Bachelorarbeit

Suchoptimierung mittels maschinellem Lernen

Zeitraum 04.07.2016 - 04.10.2016

Lukas Abegg

Matrikelnummer 798972
Sommersemester 2016
Fachsemester 6
Studiengang Medieninformatik (B.Sc.)
Beuth Hochschule für Technik

Beuth Hochschule für Technik
Luxemburger Str. 10
13353 Berlin

1. Betreuer **Prof. Dr. habil. Alexander Löser**
Fachbereich VI - Informatik und Medien
Beuth Hochschule für Technik

2. Betreuer **Prof. Dr. Martin Oellrich**
Fachbereich II - Mathematik - Physik - Chemie
Beuth Hochschule für Technik

Inhaltsverzeichnis

1 Einführung

Springer Nature ist ein weltweit führender Verlag für Forschungs-, Bildungs- und Fachliteratur mit einer breiten Palette an angesehenen und bekannten Marken. Die Verlagsgruppe bietet qualitativ hochwertige Produkte und Dienstleistungen. Springer Nature ist zudem der größte Verlag für Wissenschaftsbücher. Für Springer Nature ist es darum wichtig, auf ihren Web-Applikationen eine Suche anbieten zu können, die Suchintentionen erkennt und möglichst schnell zum gesuchten Content leitet. Die Suche wird vor allem als Hilfsmittel zur Navigation und Suche nach Literatur und Dienstleistungen genutzt. Durch die vielen von Springer Nature publizierten Zeitschriften und Querverweise in Artikeln, wird sie aber auch oft zur Suche nach Issues[1] und Artikeln verwendet sowie als Hilfestellung um Diagnosen zu Krankheitsbilder stellen zu können.

1.1 Aufbau der Suche bei Springer Nature

Damit die verschiedenen Verlage und Zeitschriften der Verlagsgruppe Springer Nature ihre Produkte und Dienstleistungen online anbieten können nutzt Springer Nature eine inhouse entwickelte White Label Applikation[2]. Die White Label Applikation verwendet *Apache Solr* als Suchplattform. Die Solr dient hierbei als eine der Schnittstellen zwischen dem Content-Pool von Springer und der Core-Applikation. Bei dem vom Content-Pool gelieferten Content, handelt es sich um vom Springer-Verlag publizierte Zeitschriften, Artikel, Bücher, Chapters und redaktionelle Inhalte.

Um das Verhalten der User auf ihren Web-Applikationen zu tracken verwendet Springer das Analysetool Webtrekk. Die daraus resultierenden Reports bieten unter anderem die Möglichkeit, *Suchquery-Logs* und *Click-Trough-Rates*[3] der User auszuwerten.

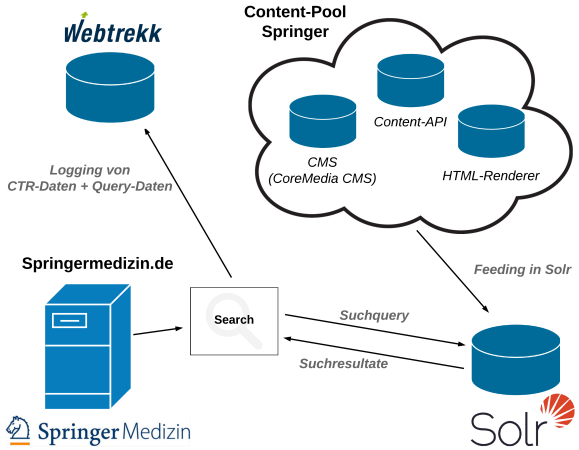

Abb. 1: Aufbau der Suche bei Springer Nature

[1] Nummer der Zeitschriftenausgabe, in der sich der Artikel befindet.
[2] „weißes Etikett“ - eine nicht beschriftete Applikation, welche von anderen Firmen unter deren Namen verwendet werden kann
[3] Kennzahl um die Anzahl der Klicks auf Links im Verhältnis zu den gesamten Impressionen darzustellen

1.2 Problemstellung

Eine konventionelle Volltextsuche ordnet Suchergebnisse danach an, wie *prominent* die Suchwörter in den einzelnen Ergebnissen vorhanden sind. Diese Art von Suche funktioniert grundsätzlich gut, solange die Prominenz der Suchwörter für die relevantesten Dokumente am höchsten ist. [Bas13]

Wird die Suche nun aber aus Usersicht beurteilt, fließen plötzlich ganz andere Faktoren in die Bewertung der Suchqualität ein, wie zum Beispiel: Welche Ergebnisse werden zu welchen Suchanfragen am meisten angeklickt? Oder, sind die Top-Suchergebnisse auch wirklich die für den User relevantesten Dokumente? Laut Studien kann unter Einbezug dieser User-Feedback Daten, die Sortierung der Top-Resultate in Suchmaschinen, signifikant beeinflusst werden. [EA06]

Springermedizin.de ist ein Fortbildungs- und Informationsportal für Ärzte. Diese suchen oft mit einschlägig, fundierten Fachbegriffen nach den neuesten und relevantesten Zeitschriften, Bücher oder Publikationen. Die zeitlich aktuellsten Suchtreffer zu finden ist für Springer kein Problem. Die für den User *relevantesten* jedoch schon.

1.3 Ziel der Arbeit

Springer Nature sammelt viele User-Tracking-Daten und dadurch viel Wissen über das Verhalten der User auf ihrer Suche, lässt dieses Wissen jedoch bisher nicht wirklich relevant in ihre Suche einfließen. Durch die *Click-Trough-Rates* der Suchergebnisse auf den Suchterm bezogen, können die für die Nutzer der Suche relevanten Dokumente im Suchresultat bevorzugt werden. Diese Click-Trough-Rates können direkt aus Webtrekk gelesen werden.

Die Click-Trough-Rate als absoluten Wert für das *Relevanzfeedback* zu nehmen, wäre jedoch falsch. Es muss davon ausgegangen werden, dass viele User der Qualität der Suchmaschine vertrauen und die Top-Suchresultate als die relevantesten Suchresultate betrachten. [TJ05] Das Relevanzfeedback muss daher in Relation zu anderen Faktoren betrachtet werden um eine wirkliche Verbesserung der Suchergebnisqualität erzielen zu können. Ein interessanter Ansatz ist hierbei das *position-based Model* (PBM). [Chu+15] Dieses geht davon aus, dass die Wahrscheinlichkeit, dass ein User ein Dokument wirklich genau analysiert bevor er es anklickt, davon abhängt wie *schlecht* dieses Dokument im Suchresultat gerankt ist. Je *schlechter* das Ranking des angeklickten Dokumentes ist, je *höher* ist das Relevanzfeedback zu bewerten.

Wird nun mittels der oben erwähnten Click-Trough-Rate in Verbindung mit dem position-based Model, die Relevanz der angeklickten Dokumente berechnet und darauf basierend ein Sortier-Algorithmus für die von der Solr zurückgegebenen Suchresultate entwickelt, müsste davon ausgegangen werden, eine Verbessung der Suchergebnisqualität erzielen zu können. Ziel dieser Arbeit ist es darum die Verbesserung der Suchqualität mittels Einbezug dieser Relevanzberechnung zu messen. Wichtig ist hierbei auch kritisch zu hinterfragen wie gut und unter welchen Voraussetzungen diese Lösung produktiv eingesetzt werden kann.

1.4 Methodik

In der aufgestellten These werden *Feedback-Strategien* für die Click-Trough-Rate Auswertung, wie in [TJ05] beschrieben, nicht verwendet. Diese werden in der Thesis auch nicht beachtet. Ebenfalls wird von komplexen Lern-Algorithmen wie in [EA06] vorgestellt, abgesehen. Der neue Lösungsansatz greift auch nicht in die Suchquery der Solr-Abfrage ein sondern sortiert das Ergebnis der Solr-Suche neu.

Die Relevanzberechnung für die im Suchresultat ausgespielten Dokumente soll auf Basis des angesprochenen *position-based Model* [Chu+15] umgesetzt werden. Die Wissensbasis für den Algorithmus bildet Webtrekk. Der Vorteil bei dieser Lösung ist, dass der Algorithmus nicht ständig neues Wissen lernen uns altes vergessen muss, sondern mit einem tagesaktuellen und über eine frei definierbare Periode terminiertes Wissens arbeiten kann. Für die Bachelorthesis wird der Webtrekk-Account von *Springermedizin.de* verwendet und pro Suchterm spezifische Analysen über einen vordefinierten Zeitraum (die letzten 30 Tage) durchgeführt. Über die Webtrekk-API können diese Analysen zur Laufzeit gelesen und verarbeitet werden.

Wie bereits in Kapitel 1.3 angesprochen, wird der Nutzer der Suche durch die Top-Suchresultate beeinflusst. Um dem entgegenzuwirken wird mithilfe eines zusätzlichen Zufallsranking der Suchergebnisliste, die Relevanz der Dokumente beeinflusst, um von der Solr-Suche schlecht gerankte Ergebnisse auch in den Fokus des Nutzers zu rücken.

Das große Kernproblem der Überprüfung wird das Messen der Qualität der erzielten Suchergebnisse sein. Um die Verbesserung der Suchqualität durch die aufgestellte These messen zu können, werden Analysen benötigt, die aussagen, wie gut die Qualität der aktuellen Suche im Vergleich zum neuen Lösungsansatz ist, wie viel Verbesserung der Ansatz bringt und an welchen Schrauben noch etwas gedreht werden muss, damit die Suche wirklich gute Ergebnisse aus Sicht der User bringt. Dazu muss eine passende Testumgebung aufgebaut werden. Auf einem Evaluationssystem sollen fachlichen Experten die Relevanz der Suchergebnisse der beiden Suchmaschinen auf Basis gleicher Suchanfragen bewerten. Mithilfe des Bewertungsalgorithmus *NDCG* kann dann anhand der Ergebnisse der Bewertungen, das Qualitätsmaß der beiden Suchen vergleichen und die Verbesserung der Suchqualität durch die neu implementierte Lösung festgestellt werden.

1.5 Gliederung und Aufbau

Wann lesen wir was und warum?

Literatur

[Bas13] Hannah Bast. „Semantische Suche“. In: *Informatik-Spektrum* 36.2 (2013), S. 136–143 (siehe S. 2).

[Chu+15] Aleksandr Chuklin, Ilya Markov und Maarten de Rijke. *Click Models for Web Search*. Morgan & Claypool, 2015 (siehe S. 2, 3).

[EA06] Susan Dumais Eugene Agichtein Eric Brill. *Improving Web Search Ranking by Incorporating User Behavior Information*. Report. Microsoft, 2006 (siehe S. 2, 3).

[HJ15] Thomas Hoppe und Horst Junghans. „Corporate Semantic Web: Wie semantische Anwendungen in Unternehmen Nutzen stiften“. In: Springer Berlin Heidelberg, 2015, S. 111–128.

[TJ05] Bing Pan Thorsten Joachims Laura Granka. *Accurately Interpreting Clickthrough Data as Implicit Feedback*. Report. Random, 2005 (siehe S. 2, 3).

Abbildungs-Verzeichnis

Tabellen-Verzeichnis

Sourcecode-Verzeichnis