Bachelorarbeit

Suchoptimierung mittels maschinellen Lernens

Zeitraum 04.07.2016 - 11.10.2016

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

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

Springer Nature ist ein weltweit führender Verlag für Forschungs-, Bildungs- und Fachliteratur mit einer breiten Palette an angesehenen und bekannten Medienmarken und zudem der weltweit größte Verlag für Wissenschaftsbücher. Für das Unternehmen Springer Nature ist es darum wichtig, auf seinen 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 zum Finden von Literatur und Dienstleistungen genutzt. Durch die vielen von Springer Nature publizierten[1] Zeitschriften und Querverweise in Artikeln, wird sie aber auch oft zur Suche nach Issues[2] 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[3] bei der Nutzung ihrer Suche, lässt dieses Wissen jedoch bisher noch nicht in ihre Suche einfließen. In dieser Arbeit wollen wir untersuchen, ob 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 eigens entwickelte White Label Applikation[4]. Die White Label Applikation verwendet *Apache Solr* (im Folgenden „Solr“ genannt, siehe [**solr**]) als Suchplattform. Die Solr dient hierbei als eine der Schnittstellen zwischen dem Content-Pool von Springer Nature und der White Label Applikation. Bei den vom Content-Pool gelieferten Inhalten, handelt es sich um von Springer Nature Verlag publizierte Zeitschriften, Artikel, Bücher und redaktionelle Inhalte.

User-Tracking mit Webtrekk

Um das Verhalten der User auf ihren Web-Applikationen zu tracken verwendet Springer Nature das Analysetool Webtrekk (siehe [**webtrekk**]). Die daraus resultierenden Berichte bieten unter anderem die Möglichkeit, *Suchquery-Logs*[5] und *Click-Trough-Rates* (CTR)[6] der User auszuwerten.

[1] Unter publizieren wird in dieser Arbeit die Veröffentlichung auf der Springermedizin-Applikation bezeichnet
[2] Nummer der Zeitschriftenausgabe, in der sich der Artikel befindet.
[3] Als User werden die Nutzer der Springer Nature Suche bezeichnet
[4] Eine White Label Applikation ist eine wiederverwendbare und agil erweiterbare Applikation
[5] Protokoll über alle ausgeführten Suchanfragen auf der Applikation
[6] Kennzahl um die Anzahl der Klicks auf Links im Verhältnis zu den gesamten Impressionen darzustellen

Architektur

In Abb. 1 ist die Suche nochmals grafisch aufbereitet:

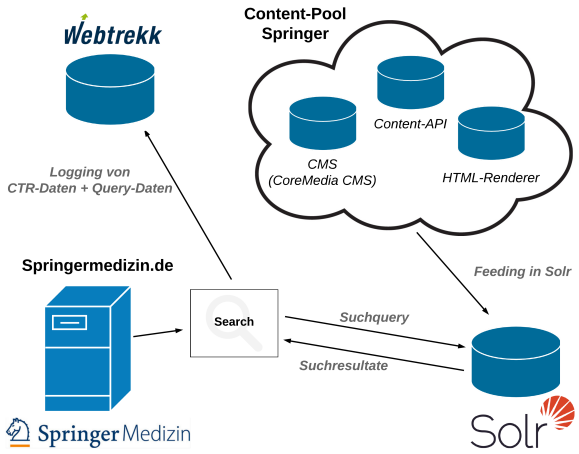

Abb. 1: Aufbau der Suche bei Springer Nature

1.2 Problemstellung: Keine Userrelevanz in der Suche

Zu den Stakeholder[7] der in Kapitel 1.1 angesprochenen White Label Applikation gehört *Springermedizin* (siehe [**SMED**]). Springermedizin betreibt ein Fortbildungs- und Informationsportal für Ärzte.

Userrelevante Dokumente werden nicht gefunden

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 Springer Nature Stakeholder: Springermedizin setzt auf Webtrekk

Mithilfe von Webanalysten und Webtrekk versucht Springermedizin das Marketing seines 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 Verwendung von Tracking-Daten in der Suche von Springermedizin gesetzt.

Der fast gläserne User

Springermedizin sammelt Tracking-Daten über jegliche Aktivitäten auf deren Applikationen und investiert Zeit und Geld in die Individualisierung[8] der Analysedaten auf Webtrekk. Mittlerweile sind knapp 30 Custom-Parameter[9] auf Webtrekk angelegt um genau die Daten zu tracken, die zur Analyse des Verhaltens der User auf ihrer Applikationen relevant sind. Dadurch entsteht ein fast „gläsernen User“. Dieses Wissen könnte zum Vorteil des Users eingesetzt werden, indem es in der Suche verwendet wird.

[7] Bezeichnet Springer Nature interne Kunden, die ein Interesse am Ergebnis der White Label Applikation haben
[8] Mit Individualisierung wird die Speicherung eigener Parameter bezeichnet
[9] Individuell erzeugte Parameter für Berichte und Analysen

1.3 Ziel der Arbeit

1.3.1 Suchoptimierung durch Click-Trough-Daten

In dieser Arbeit werden wir untersuchen, ob mithilfe der von Springermedizin gesammelten Click-Trough-Daten[10] dessen Suche verbessert werden kann. Konkret wollen wir dies anhand eines Algorithmus basierend auf dem Klick-Modell[11] *Position-Based Modell* (PBM, siehe [**pbm**]) untersuchen.

Annahmen

Wir gehen dabei von folgenden Annahmen aus. Relevante Dokumente sind wichtiger als nicht relevante Dokumente. Eine Suchergebnis ist dann gut, wenn die relevanten Ergebnisse in der verwendeten Hierarchie vor den nicht relevanten Ergebnisse auftauchen.

Anwendung auf das Springermedizin-Umfeld

Wir werden versuchen, die Click-Trough-Rates der Suchresultate mithilfe des oben angesprochenen Algorithmus zu berechnen und mit diesen ein *Reranking*[12] der Suchresultate auf das Springermedizin-Umfeld abzubilden. Die Herausforderung wird hierbei die Adaptierung des Lösungsansatzes auf das Springermedizin-Umfeld sein. Im Idealfall widerspiegeln die gesammelten Click-Trough-Daten der Userrelevanz der einzelnen Dokumente[13].

1.3.2 Abbildung auf das Springermedizin-Umfeld

Potential von Userrelevanzen in der Suchoptimierung analysieren

Die Analyse von User-Tracking-Daten bietet viel Potential bezogen auf Userrelevanzen. Sind anhand des hier umgesetzten Lösungsansatzes Verbesserungen in der Qualität der Suche zu verzeichnen, möchte Springermedizin in Zukunft vermehrt User-Tracking-Daten in die Suche einfließen lassen. Diese Arbeit könnte dann als Fundament für weitere Lösungsansätze dienen.

Bekanntes und wirkungsvolles Information Retrieval Verfahren

Suchoptimierung mittels Userrelevanz ist ein bekanntes und nicht triviales, aber relativ wirkungsvolles Information Retrieval Verfahren (siehe [**IWUSBI**]). Seit Mitte der 2000er Jahre wird mithilfe dieses Verfahrens versucht, Suchmaschinen zu verbessern. Aus dieser Zeit stammen auch die ersten Ansätze um mithilfe von Click-Trough-Daten die Userrelevanz der Suchergebnisse zu berechnen (siehe [**Joachims**]).

[10] Mit Click-Trough-Daten bezeichnen wir alle Tracking-Daten, welche während der Interaktion zwischen User und Suche protokolliert werden
[11] Als Klick-Modell wird ein Modell zur Berechnung des Userfeedbacks bzw. der Userrelevanz mithilfe von Click-Trough-Daten bezeichnet
[12] Mit Reranking bezeichnen wie die Umsortierung einer Liste von Suchresultaten
[13] Als Dokumente werden die einzelnen Suchresultate bezeichnet

Lösungsansatz basierend auf Click-Trough-Daten aus Webtrekk

Springermedizin führt ein eigenes Tracking der User durch und verwendet auf Webtrekk selbst definierte Tracking-Parameter. Dadurch hängt die Wahl des in dieser Arbeit zu untersuchenden Lösungsansatzes und dessen Umsetzung stark von den durch Webtrekk gegebenen Analyse-Daten ab.

Keine Gegenüberstellung mit anderen Lösungsansätzen

Durch den vorgegebenen Zeitraum für die Erstellung dieser Bachelorarbeit bedingt, werden wir den Lösungsansatz so wählen, dass er mit den Gegebenheiten bei Springermedizin sinnvoll und in diesem Zeitrahmen realistisch implementiert werden kann. Wir werden daher in dieser Arbeit keine Gegenüberstellung mit anderen Lösungsansätzen machen.

1.4 Methodik

Wie in Kapitel 1.3.1 angesprochen, wollen wir das Klick-Verhalten der User in der Suche analysieren, um mithilfe der daraus berechenbaren Click-Trough-Rates, die Suchergebnisse zu verbessern. Dieses Klick-Verhalten können wir aus den Click-Trough-Daten lesen.

Suchterm semantisch aufschlüsseln

Um mit den Click-Trough-Daten arbeiten zu können, müssen wir zunächst die relevanten Click-Trough-Daten herausfiltern. Dazu müssen wir die Click-Trough-Daten dem *Suchterm*[14] der Anfrage zuordnen können. Zu den Click-Trough-Daten wird immer der Suchterm gespeichert, mit dem dabei gesucht wurde. Das heißt wir können eine Relation zwischen dem Suchterm der Click-Trough-Daten und dem Suchterm unserer Anfrage herstellen.

Die Click-Trough-Daten müssen aber nicht mit dem vollständigen Suchterm in Relation stehen. Sie können auch nur mit einem Wort, einem Teil des Suchterms oder einem Synonym eines dieser Worte in Relation stehen. Wir müssen darum den Suchterm semantisch aufschlüsseln um alle relevanten Click-Trough-Daten filtern zu können.

Aufbereitung Click-Trough-Daten

Können wir alle relevanten Click-Trough-Daten zu einer Suchanfrage filtern, müssen lernen wie wir diese richtig aufbereiten, um die Click-Trough-Daten berechnen zu können. Ein wichtiger Punkt bei der Aufbereitung der Click-Trough-Daten ist die Interpretation des Relevanzfeedbacks der einzelnen Click-Trough-Daten. Nicht jeder Klick ist gleich relevant zu interpretieren. Die Relevanz eines Klicks hängt davon ab, welche Aktionen der User während dem Suchvorgang vor und nach dem Klick durchgeführt hat. Wir müssen darum zuerst analysieren, welche Informationen wir zu den Click-Trough-Daten aus Webtrekk lesen können. Reichen die Informationen für detailliertere Interpretationen nicht aus, müssen wir alle Click-Trough-Daten als gleich relevant lesen.

Result-Reranking mittels PBM basiertem Algorithmus

Wissen wir, wie wir die Click-Trough-Daten aufbereiten, können wir diese zur Berechnung der Click-Trough-Rate eines Dokumentes und somit dessen Userrelevanz einsetzen. Für die Verwendung

[14] Als Suchterm wird eine Suchanfrage bezeichnet

dieser Userrelevanz müssen wir festlegen wann und in welcher Art wir sie einsetzen. Wie bereits in Kapitel 1.3.1 erwähnt, werden wir uns in dieser Arbeit auf die *Aufbereitung* der Suchresultate aus der Solr konzentrieren und dort einen *Reranking-Algorithmus* einbauen.

Der Reranking-Algorithmus basiert auf einem *Klick-Modell*. Dieser verwendet die Click-Trough-Daten, um daraus die Click-Trough-Rate eines Dokumentes zu berechnen. Die Wahl des Klick-Modells hängt darum von den Click-Trough-Daten und den darin enthaltenen Informationen ab. In dieser Arbeit werden wir veranschaulichen, warum wir den Ansatz des Position-Based Modells (siehe [**pbm**]) gewählt haben und wie wir den Algorithmus umgesetzt haben.

Vergessen der alten Daten

Das Position-Based Modell berechnet Wahrscheinlichkeiten, um die Click-Trough-Rate eines Dokumentes zu einer Suchanfrage festzustellen. Dem Algorithmus muss dazu das notwendige Wissen entweder zur Verfügung gestellt oder antrainiert werden. Wird dem Algorithmus das Wissen antrainiert, benötigt der Algorithmus eine Möglichkeit, neues Wissen zu Lernen und altes zu Vergessen.

Mit Webtrekk haben wir eine Wissensbasis, die durch die Springermedizin-Applikation automatisch um neue Click-Trough-Daten ergänzt wird. Das heißt wir müssen uns nicht um eine Möglichkeit zum Lernen neuer Daten kümmern. Wir müssen uns aber überlegen, wie wir altes Wissen vergessen und wie wir dem User neues Wissen präsentieren, damit dieser sich durch die Click-Trough-Rate-Berechnung nicht auf alten Dokumenten festfährt.

1.5 Gliederung und Aufbau

Der Lösungsansatz und deren Grundlagen

In diesem Kapitel wurde der zu untersuchenden Lösungsansatz vorgestellt. Dabei sind wir auf die Hintergründe dieser Arbeit und die Vorgehensweise eingegangen. Im zweiten Kapitel (Grundlagen) folgt die Theorie des beschriebenen Lösungsansatzes. Hier werden wir uns auf die fachlichen Grundlagen konzentrieren.

Umsetzung des Lösungsansatzes

In Kapitel 3 (Reranking mittels Click-Trough-Rate Ergebnis) werden wir die in Kapitel 1.4 angesprochene Methodik verfeinern und detailliert die Vorgehensweise bei der Umsetzung diskutieren. Die Umsetzung selbst folgt dann in Kapitel 4 (Implementierung).

Erkenntnisse verarbeiten

Um zu prüfen ob der umgesetzte Lösungsansatz die erhofften Verbesserungen erzielt, werden wir diesen in Kapitel 5 (Evaluation und Auswertung) in einer Evaluation mit der bisherigen Springermedizin-Suche vergleichen. Aufgrund der resultierenden Erkenntnisse werden wir in Kapitel 6 (Zusammenfassung und Ausblick) ein Fazit ziehen können und einen Ausblick auf mögliche zukünftige Arbeiten geben.

2 Grundlagen

2.1 Grundbegriffe

In diesem Kapitel werden wir die fachlichen Grundlagen zu unserem in Kapitel 1.4 vorgestellten Lösungsansatz aufarbeiten. Wichtig hierfür ist das Verständnis für die Problemstellungen in der Interaktion zwischen den Nutzern der Suche und der Suche selbst und warum unser verfolgter Lösungsansatz schief gehen kann. Dazu gehört die Auseinandersetzung mit dem Klick-Verhalten der User auf der Suche von Springermedizin. Mit der Click-Trough-Rate wollen wir eine Userrelevanz bestimmen. Dazu müssen wir die Click-Trough-Daten als Relevanzfeedback deuten können. Wie wir die Click-Trough-Rates berechnen, lernen wir in den Grundlagen zu unserem Reranking-Algorithmus. Diese Grundlagen sind notwendig für die Umsetzung des Lösungsansatzes in den nachfolgenden Kapiteln 3 und 4.

2.1.1 Semantik von User-Interaktionen

Problemstellungen der Click-Trough-Daten: Was analysieren wir?

Einzelne Wörter oder Teile des Suchterms können in weiteren Suchanfragen vorkommen Ein Suchterm kann aus einem oder mehreren Wörtern bestehen. Jeder User formuliert eine Suchanfrage anders. Sei es die Wortwahl, die Zeitform oder die Verwendung von Bindewörtern. Daraus lässt sich vermuten, dass einzelne Wörter oder Teile des Suchterms in weiteren Suchanfragen vorkommen können. Folglich muss der Suchterm semantisch aufgeschlüsselt werden, um Relationen zwischen Click-Trough-Daten und der Suchanfrage herstellen zu können. Nur so können wir alle relevanten Click-Trough-Daten filtern.

Suchanfragen mit Synonymen und verwandten Begriffen beachten Nehmen wir als Beispiel die Suchanfrage „chronische Dyspnoe“. Würden wir stattdessen den sinnverwandten Suchterm „konstante Atemnot“ verwenden, würden wir für beide Fälle ähnliche Suchresultate erwarten. Folglich würden wir auch ähnliche Click-Trough-Daten vermuten. Wir sollten daher die Synonyme und verwandte Begriffe zu unserem Suchterm ebenfalls beachten und deren Click-Trough-Daten, in den Reranking-Algorithmus einfließen lassen. Dazu benötigen wir eine Wissensbasis, welche die Synonyme und verwandte Begriffe zu unserem Suchterm gespeichert hat. Eine solche Basis bieten Wörterbücher und Thesauri[1]. Beim Content von Springermedizin handelt es sich um medizinische Inhalte in deutscher Sprache. Es macht daher Sinn dies in der Wahl der richtigen Wissensbasis zu berücksichtigen.

Problemstellungen des Lösungsansatzes: Warum kann es schief gehen?

Die folgenden Faktoren leiten sich aus dem verfolgten Lösungsansatz des Reranking-Algorithmus ab, bzw. werden in diesem nicht beachtet. Wir müssen davon ausgehen dass diese das Untersuchungsergebnis des Lösungsansatzes negativ beeinflussen könnten.

[1] Thesauri sind strukturierte Verzeichnisse von Begriffen, die allesamt in irgendeiner Beziehung zueinander stehen bezeichnet

Die Relation des Suchterms zu den Click-Trough-Daten wird nicht gewichtet Die Click-Trough-Daten sind dann relevant, wenn mindestens ein Wort des aufgeschlüsselten Suchterms in Relation zu diesen Daten steht. Dadurch können falsche Relationen entstehen und nicht relevante Click-Trough-Daten die Klick-Wahrscheinlichkeiten der Dokumente im Reranking-Algorithmus negativ beeinflussen.

Intentionen und die Mehrdeutigkeit von Begriffen wird nicht beachtet Die genaue semantische Analyse eines Suchterms beinhaltet unter anderem die Erkennung von Begriffen und deren Mehrdeutigkeiten. Suchte jemand z.B. nach dem Begriff „Brücke“, hat dieser im medizinischen Kontext mehrere Bedeutungen. Es könnte ein „ein Teil des zentralen Nervensystems“ gemeint sein oder eine „Form des Zahnersatzes“. Wie bereits im vorherigen Kapitel 2.1.1 erwähnt, können wir mithilfe eines Thesaurus die verschiedenen Bedeutungen erkennen. Unser Reranking-Algorithmus ignoriert diese Mehrdeutigkeit jedoch. Er würde in diesem Fall alle Click-Trough-Daten zu beiden Begriffsbedeutungen suchen. Hier können wir zufallsbedingt drei Ausgangslagen haben. Die Click-Trough-Daten entsprechen der Suchintention (1) - das wäre der zufallsbedingte Optimalfall. Das Suchresultat würde von der Mehrfachbedeutung nicht beeinflusst werden. Tritt das Gegenteil ein (2) - Das Suchresultat wird in diesem Fall durch eine falsche Relevanz negativ beeinflusst. Keine Click-Trough-Daten vorhanden (3) - Die Volltextsuche und die Klick-Wahrscheinlichkeit der Position im Suchresultat definieren das Suchergebnis. In diesem Fall kann die Wertigkeit des Algorithmus nicht vorhergesehen werden.

Keine Aktualität in der Suche Der von uns verfolgte Reranking-Algorithmus nimmt keine Rücksicht auf die „Aktualität“ eines Beitrages sondern nur auf die Klick-Wahrscheinlichkeit und könnte dadurch aktuellere Dokumente trotz Relevanz, schlecht positionieren im Suchresultat. Die Klick-Wahrscheinlichkeit kann durch zwei Faktoren beeinflusst werden. Hohe Relevanz in der Solr-Suche (1) - wird in der Volltextsuche die Aktualität des Dokumentes in die Berechnung der Relevanz einbezogen, werden aktuelle Beiträge im Suchergebnis der Solr weit vorne eingestuft. Wie wir aus Abb. 3 erkennen können, haben niedrige Positionen eine höhere Klick-Wahrscheinlichkeit. Das könnte die Berechnung des Reranking-Algorithmus positiv beeinflussen. Reranking-Algorithmus um Zufallsfaktor erweitern (2) - mithilfe eines Zufallsfaktor ist die Reihenfolge der Suchresultate weniger vom Algorithmus abhängig und die Wahrscheinlichkeit, aktuelle Dokumente ohne Click-Trough-Daten weit vorne im Suchergebnis zu finden wird erhöht.

Interessante Dokumente werden nie gesehen Wie bereits oben erwähnt, beachtet der Reranking-Algorithmus Dokumente ohne Click-Trough-Daten nur, wenn die Position im Suchresultat eine Klick-Wahrscheinlichkeit aufweist. Dadurch kann es sein, dass interessante Dokumente nie gesehen werden. Dem entgegenwirken können wir ebenfalls mit dem oben erwähnten Zufallsfaktor. Dadurch wird die Klick-Wahrscheinlichkeit für interessante Dokumente zufallsbedingt erhöht.

Reranking-Algorithmus beachtet nur die Top-N-Ergebnisse Der Reranking-Algorithmus wird in die Aufbereitung der Suchresultate aus der Solr-Suche integriert. Wir müssen darum beachten, dass die Solr durch die Pagination-Funktion (siehe [**Pagination**]) nur die Top-N-Ergebnisse zurückgibt. Dadurch sehen wir nur einen Teil der Suchergebnisse. Wenn wir die Abb. 3 betrachten, sehen wir, dass die Klick-Wahrscheinlichkeit mit zunehmender Positionen kleiner wird und bei Position 20 bereits relativ klein ist. Um sicherzustellen, dass wir möglichst alle relevanten Suchergebnisse berücksichtigen, werden wir jeweils nach Relevanz der Solr in absteigender Form 100 Suchergebnisse im Reranking-Algorithmus verarbeiten. Für die Untersuchung des Reranking-Algorithmus werden wir uns bei der Auswertung jeweils auf die Seite 1 der Suchergebnisse konzentrieren. Bei Springermedizin somit auf die ersten 20 Suchresultate. Die Pagination der Folgeseiten der Suchresultate werden wir nicht untersuchen.

Würden wir dies aber implementieren, müssten für die Ausspielung der Folgeseiten des Suchresultats einen Lösungsansatz überlegen, damit die Solr die durch den Reranking-Algorithmus ausgespielten Suchresultate nicht mehrfach ausspielt.

Nicht beeinflussbare Faktoren: Fehlerhafter Content verfälscht die Suchergebnisse

Die folgenden Faktoren beeinflussen das Suchergebnis negativ, sind aber vom Content so vorgegeben. Der von uns verfolgte Lösungsansatz des Reranking-Algorithmus kann diese nicht beeinflussen. Wir beachten diese Faktoren in unserer Arbeit darum nicht.

Mehrfachverwertung des Contents Auf der Springermedizin Suche wird teilweise im Suchergebnis auf denselben Artikel mehrfach verwiesen. Das liegt an der bei Springermedizin praktizierten Mehrfachverwertung des Contents. Es gibt *Journal-Artikel*, das sind aus Journalen, Zeitschriften oder Magazinen stammende Artikel, die auf Springermedizin direkt online[2] gelesen werden können. Bei Neuerscheinung des Artikels, werden dazu oft redaktionelle Artikel publiziert, welche auf den Journal-Artikel verweisen sollen. Diese können im CMS (siehe Abb. 1) von der Suche exkludiert werden. Werden diese nicht exkludiert, können beide Artikel im Suchergebnis erscheinen.

Ausspielung von Teaser Springermedizin verwendet Teaser[3] auf der Startseite und auf Übersichtsseiten zu Rubriken als Einstieg in den nachfolgenden ausführlichen Beitrag. Diese werden auch in der Suche ausgespielt. Teaser sagen nichts über die Wertigkeit des Beitrages aus. Man weiß nicht, auf welche Art von Beitrag (z.B wissenschaftliche Publikation oder ein Artikel aus ein Journal) verwiesen wird und von welchem Autor der Beitrag stammt. Sie können darum nicht nach Relevanz eingestuft werden und sollten darum nicht im Suchergebnis erscheinen.

Fehlerhafte Importe der Daten Viele Beiträge sind falschen Rubriken zugeordnet. Beispielsweise werden Beiträge fälschlicherweise als wissenschaftliche Publikationen publiziert, obwohl sie aus einem Journal oder einer Fachzeitschrift stammen. Diese Fehler sind auf fehlerhafte Importe der Daten zurückzuführen und verfälschen die Wertigkeit des Suchergebnisses.

Wenige Dokumente erhalten viele Klicks

Um das Klick-Verhalten der User auf der Springermedizin-Suche zu verstehen, ist es wichtig anhand oft gesuchter Suchphrasen dieses Verhalten zu analysieren. Dazu wurde eine Analyse über einen Zeitraum von 30 Tagen erstellt und die zehn am häufigsten gesuchten Suchphrasen verwendet. Die Analyse vergleicht für jede Suchphrase die 20 Dokumente mit den meisten Klicks. Die Dokumente wurden hierbei nicht nach Position im Suchergebnis sondern nach Klick-Häufigkeit selektiert. Jeder Graph der folgenden Abbildung 2 stellt eine Suchphrase dar. Wie wir sehen, zeigen die meisten Graphen ein exponentiell stark abnehmendes Verhalten der Klick-Häufigkeiten. Dieses exponentielle Verhalten zeigt, dass einzelne Dokumente häufig und viele Dokumente selten bis nie angeklickt werden. Dieser Effekt kann wie in vielen natürlichen Phänomenen mit exponentiellem Verhalten, durch das Potenzgesetz (Power Law, siehe [**PowerLaw**]) beschrieben werden.

[2] Der Begriff „online“ wird hier als Verweis auf die Springermedizin.de-Webseite verwendet
[3] Als Teaser wird ein kurzer Texte bezeichnet, der das Interesse für den nachfolgenden Beitrag wecken soll

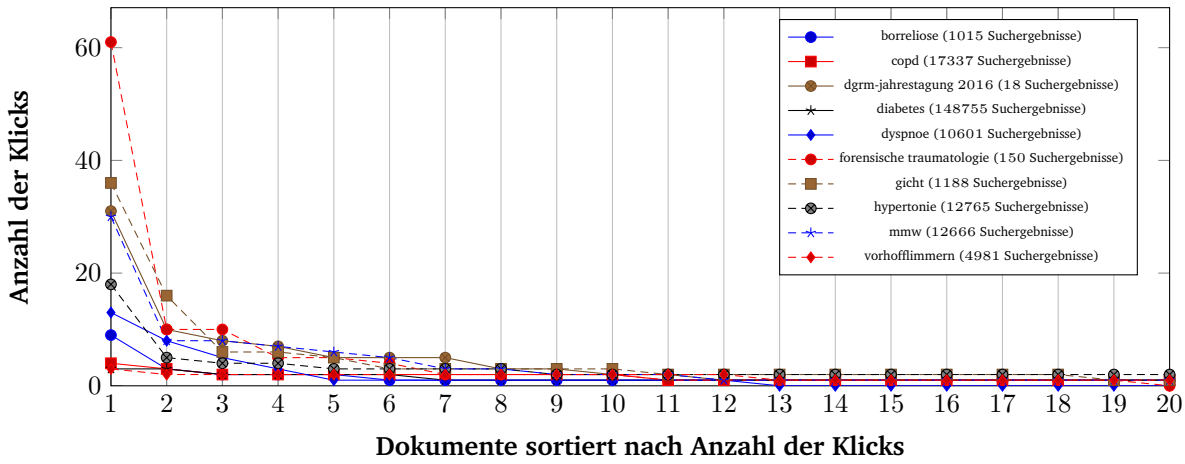

Abb. 2: Analyse der 20 am häufigsten angeklickten Dokumente der zehn meistgesuchten Suchphrasen. *Zeitraum der Analyse: 19.08.16 - 19.09.16*

Betrachten wir die Graphen, können wir vor allem für die ersten fünf analysierten Dokumente verglichen mit den restlichen analysierten Dokumenten, hohe Klick-Häufigkeiten feststellen. Daraus lässt sich die Vermutung ableiten, dass einzelne Dokumente eine sehr hohe Relevanz für die entsprechende Suchanfrage aufweisen und nur wenige Dokumente auf die User als relevant wirken. Ein weitere Vermutung ist, dass der zu durchsuchende Content wenig relevante Dokumente hat. Die Suchphrasen lassen auf sehr diverse Suchintentionen deuten. Es handelt sich hierbei unter anderem um Krankheiten, Zeitschriften und Behandlungen mit mehreren tausend Suchergebnissen. Die Wahrscheinlichkeit, dass wenig relevanter Content für die meisten der analysierten Suchphrasen zutrifft, sollte aufgrund der hohen Anzahl an gefundenen Suchergebnissen zu diesen Suchphrasen, relativ gering sein. Wir müssen darum eher davon ausgehen, dass sich die User auf einzelne im Suchresultat weit oben stehende Dokumente festfahren. Das könnte an schlechten Suchergebnissen und somit an einer schlechten Suchqualität liegen. Um jedoch ein genaueres Bild über das Verhalten erstellen zu können müssen wir einen Vergleich mit der nachfolgenden Analyse in Abbildung 3 ziehen.

Niedrige Positionen werden häufiger angeklickt

In der unten folgenden Analyse sehen wir das positionsbezogene Klick-Verhalten der User der Springermedizin-Suche. Dazu wurden über den Zeitraum von einem Monat, die letzten 1000 Suchanfragen ausgewertet. Dargestellt sehen wir die Häufigkeitsverteilung der Klicks als Graph. Wir beschränken uns hierbei auf die ersten 20 Positionen der Suchresultate. Wie wir sehen, nimmt die Anzahl der Klicks mit zunehmender Position exponentiell ab. Dieser Effekt kann ebenfalls, wie in Abbildung 2, durch das Potenzgesetz (Power Law, siehe [**PowerLaw**]) beschrieben werden.

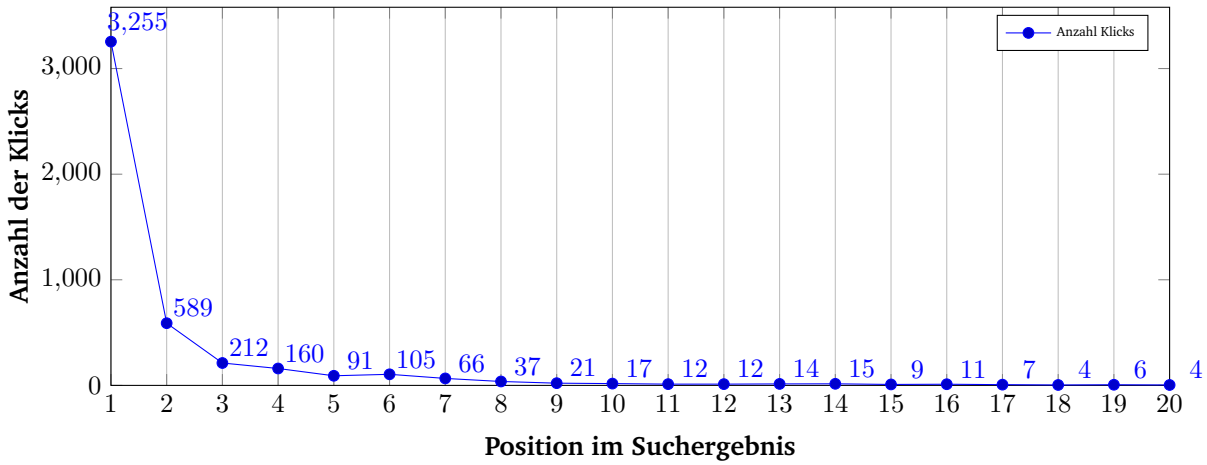

Abb. 3: Analyse der Klicks auf die ersten 20 Positionen der Suchergebnisse aller Suchanfragen.
Zeitraum der Analyse: 19.08.16 - 19.09.16

Betrachten wir den Graphen, sehen wir, dass besonders die erste Position auffällig oft angeklickt wird. Daraus könnten wir die Vermutungen ableiten, dass die Suche eine sehr gute Qualität besitzt, weil die zu oberst angezeigten Dokumente sehr relevant sind und die meisten User der Suchmaschine vertrauen. Wie wir aus den Analysen von [**Joachims**] lesen können, müssen wir davon ausgehen, dass die Häufigkeit des Klicks auf die ersten Positionen des Suchresultates eher dem Vertrauen der User der Suchmaschine, als der Qualität der Suche geschuldet ist. Vergleichen wir die Analyse aus Abbildung 2 mit dieser Analyse, sehen wir ein sehr ähnliches Muster in der Häufigkeitsverteilung der Klicks. Wir können anhand der Klick-Zahlen ebenfalls vermuten, dass die am häufigsten angeklickten Dokumente sich dabei auf den kleinsten Positionen befunden haben.

2.1.2 Userrelevanz mittels Click-Trough-Rate (CTR)

Um mit Click-Trough-Daten arbeiten zu können, müssen wir zuerst verstehen, was Click-Trough-Daten sind und wie sie entstehen.

Was sind Click-Trough-Daten und wie entstehen diese?

Click-Trough-Daten sind Tracking-Daten. Tracking-Daten entstehen durch die Interaktion zwischen dem User der Applikation und der Applikation selbst. Sie verfolgen das Verhalten der User auf der Applikation und speichern diese in einer Datenbank, in unserem Fall in Webtrekk ab. Die für uns interessanten Tracking-Daten entstehen, wenn der User auf der Suche von Springermedizin ein Anfrage stellt und darauf folgend, ein Element aus dem Suchresultat anklickt.

Wie werden die Click-Trough-Daten in Webtrekk gespeichert?

Die Speicherung der Daten auf Webtrekk übernimmt die Springermedizin-Applikation. Führt ein User eine Suche durch und klickt dabei ein Resultat an, sendet die Springermedizin-Applikation die Tracking-Informationen an Webtrekk. Die Tracking-Daten für diese Aktion, setzen sich zusammen aus der Suchanfrage, dem Zeitpunkt der Suche, den Userdaten, der angeklickten Position im Suchresultat und den Dokumentinformationen zum angeklickten Dokument.

Wie können wir Click-Trough-Daten aus Webtrekk lesen?

Webtrekk ist ein Analysetool. Das heißt für uns, wir können nicht direkt auf die Datenbank mit den Tracking-Daten zugreifen. Um die Tracking-Daten lesen zu können, müssen wir eine Analyse auf Webtrekk ausführen. Mithilfe dieser Analyse können wir uns die Click-Trough-Daten so zusammenstellen lassen, wie wir sie für die Berechnung der Click-Trough-Rate benötigen.

Die Click-Trough-Daten bestehen aus einzelnen *Klick-Häufigkeiten*. Eine Klick-Häufigkeit beschreibt die Anzahl der Klicks, die zu einer bestimmten Suchanfrage auf ein bestimmtes Dokument gemacht wurden und auf welcher Position im Suchresultat sich dieses Dokument dabei befunden hat. Die Webtrekk-Analysen geben uns eine Sammlung von Klick-Häufigkeiten zurück. Wir können bei diesen Analysen die Klick-Häufigkeiten nach Suchbegriffen oder auch Suchtermen filtern und den Zeitraum mitgeben, in welchen die Suchanfragen durchgeführt wurden. Des weiteren gibt es die Möglichkeit, weitere Filter wie die Anzahl zurückzugebender Klick-Häufigkeiten oder auch den „Login-Status[4] des Users“ zu setzen.

Wie sehen die Click-Trough-Daten aus?

Eine Beispiel für eine Klick-Häufigkeit wie er von einer Webtrekk-Analyse ausgespielt wird, sieht wie folgt aus:

Click-Trough-Daten	**Klick-Häufigkeit**
searchresult-1.Course.chronische Dyspnoe bei Erwachsenen.10621768.chronische Dyspnoe	5

Tab. 2.1: Beispiel Click-Trough-Daten

Hier die Aufschlüsselung der Click-Trough-Daten:

Position	**Dokumenttyp**	**Titel**	**ID**	**Suchterm**
searchresult-1	Course	chronische Dyspnoe bei Erwachsenen	10621768	chronische Dyspnoe

Tab. 2.2: Beispielhafte Aufschlüsselung der Click-Trough-Daten

Die Click-Trough-Daten lassen sich wie folgt lesen. In diesem Beispiel haben die User mit der Suchanfrage „chronische Dyspnoe[5]“ gesucht. Dabei haben sie das Dokument mit der ID 10621768 angeklickt. Dieses hat sich dabei auf der Position eins der Suchresultate befunden. Es wurde insgesamt fünfmal angeklickt in der gesuchten Periode.

Aus Merkmalen und Eigenschaften des Userverhaltens ein implizites Feedback bilden

Mit dem Tracking der User auf einer Suchmaschine verfolgen wir die Idee, ein implizites Feedback aus deren Verhalten interpretieren zu können. Das machen wir, indem wir Merkmale und Eigenschaften des Verhaltens lesen und daraus ein Feature-Set[6], wie in [**IWUSBI**] beschrieben erzeugen. Dieses Feature-Set setzt sich zusammen aus den Informationen des *Klick-Verhaltens* der User (Click-Trough Features) und deren *Browsing-Verhalten*[7] (Browsing Features) während einer Suchanfrage und den *semantischen*

[4] Mit Login-Status wird zwischen einem zum Zeitpunkt der Suche auf der Springermedizin-Applikation angemeldeten und nicht angemeldeten User unterschieden
[5] Als Dyspnoe wird eine unangenehm erschwerte Atemtätigkeit bezeichnet
[6] Mit Feature-Set bezeichnen wir eine Sammlung von Merkmalen und Eigenschaften zum Userverhalten auf der Suchmaschine
[7] Mit Browsing wird hier das Verhalten des Users bei der Navigation durch die Suche beschrieben

Relationen zwischen der Suchanfrage und den dazu ausgespielten Suchresultaten (Query-Text Features). Mithilfe des Feature-Set lassen sich dann Schlussfolgerungen zum Relevanzfeedback ziehen. Auf diesem Feature-Set werden wir bei der Auswertungen unserer Click-Trough-Daten aufbauen, um damit unsere Click-Trough-Rates zu berechnen.

2.1.3 Result-Reranking mittels PBM basierten Algorithmus

Alternative Ansätze um Click-Trough-Daten in den Suchprozess einzubinden

Kurzanalyse der möglichen Ansätze um Click-Trough-Daten in Suchprozess einzubinden Wir untersuchen in dieser Arbeit die Verwendung der Click-Trough-Daten in der Aufbereitung der Suchresultate der Springermedizin-Applikation. Es gibt aber auch andere mögliche Eingriffspunkte während des Suchprozesses, um die Click-Trough-Daten zu verwenden. Eine Alternative wäre die Verwendung der Click-Trough-Rate in der Aufbereitung der Suchanfrage auf der Springermedizin-Applikation. Denkbar wäre auch, die Berechnung der Click-Trough-Rate in den Suchindex der Solr einzubauen. Wir werden die verschiedenen Ansätze kurz durchgehen und am Ende erläutern, weshalb wir uns für den gewählten Ansatz mit dem PBM basierten Algorithmus entschieden haben.

Ansatz: Suchindex-Erweiterung in der Solr-Suche Um die Click-Trough-Rate direkt in die Solr einzubeziehen gibt es zwei Varianten. Wir können das *Schema des Suchindexes* über die Schema API (siehe [**SchemaAPISolr**]) erweitern (1) und alle Einträge neu indexieren, oder wir ergänzen den Index um ein *externes Feld* (ExternalFileField, siehe [**ExtFieldSolr**]) (2).

Beide Lösungsansätze ergeben nur bei der Speicherung einer einfachen *Click-Count Popularität*[8] Sinn. Diese genügen allerdings den hier gegebenen Anforderungen nicht, da die Click-Trough-Rate abhängig vom Suchterm ist. Der erste Lösungsansatz ist zudem besonders heikel, weil bei jeder Änderung des Click-Count-Wertes, das Dokument in der Solr neu indexiert werden.

Ansatz: Aufbereitung der Suchanfrage Die Solr-Suche bietet eine Boost-Funktion namens *DisMax Query Parser* (siehe [**DisMax**]). Mit dieser können basierend auf Feldwerten, einzelne Dokumente besser im Suchergebnis positioniert werden. Die Boost-Funktion müssten wir in den Aufbau der Suchanfrage für die Suche auf der Springermedizin-Applikation einbauen. Dieser Ansatz beinhaltet einige Gefahren die wir beachten müssen.

Dazu zählen beispielsweise die Abhängigkeiten von anderen *Boost-Faktoren*[9]. Alle Boost-Faktoren hängen voneinander ab und müssten bei jeder Ergänzung um neue Faktoren normalisiert werden, um kein „über-Boosting“[10] einzelner Faktoren zu riskieren. Zudem besteht die Gefahr des „blinden Boosting“ von Dokumenten. Die Solr-Relevanzberechnung ist komplex und der Einfluss des *Boosting* in die Solr-Relevanzberechnung schwer erkennbar. Auch hat Springermedizin bereits sehr schlechte Erfahrungen mit Boosting gemacht und bevorzugt einen Lösungsansatz ohne Boosting.

[8] Kennzahl für alle Klicks auf ein Dokument unabhängig des Suchterms
[9] Die Solr besitzt eine Boosting-Funktion, um bestimmte Wertübereinstimmungen in der Suche höher gewichtet zu können
[10] Bezeichnet die über-priorisierte Bewertung einzelner Faktoren

Der in dieser Arbeit verfolgte Ansatz: Aufbereitung der Suchresultate anhand eines Klick-Modell basierten Algorithmus

Wir verfolgen in dieser Arbeit den Ansatz der Aufbereitung der Suchresultate aus der Solr-Suche mithilfe des PBM basierten Algorithmus. Dieser soll die Suchergebnisliste analysieren, die Click-Trough-Rate der Dokumente berechnen und die Liste neu sortieren.

Mithilfe der Click-Trough-Daten aus Webtrekk, können wir zwei wichtige Informationen zu jeder Suchanfrage ermitteln. Wir wissen welches Dokument und welche Position im Suchresultat angeklickt worden ist. Zudem kennen wir die Reihenfolge der Dokumente im Suchresultat der Solr.

Der *Position-based Modell* (PBM) (siehe [**pbm**]) basierte Algorithmus baut genau auf diesen Click-Trough-Informationen auf. Es berechnet die Wahrscheinlichkeit dafür, dass ein User ein Dokument wirklich genau analysiert, bevor er es anklickt. Es setzt sich aus zwei Wahrscheinlichkeiten zusammen. Die Wahrscheinlichkeit für einen Klick auf die Position im Suchresultat und die Wahrscheinlichkeit für einen Klick auf das Dokument. Diesen Ansatz werden wir in dieser Arbeit implementieren.

Warum verwenden wir den PBM basierten Reranking-Algorithmus? Den PBM basierten Algorithmus können wir relativ einfach in die Springermedizin-Applikation integrieren, ohne die restliche Suchlogik[11] zu beeinflussen.

Wägen wir die besprochenen Fakten ab, wirkt der Ansatz mit der Aufbereitung der Suchresultate durch einen Klick-Modell basierten Algorithmus am sinnvollsten. Wir wissen bei diesem Ansatz, welche Dokumente für die Click-Trough-Rate-Berechnung überhaupt in Frage kommen. Zudem kennen wir alle Einfluss-Faktoren für den Algorithmus und wir sind unabhängig von der Suchlogik auf der Solr. Dadurch können wir Änderungen in unserer Logik schnell und einfach implementieren.

Grundlagen des Algorithmus

Worauf basiert unser Ansatz? Publikationen blablabla

Verwendete Formeln Formeln blablabla

2.2 Zusammenfassung

[11] Dazu gehört die Aufbereitung der Suchanfrage für die Solr und die Suche auf der Solr

3 Reranking mittels Click-Trough-Rate Ergebnis

3.1 Prozessaufbau des Lösungsansatzes

3.1.1 Prozessaufbau als Bild

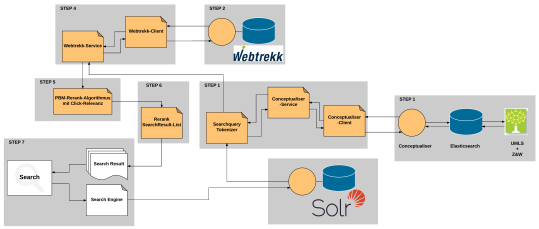

Abb. 4: Prozessaufbau des Lösungsansatzes

3.1.2 Probleme des Lösungsansatzes

3.1.3 Suchterm Segmentierung

3.1.4 Aufbereitung Click-Trough-Daten

Kein Einfluss auf Suchergebnisqualität während der Klicks

siehe Grundlagen

Userverhalten

Die Webtrekk-Analysen bieten uns nur beschränkte Informationen zum Klick-Verhalten der User. Wichtige Informationen wie die Verweildauer auf einem Dokument oder ob nach diesem Dokument ein weiteres Dokument zum gleichen Suchterm angeklickt worden ist, lassen diese Analysen nicht zu.

3.1.5 Click-Trough-Rate in Suchprozess einbinden

- Solr gebundene Suchresultatmenge - Pagination (Einfluss von Reranking)

3.1.6 Result-Reranking mittels PBM Algorithmus

- Smoothing

- Mehrfachverwertung von Content => mehrfache Auflistung in Suchergebnissen (nicht Teil dieser Arbeit) - Algorithmus nicht festfahren (Overfitting)

- User nicht festfahren auf altem Wissen - aktuelle und neue Publikationen werden nicht berücksichtigt => keine Userrelevanz (nicht Teil dieser Arbeit)

3.2 Methodik

In Kapitel 2.1 haben wir gelernt wie Click-Trough-Daten entstehen und wie sie zu lesen sind. Nun können wir mit diesem Wissen die Click-Trough-Rate der Dokumente berechnen. Mithilfe der berechneten Click-Trough-Daten werden wir dann ein *Reranking* der Suchresultate durchführen. So wollen wir die Userrelevanz in die Suche einbinden. Die Vorgehensweise dazu sieht wie folgt aus.

3.2.1 Suchterm Segmentierung

Suchterm semantisch aufschlüsseln mittels Segmentierung

Um alle relevanten Click-Trough-Daten lesen zu können, müssen wir zunächst den Suchterm auftrennen, wie in Kapitel 2.1.1 angesprochen. Die Auftrennung des Suchterms in die einzelne Worte können wir mithilfe einer Segmentierung[1] durchführen. Hier könnten wir uns überlegen, zusätzlich mit Stoppwörtern[2] nicht relevante Wörter aus dem Suchterm zu entfernen. Dieses Verfahren macht aber im Springermedizin-Kontext keinen Sinn. Wie in Kapitel 1.2 angesprochen, suchen die User der Springermedizin-Applikation oft mit einschlägig, fundierten Fachbegriffen. Wir gehen darum davon aus, dass alle Wörter des verwendeten Suchterms für das Suchergebnis relevant sind. Diese Erkenntnis basiert auf Aussagen der Redakteure von Springermedizin und Webtrekk-Analysen der meist gesuchtesten Suchtermen der letzten Monate. Auch sind Stoppwörter veraltet und werden in modernen Information Retrieval Verfahren nicht mehr eingesetzt. Wir verzichten darum auf den Einstatz von Stoppwörtern.

Suchterm semantisch erweitern mittels Thesaurus

Wie in Kapitel 2.1.1 thematisiert, wollen wir die Click-Trough-Daten zu unserem Suchterm um Click-Trough-Daten zu verwandten Begriffen erweitern. Für diese semantische Erweiterung eines Suchwortes werden wir einen Thesaurus verwenden. Die Erweiterung umfasst zum Suchterm gleichbedeutende Begriffe (Synonyme), sehr ähnliche Begriffe (Narrow Terms), ähnliche Begriffe im weiteren Sinne (Broader Terms) und verwandte Begriffe (Related Terms).

Springer Nature besitzt einen Webservice mit welchem auf den Thesaurus *Unified Medical Language System* (UMLS) (siehe [**UMLS**]) zugegriffen werden kann. Der Webservice nimmt einzelne Wörter und Wort-Listen entgegen. Zu jedem dieser Wörter durchsucht der Webservice den Thesaurus nach den oben erwähnten Arten von verwandten Begriffen. Der Webservice verwendet für diese Suche eine Elasticsearch (siehe [**elasticsearch**])[3]. Die dabei gefundenen Begriffe, liefert der Webservice als Antwort zurück. Um alle relevanten Click-Trough-Daten zu finden, werden wir mit dem segmentierten Suchterm eine Anfrage gegen diesen Webservice stellen und anschließend den segmentierten Suchterm um die

[1] Bezeichnet die Aufteilung in Abschnitte, in diesem Fall in einzelne Worte
[2] Stoppwörter sind Wörter, die sehr häufig auftreten und für gewöhnlich keine Relevanz für den Dokumentinhalt besitzen
[3] Eine Elasticsearch ist eine Volltextsuchmaschine

gefundenen Begriffe erweitern. Mithilfe des erweiterten Suchterms können wir dann anschließend eine Analyse in Webtrekk starten und alle relevanten Click-Trough-Daten lesen.

3.2.2 Aufbereitung Click-Trough-Daten

Jeder Klick auf ein Dokument ist relevant

Wie in Kapitel **??** beschrieben, reichen Webtrekk-Analysen für komplexe Auswertungen der Click-Trough-Daten nicht aus. Wir können darum in dieser Arbeit *Feedback-Strategien* für die Click-Trough-Rate Auswertung, wie in [**Joachims**] beschrieben, nicht verwenden. Stattdessen greifen wir wie ebenfalls in Kapitel **??** beschrieben auf die Click-Trough Features zu, die uns Webtrekk zur Verfügung stellt. Daraus entsteht ein interpretierbares Feature-Set. Dieses ist leider sehr klein und enthält keine Informationen um ein Relevanzfeedback zu den Klick-Häufigkeiten daraus lesen zu können. Wir müssen wir darum davon ausgehen, dass jeder Klick auf ein Dokument relevant ist.

Gewichtung der Click-Trough-Daten

Durch die semantische Aufschlüsselung des Suchterms haben wir verschieden starke Relationen zwischen Click-Trough-Daten und dem Suchterm. Die Gewichtung der Stärke dieser Relation ist aber nicht Kern dieser Arbeit. Wir gehen darum davon aus, dass unabhängig der stärke der Relation zum Suchterm, alle Click-Trough-Daten eine gleiche Relevanz besitzen.

Berechnung der Click-Trough-Rate

Die Click-Trough-Rate wird vor allem im Bereich des Internet-Marketing verwendet und stellt grundsätzlich die Anzahl der Klicks auf ein Dokument oder Link im Verhältnis zu den gesamten Impressionen dar. Bezogen auf das in Kapitel **??** angesprochene Feature-Set, würden die *ClickProbability* (Klick-Wahrscheinlichkeit) direkt als Click-Trough-Rate verwenden. Dazu müssten wir nur die Click-Trough-Daten eines Dokuments ins Verhältnis zu allen Click-Trough-Daten für einer Suchanfrage stellen. Wie wir aber bereits in Kapitel 2.1.1 gelernt haben, würden wir damit viele Problemstellungen der Interaktion der User mit der Suche ignorieren. Deswegen verwenden wir den Position-Based Modell basierten Algorithmus um mithilfe dieses angesprochenen Feature-Sets die Click-Trough-Rate zu berechnen.

3.2.3 Result-Reranking mittels PBM basiertem Algorithmus

Klick-Wahrscheinlichkeit mit Position-based Modell berechnen

Wie in Kapitel 2.1.3 angesprochen, werden wir unseren Reranking-Algorithmus in die Aufbereitung der Suchresultate aus der Solr-Suche integrieren.

Dieser soll die Suchergebnisliste analysieren, die Click-Trough-Rate der Dokumente berechnen und die Liste neu sortieren.

Wir müssen jedoch beachten, dass die Solr durch die Pagination-Funktion (siehe [**Pagination**]) nur die Top-N-Ergebnisse (bei Springermedizin sind es 20 Ergebnisse) zurückgibt. Dadurch sehen wir nur einen Teil der Suchergebnisse.

Diese Logik liegt in der Springermedizin-Applikation im Aufbau der Suchanfrage. Daher können wir diese selber steuern und uns statt 20 beispielsweise die nächsten 100 Ergebnisse zurückgeben lassen. Am Ende filtern wir die ersten 20 Ergebnisse und stellen diese dar. Außerdem wissen wir bei diesem Lösungsansatz, in welcher Reihenfolge die Ergebnisse aus der Solr zurückgegeben werden. Wir kennen die Dokumente und deren Rang. Dadurch haben wir hilfreiches Zusatzwissen, welches wir in den Klick-Modell basierten Algorithmus einfließen lassen können.

Verhältnis zwischen den Klick-Wahrscheinlichkeiten abhängig der Position im Suchresultat definieren

Aus eigener Erfahrung wissen wir, dass die ersten Dokumente im Suchresultat immer zuerst gesehen werden. Die dahinter gelisteten Dokumente werden fortlaufend analysiert. Dies bestätigt die in Abb. 3 dargestellte Analyse der Klicks auf die ersten 20 Positionen eines Suchergebnisses. Wir sollten darum darauf achten, dass je *schlechter* der Rang des angeklickten Dokumentes im Suchresultat der Solr ist, desto *höher* das Relevanzfeedback zu bewerten ist.

Das machen wir, indem wir für die Berechnung der Click-Trough-Rate das Verhältnis zwischen Klick-Wahrscheinlichkeit der Position und Klick-Wahrscheinlichkeit des Dokumentes abhängig der Position im Suchresultat definieren. Als Grundlage hierbei dient uns die Position des Suchresultats der Solr. In der folgenden Tabelle sehen wir die Aufteilung der Verhältnisse abhängig der Position:

Position	Verhältnis Position zu Dokument
1 bis 10	1:1
11 bis 20	1:2
größer 20	1:3

Tab. 3.1: Verhältnis Klick-Wahrscheinlichkeiten der Position zu der des Dokumentes

Für die Suchresultate mit einer Position über 20, verstärken wir die Gewichtung der Klick-Wahrscheinlichkeit des Dokumentes erheblich. Das liegt daran, dass bei Klicks auf Dokumente mit einer solch hohen Position wir davon ausgehen können, dass die suchende Person die Suchresultate genau analysiert hat, bevor sie ein Dokument angeklickt hat. Haben wir die Verhältnisse definiert, müssen wir diese in den Algorithmus einbauen. Wie wir dies machen, werden wir im folgenden Abschnitt anschauen.

Smoothing Faktor in Position-based Modell

Wir wissen dass eine Wahrscheinlichkeit einen Wert zwischen 1 und 0 besitzt. Dadurch können Nullwerte entstehen. Das PBM multipliziert die Positions- und Dokument-Wahrscheinlichkeit miteinander, um die Klick-Wahrscheinlichkeit zu berechnen. Wir müssen aber davon ausgehen, dass es Dokumente geben kann, deren Rang nie angeklickt worden ist und umgekehrt.

Multiplikationen mit Null ergeben immer einen Nullwert. An dieser Stelle führen wir einen *Smoothing-Faktor* ein. Der Smoothing-Faktor soll zwei Probleme lösen. Zum einen wollen wir einen Wahrscheinlichkeitswert trotz der Multiplikation mit Null beachten. Zum anderen wollen wir die im vorherigen Absatz beschriebene Gewichtung abhängig des Relevanzfeedbacks in den Algorithmus einbeziehen. Wir transformieren dazu das Produkt der beiden Wahrscheinlichkeiten in eine gewichtete Summe, dem sogenannten *Weighted Moving Average* (siehe [**weightedAVG**]), dessen Gewichte sich zu Eins aufsummieren. Diese Gewichte sind die Smoothing-Faktoren, weshalb das Verfahren zählt zu den Smoothing-Algorithmen zählt.

3.2.4 Vergessen der alten Daten

Ein Algorithmus zur Berechnung von Wahrscheinlichkeiten muss sich ein gewisses Grundwissen aneignen. Dies geschieht üblicherweise durch Trainingsdaten. Genauso muss er alte Daten wieder vergessen können, um Overfitting[4] zu vermeiden.

Durch Webtrekk ist kein komplexer Lern-Algorithmus notwendig

Durch Webtrekk haben wir eine Wissensbasis, die sich stetig und zeitnah aktualisiert. So muss der Algorithmus nicht stetig neues Wissen lernen und altes vergessen, sondern er kann direkt diese Wissensbasis zugreifen. Dies geschieht, indem zur Laufzeit[5] Analysen gegen Webtrekk über eine frei definierbare Periode gemacht werden. Dadurch kann *Overfitting* vermieden werden. Deshalb verwenden wir keinen komplexen Lern-Algorithmen wie in [**IWUSBI**] vorgestellt.

Die Klick-Wahrscheinlichkeit ist kein absoluter Wert für die Userrelevanz

Nun könnten wir die Klick-Wahrscheinlichkeit als absoluten Wert für die *Userrelevanz* betrachten. Dies wäre jedoch falsch, wie in Kapitel 2.1.1 analysiert, müssen wir davon ausgehen, dass viele User der Qualität der Suchmaschine vertrauen. Diese betrachten die Top-Suchresultate als die relevanten Suchresultate. Denkbar wäre auch, dass User unabsichtlich das falsche Dokument anklicken und dadurch die Click-Trough-Rate eines Dokumentes verfälschen. Dadurch kann ein *Overfitting* des Algorithmus entstehen.

Overfitting vermeiden

Um ein Overfitting zu vermeiden, darf der Algorithmus nicht immer anschlagen. Wir müssen sicherstellen, dass vereinzelt zufällige Dokumente in den „Top-Suchresultaten“ angezeigt werden. So können auch andere Dokumente in den Fokus des Users gerückt werden. Das System fährt sich dadurch nicht auf falschen Annotationen fest.

Zusätzliche Varianz durch Zufallsfaktor

Mithilfe eines Zufallsfaktors kann eine solche Varianz in den Klick-Modell basierten Algorithmus gebracht werden. Wie bereits weiter oben erwähnt, werden viele Suchresultate nie und deren Rang selten bis gar nicht angeklickt. Sie haben darum keine Click-Trough-Daten. Deren Klick-Wahrscheinlichkeit ist entweder Null oder sehr klein. Der Zufallsfaktor soll darum nur leichte Einflüsse in die Klick-Wahrscheinlichkeitsberechnung haben. Auch hier können wir wieder mit dem oben eingeführten *Weighted Moving Average* arbeiten.

3.3 Der PBM-Algorithmus

3.4 Zusammenfassung

[4] Überanpassung des Algorithmus durch zu viele (falsche oder veraltete) Daten
[5] Unter Laufzeit wird in diesem Fall der Zeitpunkt der direkte Abfrage während der Suchanfrage bezeichnet

4 Implementierung

5 Evaluation und Auswertung

5.1 Einführung

Suchvarianten mithilfe eines Evaluationssystems vergleichen

Das große Kernproblem der Überprüfung der Verbesserungen durch den untersuchten Lösungsansatz wird das Messen der Qualität der erzielten Suchergebnisse sein. Mithilfe einer Evaluation wollen wir messen, wie gut die Suchergebnis-Qualität der aktuellen Springermedizin-Suche im Vergleich zur im Zuge dieser Arbeit entwickelten Lösung ist.

Ziel der Evaluation

Die Evaluation soll Informationen darüber liefern, wie viel Verbesserung der neue Lösungsansatz bringt. Aus den Ergebnissen wollen wir erkennen, an welchen „Schrauben“ etwas gedreht werden muss, damit die Suche wirklich gute Ergebnisse aus Sicht der User bringt.

5.2 Aufbau der Analyse

5.2.1 Datengrundlage

Filterung der nutzbaren Daten mittels Cohens Kappa

Um die Zuverlässigkeit der Relevanzbewertungen zu messen, werden wir die gleichen Suchterme jeweils von zwei fachlichen Experten bewerten lassen. Das meist verwendete Maß zur Bewertung der Übereinstimmungsgüte ist der *Cohens Kappa Koeffizient* (siehe [**Kappa**]). Diese Zahl misst den Anteil übereinstimmender Bewertungen. Hierbei können aber auch zufällige Übereinstimmungen entstehen. Der Cohens Kappa Koeffizient korrigiert das Maß an Übereinstimmung um diesen Zufallsfaktor. Anhand der Auswertungen werden wir ein Mindestmaß der Übereinstimmungsgüte definieren. Die darunter liegenden Bewertung werden wir in der Auswertung ignorieren.

5.2.2 Metrik

Evaluationsdaten mittels NDCG-Algorithmus auswerten

Um das Qualitätsmaß der beiden Suchen vergleichen zu können werden wir den Bewertungsalgorithmus *NDCG* (siehe [**NDCG**]) einsetzen. Dieser geht davon aus, dass besser positionierte Suchergebnisse eine höhere Relevanz als schlechter positionierte haben. Der NDCG vergleicht die Reihenfolge der Relevanzbewertungen der Suchergebnisse mit der idealen Reihenfolge derselben Relevanzbewertungen. Im Idealfall entspricht die Reihenfolge der Suchergebnisse der Relevanz der Suchergebnisse.

Qualitätsmaß einer Suchvariante bestimmen

In der Evaluation werden zu jedem Suchterm zwei Bewertungen für die Springermedizin-Suche und zwei Bewertungen für die Suche mit dem hier zu untersuchenden Lösungsansatz abgegeben. Um das Qualitätsmaß einer Suchvariante zu einem Suchterm zu bestimmen, berechnen wir den NDCG der beiden Bewertungen. Nehmen wir den Mittelwert der beiden resultierenden NDCG-Werte, erhalten wir den effektiven NDCG-Wert. Die NDCG-Werte der beiden Suchen können wir dann miteinander vergleichen.

5.2.3 Vorgehen

Evaluationssystem aufbauen

Um eine Evaluation durchführen zu können, müssen wir eine passende Testumgebung aufbauen. Diese besteht aus einem Evaluationssystem, einer Instanz der aktuellen Springermedizin-Applikation und einer Instanz des neu implementierten Lösungsansatzes. Auf dem Evaluationssystem sollen fachliche Experten (Redakteure von Springermedizin) die Relevanz der Suchergebnisse der beiden Suchmaschinen vergleichen. Dazu sollen die jeweils besten zehn Suchergebnisse nach Relevanz zum Suchterm bewertet werden. Der Ergebnisse werden in einer Datenbank gespeichert, um sie später auszuwerten.

Evaluationssystem auswerten

Nach Ablauf der Evaluationsphase werden wir die Evaluations-Daten auswerten. Die Auswertung der Daten findet direkt im Evaluationssystem statt.

Dazu werden die Daten aus der Datenbank gelesen und mit dem Cohens Kappa Koeffizienten die nutzbaren Daten gefiltert.

5.2.4 Durchführung

Verschiedene Varianten des neuen Lösungsansatzes werden evaluiert

Der in dieser Arbeit zu untersuchende Lösungsansatz kann verschieden konfiguriert werden. Wir können den Einfluss des Zufallsfaktors bestimmen. Um verschiedene Konstellationen testen zu können, werden wir mit zwei verschiedenen Werten für den Einfluss des Zufallsfaktor evaluieren. Die Click-Trough-Daten von an der Applikation angemeldeten Benutzern können wir von den Click-Trough-Daten von anonymen Benutzern unterscheiden.

Aus den beiden Einflusswerten des Zufallsfaktors und der Unterscheidung zwischen angemeldeten und anonymen Benutzern, ergeben sich vier Konstellationen, die evaluiert werden können. Jeder Konstellation werden wir jeweils 25 Prozent der Suchterme zuteilen. Mithilfe des Evaluationssystems werden wir die Zuteilung der Suchterme zufällig generieren lassen.

Hier folgend die Aufteilung der generierten Analysen:

- Springermedizin.de Suche
  - 50% aller Analysen

- Reranking Suche
  - CTR-Daten der angemeldeten User: 25% aller Analysen
    - Einfluss Zufallsranking 0.1: 50% dieser Analysen
    - Einfluss Zufallsranking 0.01: 50% dieser Analysen
  - CTR-Daten aller User: 25% aller Analysen
    - Einfluss Zufallsranking 0.1: 50% dieser Analysen
    - Einfluss Zufallsranking 0.01: 50% dieser Analysen

5.3 Auswertung der Suchergebnis-Qualität

5.3.1 Quantitative Auswertung

5.3.2 Diskussion

5.4 Zusammenfassung

6 Zusammenfassung und Ausblick

Anhang

7

Abbildungs-Verzeichnis

Tabellen-Verzeichnis