

Bachelorarbeit

Suchoptimierung mittels maschinellen Lernens

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 BERLIN
University of Applied Sciences

Beuth Hochschule für Technik

Luxemburger Str. 10
13353 Berlin

Betreuer

Prof. Dr. habil. Alexander Löser

Fachbereich VI - Informatik und Medien

Beuth Hochschule für Technik

Gutachter

Prof. Dr. Martin Oellrich

Fachbereich II - Mathematik - Physik - Chemie

Beuth Hochschule für Technik

Inhaltsverzeichnis

1 Einführung	1
1.1 Aufbau der Suche bei Springer Nature	1
White Label Applikation mit Solr-Suche	1
User-Tracking mit Webtrekk	1
Architektur	2
1.2 Das große Problem	2
Der Springer Nature Stakeholder Springermedizin setzt auf Webtrekk	2
Keine Userrelevanz in der Suche	2
Der fast gläserne User	2
1.3 Ziel der Arbeit	3
1.4 Methodik	3
1.5 Gliederung und Aufbau	4
Abbildungs-Verzeichnis	5
Tabellen-Verzeichnis	6
Sourcecode-Verzeichnis	7

Einführung

Springer Nature ist ein weltweit führender Verlag für Forschungs-, Bildungs- und Fachliteratur mit einer breiten Palette an angesehenen und bekannten Marken und 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 Issues¹ und Artikeln verwendet sowie als Hilfestellung um Diagnosen zu Krankheitsbilder stellen zu können.

Springer Nature sammelt viele User-Tracking-Daten und dadurch viel Wissen über das Verhalten der User auf ihrer Suche, lässt dieses Wissen jedoch nicht in ihre Suche einfließen. In dieser Arbeit wird untersucht, wie mithilfe dieses Wissens, die Suche optimiert werden kann.

1.1 Aufbau der Suche bei Springer Nature

White Label Applikation mit Solr-Suche

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 Applikation². 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.

User-Tracking mit Webtrekk

Um das Verhalten der User auf ihren Web-Applikationen zu tracken verwendet Springer das Analysetool Webtrekk³. Die daraus resultierenden Reports bieten unter anderem die Möglichkeit, *Suchquery-Logs* und *Click-Trough-Rates*⁴ der User auszuwerten.

¹Nummer 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<https://www.webtrekk.com>

⁴Kennzahl um die Anzahl der Klicks auf Links im Verhältnis zu den gesamten Impressionen darzustellen

Architektur

In Abb. 1 ist die Suche nochmals grafisch aufbereitet:

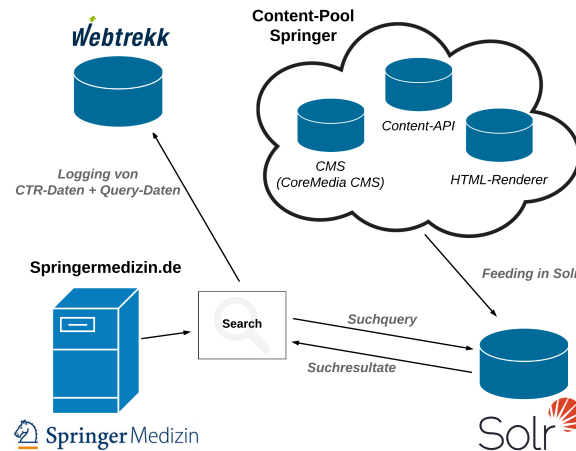


Abb. 1: Aufbau der Suche bei Springer Nature

1.2 Das große Problem

Der Springer Nature Stakeholder Springermedizin setzt auf Webtrekk

Zu den Stakeholder⁵ der in 1.1 angesprochenen White Label Applikation gehört *Springermedizin*⁶. Springermedizin ist ein Fortbildungs- und Informationsportal für Ärzte. Mithilfe von Webanalysten und Webtrekk versucht Springermedizin das Marketing ihres Webauftrittes zu verbessern und ist sehr interessiert an neuen Ansätzen, um die gesammelten Tracking-Daten besser einzusetzen. In dieser Arbeit wird darum der Fokus auf die Suche von Springermedizin gesetzt.

Keine Userrelevanz in der Suche

Die User von Springermedizin suchen oft mit einschlägig, fundierten Fachbegriffen nach den neuesten und relevantesten Zeitschriften, Bücher oder Publikationen. Die zeitlich aktuellsten Suchtreffer zu finden ist für Springermedizin kein Problem. Die für den User *relevantesten* jedoch schon.

Der fast gläserne User

Springermedizin sammelt Tracking-Daten über jegliche Aktivitäten auf deren Applikationen und investiert Zeit und Geld in die Individualisierung der Analysedaten auf Webtrekk. Mittlerweile sind knapp 30 Custom-Parameter⁷ auf Webtrekk angelegt um genau die Daten zu tracken, die sie über das Verhalten der User auf ihrer Applikationen wissen wollen. Dadurch entsteht ein fast „gläsernen User“.

⁵bezeichnet Springer-interne Kunden, die ein Interesse am Ergebnis der White Label Applikation haben

⁶<https://www.springermedizin.de/>

⁷Individuell erzeugte Parameter für Reports und Analysen

1.3 Ziel der Arbeit

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. [Joachims] 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). [chuklin2015] 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 Verbesserung 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 [Joachims] beschrieben, nicht verwendet. Diese werden in der Thesis auch nicht beachtet. Ebenfalls wird von komplexen Lern-Algorithmen wie in [IWUSBI] 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* [chuklin2015] umgesetzt werden. Die Wissensbasis für den Algorithmus bildet Webtrekk. Der Vorteil bei dieser Lösung ist, dass der Algorithmus nicht ständig neues Wissen lernen und 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?

Abbildungs-Verzeichnis

1	Aufbau der Suche bei Springer Nature	2
---	--	---

Tabellen-Verzeichnis

Sourcecode-Verzeichnis