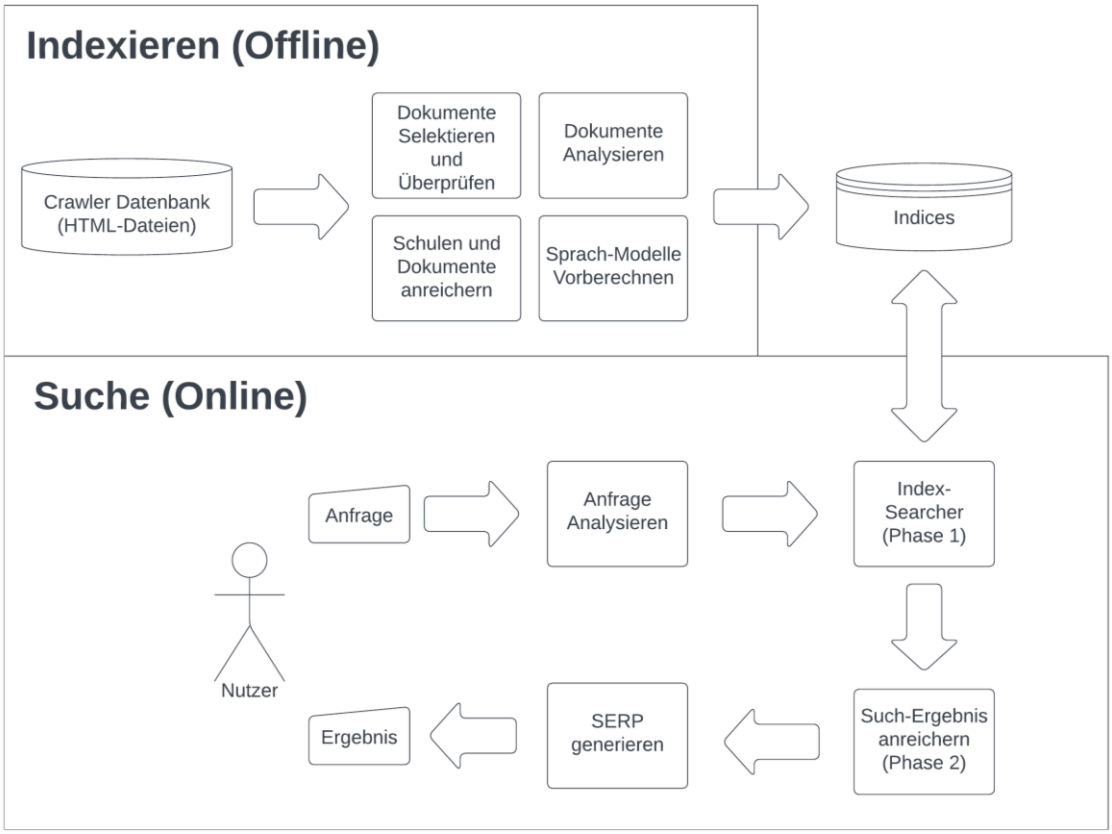

Abbildung 4Teilschritte und Aufbau der vertikalen Suchmaschine

Der Ablauf einer Suche verläuft immer nach dem gleichen Schema: Ein Nutzer schickt eine Anfrage an die Suchmaschine, diese analysiert die Anfrage und gibt die verarbeitete Anfrage an den „Index-Searcher“ weiter. Der Index-Searcher verwendet die Anfrage, um die Dokumente in einer gerankten Reihenfolge zurückzugebnen. Anschließend werden die Ergebnisdokumente in der SERP (Search Engine Result Page) dargestellt. Dabei kann eine vertikale Suchmaschine z. B. der Tatsache Rechnung tragen, dass in diesem Fall sowohl nach Dokumenten (Webseiten, PDF-Dateien …) als auch nach Schulen gesucht werden soll. Die SERP kann daher erlauben auch eine zu Schulen aggregierte Ergebnisdarstellung und ein entsprechendes Ranking zu wählen.

Herausforderungen bestehen hier in der Aufbereitung der Dokumente (zum Teil parallel zur Analyse) und in der Optimierung der SERP.

Mit der Suchlösung soll dann die Beantwortung konkreterer Fragen (z. B. „Welche Schulen bieten Schulorchester an?“) zu unterstutzen. Wie bei allen anderen Teilen der Pipeline ist dabei vor und neben der technischen Umsetzung zunächst eine umfangreiche Anforderungsanalyse erforderlich.