Messung des Nutzens semantischer Suche

Dr. Thomas Hoppe

Datenlabor Berlin, Email: thomas.hoppe@datenlabor.berlin und

Ontonym GmbH, Berlin, Email: thomas.hoppe@ontonym.de

Horst Junghans

EUROPUBLIC GmbH Werbeagentur, Berlin, Email: Junghans@europublic.de

<Kasten Kernaussagen / Empfehlungen>

1. Precision und Recall sind zur Beurteilung von Suchfunktionen durch die Rahmenbedingungen im Unternehmenskontext ungeeignet.
2. Zur quantitativen Bewertung von Effizienzsteigerungen wird ein neues Maß benötigt.
3. Semantische Suche liefert einerseits weniger, dafür aber genauere Treffer, andererseits aber auch zusätzliche Treffer.
4. Die Bewertung einer semantischen Suche muss diesen vermeintlichen Widerspruch auflösen.
5. Zwei empirische Vergleiche zeigen, dass durch semantische Suche Effizienzsteigerungen seitens des Benutzers in der Größenordnung von 10-15% möglich sind.

<Kasten Ende>

Zusammenfassung Der Nutzen semantischer Technologien wird in der Regel durch qualitative Aussagen beschrieben: welche Effekte semantische Technologien im praktischen Einsatz haben oder welche Gewinnzuwächse bzw. Einsparungen Unternehmen erzielten, die semantische Technologien bereits eingesetzt haben. Unternehmensintern helfen solche Aussagen, um Entscheider dazu zu bewegen, sich mit der Frage des Einsatzes semantischer Anwendungen auseinanderzusetzen. Für die Ermittlung des konkreten Nutzens einer geplanten semantischen Anwendung sind sie jedoch in der Regel kaum hilfreich. Um einen betriebswirtschaftlichen „return of investment“ herleiten zu können, müssen oft Hypothesen über erzielbare Gewinn- oder Effizienzsteigerungen herangezogen werden, denen es an einer ausreichenden Fundierung fehlt. Auf diesen Hypothesen

aufbauende ROI-Betrachtungen stehen auf wackligen Beinen und Entscheidungen, die auf ihnen aufbauen, basieren eher auf Hoffnung und Glauben als auf gesicherten Erkenntnissen. In diesem Kapitel stellen wir eine Vorgehensweise vor, mit der verlässliche quantitative Aussagen über den Nutzen von Suchtechnologien durch den direkten Vergleich ihrer Suchergebnisse möglich werden. Wir wenden diese Vorgehensweise zum Vergleich einer semantischen Suche mit zwei unterschiedlichen, konventionellen Volltextsuchen an und zeigen welche konkreten Effizienzsteigerungen durch den Einsatz einer Thesaurus-basierten semantischen Suche gegenüber Volltextsuchen erzielbar sind.

Motivation

Eine kleine Eingangsfrage: Haben Sie Kinder oder können Sie sich noch an Ihre Schulzeit erinnern? Dann wissen Sie, dass spätestens bei den Abschlusszeugnissen und evtl. beim Antritt einer Ausbildung oder eines Studiums die Frage nach dem Notendurchschnitt eine wichtige Rolle spielt. Sagt dieser Notendurchschnitt aber etwas über spezifische Fachkenntnisse Ihrer Kinder oder von Ihnen aus?

Bei einem Notendurchschnitt von 1,3 ist die Wahrscheinlichkeit, dass Ihr Kind oder Sie gut in den Naturwissenschaften sind, zwar hoch, sie ist aber nicht garantiert. U.U. ist ein Mitschüler, der nur einen Schnitt von 2,0 hat, viel besser in den entsprechenden Fächern. Der Notendurchschnitt ist eben nur ein Durchschnitt und gibt kaum Auskunft über spezifische Kenntnisse, Fähigkeiten oder Fertigkeiten. Wir kommen später nochmals auf dieses Beispiel zurück, lassen Sie uns jetzt zum eigentlichen Thema dieses Kapitels kommen.

Bewertung von Information Retrieval Systemen

Im Information Retrieval sind *Recall*, *Precision* und *F-Maß* die Maße der Wahl, wenn es darum geht Information Retrieval Systeme zu bewerten [1,2]. Diese Maße werden herkömmlicherweise auch zur Bewertung von Suchmaschinen herangezogen, um die Relevanz der Suchergebnisse, den auf Anfragen zurückgelieferten Dokumenten und damit das Suchverfahren selber zu beschreiben und zu beurteilen.

Unter dem *Recall* versteht man den Anteil der korrekt als positiv (relevant) erkannten Ergebnisse an der Gesamtheit der tatsächlich positiven (relevanten) Ergebnisse. Der Recall beschreibt somit, wie viel Prozent der relevanten Ergebnisse erkannt wurden. Die *Precision* beschreibt den Anteil der korrekt erkannten, positiven (relevanten) Ergebnisse an allen korrekt erkannten Ergebnissen. Sie beschreibt somit, wie groß die Erkennungsgenauigkeit ist. Sowohl Recall als auch Precision

sind nicht unabhängig voneinander und werden durch das *F-Maß* zu einem einzigen Wert kombiniert.[1]

Die mathematischen Details der Berechnung dieser Werte können in der einschlägigen Literatur nachgelesen werden (beispielsweise [1]). An dieser Stelle interessiert uns der Prozess, mit dem diese Werte bestimmt werden, insbesondere in Hinblick auf die Nutzung dieser Maße im Unternehmenskontext im Allgemeinen und zur Beurteilung von semantischen Suchverfahren im Besonderen.

Relevanz, Textkorpora und Goldstandards

Um die Basismaße Recall und Precision bestimmen zu können, die auf die Messung der *Relevanz von Ergebnissen* abzielen, muss natürlich festgelegt werden, was denn die *Relevanz eines Ergebnisses* ist.

Generell kann die Relevanz eines Retrieval-Ergebnisses (i.E. eines gefundenen Dokuments) nur bezogen auf das konkrete Informationsbedürfnis eines Benutzers festgelegt werden; sie ist damit eine subjektive Beurteilung. Natürlich können bei einer Bewertung eines Systems niemals die Informationsbedürfnisse aller Benutzer berücksichtigt werden, sie sind weder bekannt noch verfügbar. Bei der konventionellen Bewertung von Informationssystemen werden daher repräsentative Anfragen (im Kontext von TREC auch als *Topics* bezeichnet[2]) einer Teilmenge der Benutzer im Voraus zusammengetragen.

Um zu einer korrekten Bewertung eines Informationssystems zu gelangen, müsste natürlich die Relevanz aller Dokumente des Anwendungsgebiets hinsichtlich dieser Anfragen bewertet werden. Je nach Umfang der Dokumentenmenge ist dies aus Aufwandsgründen nicht möglich, so dass für die Bewertung nur eine Teilmenge der Dokumente des Anwendungsbereichs herangezogen werden kann. Da die Bewertung weiterer Dokumente zu einem späteren Zeitpunkt Zusatzaufwände nach sich zieht, wird die bewertete Dokumentenmenge in der Regel einmalig festgelegt und hinsichtlich der Relevanz bewertet. Sie ist somit statisch.

Die Relevanzbewertung der Dokumententeilmenge hinsichtlich der Teilmenge der festgelegten Anfragen erfordert es, dass diese Bewertung durch Menschen vorgenommen wird. Da – wiederum aus Aufwandsgründen – nicht alle Benutzer eines Informationssystems diese Bewertung durchführen können und die Bewertung durch einzelne Personen stark subjektiv wäre, muss eine repräsentative, halbwegs objektive Bewertung der Relevanz von einer kleinen Benutzergruppe vorgenommen werden, deren Einzelbewertungen miteinander verrechnet werden.

Zusammengefasst: Um zu einer Relevanzbewertung einer Teilmenge der Dokumente eines Anwendungsgebiets hinsichtlich einer Teilmenge aller Anfragen zu

---

[1] Für eine ausführliche, anschauliche Darstellung siehe z.B. http://de.wikipedia.org/wiki/Recall_und_Precision#Anwendung_im_Information_Retrieval

[2] TREC (Text REtrieval Conference) ist eine Konferenzreihe, http://trec.nist.gov/

gelangen, muss eine kleine Menge von Personen (die als Teilmenge aller Benutzer betrachtet werden kann) jedes der Dokumente hinsichtlich seiner Relevanz bzgl. der festgelegten Anfragen beurteilen. Diese bewertete Dokumententeilmenge wird in der Regel als *Textkorpus* bezeichnet und definiert zusammen mit der Relevanzbewertung einen *Goldstandard,* mit dem unterschiedliche Verfahren verglichen werden können.

Offensichtlich erzeugt eine solche Relevanzbewertung Aufwand, der nicht für jedes Anwendungsgebiet betrieben werden kann. Für ein neues Anwendungsgebiet ist somit nicht garantiert, dass ein entsprechender Textkorpus und damit der entsprechende Goldstandard existiert.

Für die Bewertung von Informationssystemen ergibt sich damit, dass lediglich eine begrenzte Menge Textkorpora für eine begrenzte Menge von Anwendungsgebieten zur Verfügung stehen, für die die Relevanz bzgl. einer begrenzten Menge von repräsentativen Anfragen von einer Teilmenge von Benutzern bewertet wurde.

Rahmenbedingungen des Unternehmenskontextes

Entscheidungen im Unternehmenskontext haben in der Regel betriebswirtschaftliche Konsequenzen, einerseits was Kostenfaktoren, andererseits was Gewinne betrifft. Um sich zwischen mehreren gleichartigen Technologien zu entscheiden oder eine existierende Technologie durch eine neue abzulösen, insbesondere wenn diese auf einer neuen, noch unbekannten technologischen Grundlage basiert, werden fundierte Entscheidungsgrundlagen benötigt. Neben den Vorteilen, die eine neue Technologie bietet, muss daher ihr Nutzen quantifiziert werden können, so dass eine rationale Investitionsentscheidung möglich wird. Unter Umständen müssen dabei – wenn es sich um den Vergleich von funktional äquivalenten, auf unterschiedlichen Technologien basierenden Systemen handelt – auch mal Äpfel mit Birnen verglichen werden, wobei es nicht so sehr auf subjektive Faktoren wie Farbe, Form und Geschmack ankommt, sondern objektivierbare Vergleichsparameter gefragt sind, wie Nährwert, Kosten und Mengen.

Um Aussagekraft für die spezifischen Anforderungen eines Unternehmens zu besitzen, ist es wesentlich, dass ein solcher Vergleich möglichst auf den Dokumentenbeständen des Unternehmens oder in Ausnahmefällen zu mindestens auf Dokumentenbeständen des gleichen Anwendungsbereichs erfolgt. Darüber hinaus sollten Anfragen die für das Unternehmen typischen Benutzerinteressen reflektieren, um ein glaubwürdiges Vergleichsszenario zu schaffen.

Offensichtlich wird kaum einer der existierenden Goldstandards diesen Anforderungen gerecht. Selbst wenn ein Textkorpus des fraglichen Anwendungsbereichs existiert, ist es extrem unwahrscheinlich, dass die für das Korpus genutzten Anfragen Relevanz für das Unternehmen besitzen und gleichzeitig die

Relevanzbewertung der Dokumente bzgl. der Anfragen das Informationsbedürfnisse der Unternehmensnutzer widerspiegelt.

Die Nutzung existierender Goldstandards zum Vergleich von Information Retrieval Systemen ist daher kaum möglich, noch lässt es der Unternehmenskontext zu – bedingt durch den damit verbundenen Aufwand – einen eigenen Goldstandard für einen solchen – unter Umständen einmaligen – Vergleich aufzubauen.

Rahmenbedingungen bei der Bewertung von Suchmaschinen

Sie kennen es, die Ergebnisse von Suchmaschinen werden in der Regel in einer linearen Anordnung sortiert angezeigt, wobei die Sortierung auf einem komplexen Ordnungskriterium basiert, welches „Relevanz“, „Aktualität“, „Wichtigkeit“ oder einen vom „Kunden bezahlten Preis für die Position seiner Information“ berücksichtigt. Haben wir als Benutzer ein spezielles Informationsbedürfnis, inspizieren wir die Treffer in der angezeigten Reihenfolge mehr oder weniger intensiv, in der Regel aber in ihrer linearen Anordnung, wobei wir ggf. auf Suchergebnisfolgeseiten wechseln, so lange, bis wir entweder den für unser Informationsbedürfnis passenden Treffer finden oder die Suche abbrechen – mit der Schlussfolgerung „dazu gibt es nichts“.

Die Maße Recall, Precision und F-Maß nehmen auf die Anordnung der Suchergebnisse keine Rücksicht, sie messen lediglich den Durchschnitt über alle Ergebnistreffer, ohne diese Menge in Blöcke gleicher Anzahl zu zerlegen und sie via Paginierungsmechanismen auf jeweils separaten „Search Engine Result Pages“ (SERP) anzuzeigen, zwischen denen der Benutzer navigieren kann.

Betreiber von Suchmaschinen haben anhand des Benutzerverhaltens herausgefunden, dass Benutzer hierbei nicht alle SERPs inspizieren. Während die erste SERP noch von 68% aller Benutzer inspiziert wird, inspizieren 17% noch die zweite SERP. Lediglich 7% aller Benutzer inspizieren auch noch die dritte SERP. Die folgenden SERPs werden so selten inspiziert, dass die Suchergebnisse auf diesen Seiten als quasi nicht existent betrachtet werden können [3,4,5,6].[3]

Darüber hinaus sind Recall und Precision lediglich Durchschnittswerte – Sie erinnern sich an das Notendurchschnittsbeispiel vom Anfang des Kapitels – mit denen keine detaillierteren Aussagen über das Verhalten bei bestimmten Anfrageklassen oder Verarbeitungsarten getroffen werden können.

Semantische Suche

[3] Merke: „the best place to hide a dead body is on page two of Google’s search results“ (http://digitalsynopsis.com/tools/google-serp-design/)

Wenn wir im Folgenden von semantischer Suche reden, ist hiermit immer eine Thesaurus-basierte semantische Suche gemeint, die Hintergrundwissen über ein Anwendungsgebiet in Form eines Thesaurus oder einer Ontologie verwendet, um die auf Anfragen passenden Suchergebnisse (im weitesten Sinn alle Formen von Dokumenten) zu ermitteln. Zweckmäßigerweise werden diese Suchergebnisse in der Regel nach ihrer „Passgenauigkeit“ auf die Suchanfrage geordnet. Wie konkret diese „Passgenauigkeit“ ermittelt wird, ist an dieser Stelle irrelevant. Wir stellen an das entsprechende Ordnungskriterium lediglich die Anforderung, dass die am besten passenden Ergebnisse vor den weniger passenden Ergebnissen präsentiert werden.

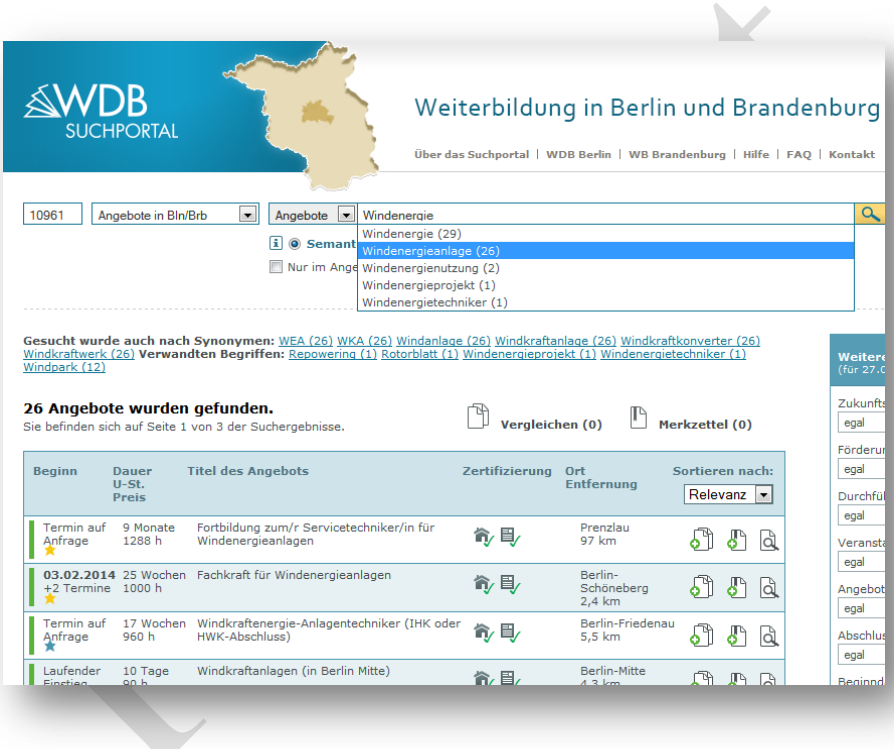

Abb. Y.1: Semantische Suche nach Weiterbildungsangeboten

Als Beispiel für eine solche Thesaurus-basierte semantische Suche sei hier die semantische Suche im Suchportal der Weiterbildungsdatenbank Berlin-Brandenburg genannt, die Ontonym in das von EUROPUBLIC im Auftrag des Berliner Senats und der Landesregierung Brandenburgs entwickelte gemeinsame Suchportal integriert hat.[4] Der verwendete Thesaurus über recruitment- und

[4] Die Entwicklung des WDB Suchportals wird gefördert aus Mitteln des Europäischen Sozialfonds und der Länder Berlin und Brandenburg. "Investition in die Zukunft" – Ein Angebot der Weiterbildungsdatenbanken Berlin und Brandenburg.

weiterbildungsspezifische Themen, der von Ontonym modelliert wurde und weiter gepflegt wird, umfasst derzeit[5] 9.470 in einer Ober-/Unterbegriffshierarchie angeordnete Konzepte mit 14.680 Bezeichnungen und rund 5.290 weitergehende, nicht-hierarchische Beziehungen. Durch automatische Mechanismen werden aus den Bezeichnungen 58.635 unterschiedliche Schreibvarianten (getrennt geschrieben, mit und ohne Bindestriche, zusammen geschrieben, mit und ohne Abkürzungen und mit unterschiedlichen Umlautschreibweisen) generiert, mit denen durch Wordstammbildung eine nicht näher bestimmbare, mindestens 6-stellige Anzahl von Schreibweisen in den Weiterbildungsangeboten erkannt werden kann.

Abbildung Y.1 zeigt eine typische Anfrage zu Weiterbildungsangeboten aus dem Bereich „Erneuerbare Energien“. Neben der Thesaurus-basierten, semantischen Autocompletion, die den Benutzer bei der Eingabe von Suchanfragen unterstützt, in dem potentielle Begriffsverfeinerungen anhand des Thesaurus vorgeschlagen und die Anzahl der potentiell passenden Weiterbildungsangebote angezeigt werden, zeigt dieser Bildschirmausschnitt darüber hinaus die zusätzlichen Begriffe, die bei der Suche berücksichtigt wurden. Dies sind Synonyme, Unterbegriffe und Begriffe, die zu der Anfrage in enger Beziehung stehen.

Mehr Ergebnisse

Offensichtlich liefern semantische Anwendungen dadurch, dass sie die Bedeutung einer Anfrage anhand von Hintergrundwissen ermitteln, neben Treffern, die allein über die Anfrage findbar sind (Treffer 1 und 2 in Abb. Y.1, die die Bezeichnung „Windenergieanlage“ direkt enthalten), auch Treffer die ähnliche Bezeichnungen oder Synonyme verwenden (Treffer 3 und 4, die die Synonyme „Windenergietechniker“ und “Windkraftanlage“ enthalten, siehe hierzu auch Abb. X3 in Kapitel 8). Diese zusätzliche Intelligenz verschafft semantischen Anwendungen einen Mehrwert. Benutzer werden davon entlastet, andere Bezeichnungen selber zu (er-)finden und Anfragen mehrfach zu stellen. Darüber hinaus werden die Einzelergebnisse der Anfragen zu einem einzigen Gesamtergebnis integriert.

Konventionelle Volltextsuchen hingegen probieren ähnliche Effekte durch rein syntaktische Mechanismen zu erzielen, wie z.B. Suche über Wortstämmen, Prefix- oder Wildcard-Matching (bei dem alle Bezeichnungen erkannt werden, die mit dem angefragten Term beginnen) oder Substring-Matching (bei dem auch alle Bezeichnungen gefunden werden, in denen der angefragte Term irgendwo auftritt).

[5] Stand 15.2.2014

Genauere Ergebnisse

Für jeden Suchenden besteht die Crux reiner Volltextsuche im Allgemeinen jedoch aus den nicht-passenden Treffern. Wenn beispielsweise zu einer Suche nach einer Stelle als „Maler“ von einem Stellenportal, dessen Suche sich Wortstammbildung und Substring-Matching bedient, ein Job als Finanzberater vorgeschlagen wird, weil dieser „Beratung hinsichtlich opti*maler* Anlagestrategie“ geben soll, oder für jemanden, der eine Stelle als „MTA“ sucht, Stellen als Informatiker angezeigt werden, weil in diesen die Rede von einer „Gesa*mta*rchitektur“ ist, dann sind diese Treffer weder „passgenau“ noch „relevant“, sondern schlichtweg falsch. Neben diesen Extremfällen, bei denen zu einfache, veraltete Suchtechniken genutzt werden, finden sich aber auch Fälle, die selbst mit besseren Verfahren kaum zu vermeiden sind. Unser Standardbeispiel ist hier der „Assistent des Geschäftsführers“, der bei einer Suche nach „Geschäftsführer“ oder „CEO“ (falls die entsprechende Suche schon mit Synonymen umgehen kann) zurückgeliefert wird.

Durch die Berücksichtigung von Begriffskontexten, durch linguistische Analysen und durch die Auflösung von mehrdeutigen Begriffen (wie *Bank*, *Jaguar* oder *Kröten*) können semantische Suchen genauere Ergebnisse erzielen, in dem gravierende Interpretationsfehler der Eingaben vermieden werden.

Widersprüchliche Ziele?

Spätestens bei der Präsentation dieser Argumente in einem Unternehmen fragt sich der Entscheider: Was gilt denn nun? Mehr Ergebnisse? Bessere Kundenzufriedenheit durch weniger Anfrageaufwand und ein integriertes Suchergebnis, das mit höherem Aufwand bei der Trefferinspektion erkauft werden muss? Oder weniger Ergebnisse durch verbesserte Erkennungsverfahren und Auflösung von Mehrdeutigkeiten, was aber das Risiko beinhaltet, dass Benutzer unzufrieden sind, da sie sich bereits an die unzulänglichen, syntaktischen Vergleichsverfahren gewöhnt haben und unpassende Treffer vermissen?

Allein durch die obigen qualitativen Argumente kann der Nutzen semantischer Technologien offensichtlich nicht transportiert werden.

Konsequenzen für die Bewertung semantischer Suche

Um einerseits Synonyme und verwandte Begriffe erkennen und andererseits Fehler bei der Erkennung von Bezeichnungen vermeiden zu können, ist – wie in Kapitel 8 argumentiert – die Verwendung von Hintergrundwissen über den jeweiligen Anwendungsbereich bzw. die Anwendung notwendig. Ob dieses Hintergrundwis-

sen nun manuell in Form eines Thesaurus oder einer Ontologie modelliert wurde oder spezielle Zugriffs- und Auswahlalgorithmen für die Nutzung von „linked open data“ (siehe z.B. Kapitel ?) genutzt und implementiert werden, für die Bewertung einer semantischen Suche mit den Standardmaßen Recall und Precision müsste zusätzlicher Aufwand getrieben werden. Jedes Textkorpus müsste *zusätzlich* um das für den Anwendungsbereich benötigte Hintergrundwissen erweitert werden. Der bereits hohe Aufwand zur Entwicklung des Goldstandards würde nochmals steigen.

Vergleich und Bewertung durch Gegenüberstellung

Kommen wir nochmals auf das Eingangsbeispiel zurück: Was können wir unternehmen, wenn wir zwei Personen mit unterschiedlichen Notendurchschnitten vergleichen müssen? Wir können die einzelnen Fachnoten durchgehen und vergleichen, wer in welchem Fach besser abschnitt. Dann stellt sich etwa heraus, dass beide in Geschichte, Erdkunde und Musik gleich gut waren, der 1,3er Kandidat jedoch in Deutsch, Englisch, Latein, Biologie und Kunst viel besser war, während der 2,0-Kandidat in Mathematik, Physik, Biologie und Chemie die Nase vorn hatte. Der 1,3er Kandidat hat in den Wahlfächern Philosophie, Chinesisch und Religion Spitzennoten erhalten, während der 2,0er sich in den Wahlfächern Informatik, Astronomie und Sport seinem Durchschnitt entsprechend gut war.

Wollten Sie mit diesen Kandidaten die Stelle eines naturwissenschaftlich orientierten Mitarbeiters besetzen, wird Ihre Wahl durch die Berücksichtigung der Einzelnoten anders ausfallen.

Vergleichsmodell

Wir legen dem Vergleich von Suchmaschinen anhand ihrer Suchergebnisse ein ähnliches Modell zugrunde. Wir setzen voraus, dass zwei Suchmaschinen $\mathbf{S_1}$ und $\mathbf{S_2}$ miteinander verglichen werden sollen, dass beide Suchmaschinen die gleichen Dokumente indexieren und dass die gleichen Anfragen an beide Suchmaschinen gestellt werden. Rahmenbedingungen die im Unternehmenskontext ohne weiteres erfüllbar sind.

Annahmen: Wir gehen davon aus,

- dass $\mathbf{S_2}$ qualitativ besser passende Ergebnisse liefert als $\mathbf{S_1}$ und
- das unabhängig von einzelnen, konkreten Suchanfragen, die durch die Ersetzung von $\mathbf{S_1}$ durch $\mathbf{S_2}$ zusammenfassend beschrieben und dargestellt werden soll.

Wir gehen davon aus, das $\mathbf{S_1}$ die „Referenzsuchmaschine“ ist, die entweder bereits im Unternehmen eingesetzt wird oder die im Voraus festgelegt wurde, und dass $\mathbf{S_2}$ gegen diese Referenzsuchmaschine verglichen werden soll.

Bezeichnen wir mit $\mathbf{E_1}$ und $\mathbf{E_2}$ die geordneten Ergebnismengen der Suchmaschinen $\mathbf{S_1}$ und $\mathbf{S_2}$ hinsichtlich einer Anfrage. Dann kann sich sowohl die Gesamtzahl der Treffer beider Suchmaschinen **n** bzw. **m** als auch die Anordnung der Treffer unterscheiden. Von Interesse sind hierbei insbesondere die folgenden Treffermengen:

- $\mathbf{E_1 \cap E_2 = E}$
- $\mathbf{E_1 \backslash E_2 = A}$
- $\mathbf{E_2 \backslash E_1 = B}$

Die Menge **E** umfasst die Treffer, die von beiden Suchmaschinen zurückgeliefert werden. **A** umfasst die Treffer, die lediglich von $\mathbf{S_1}$ gefunden werden, und **B** die Treffer, die nur von $\mathbf{S_2}$ gefunden werden. Diese Werte sagen uns bereits einiges über die Gemeinsamkeiten und Unterschiede der Ergebnisse beider Suchmaschinen und damit über deren Suchfunktionen selber aus.

Aus eigener Erfahrung kennen Sie auch diese einfache Weisheit: „Das Gesuchte findet man immer zum Schluss“. Oder anders ausgedrückt, wenn Sie das Gesuchte finden, war es immer das Letzte. Diese einfache Wahrheit nutzen wir bei unserem Verfahren ebenfalls. Hierzu gehen wir davon aus, dass es für jeden Treffer mindestens einen „virtuellen Benutzer“ gibt, den dieser Treffer interessiert, der diesen Treffer finden möchte und der ihn als richtig hinsichtlich seines Informationsbedarfs beurteilt, den er mit der von ihm gestellten Frage ausdrückt.[6] Um diesen Treffer in einem der Suchergebnisse $\mathbf{E_1}$ oder $\mathbf{E_2}$ zu identifizieren, muss dieser Benutzer im schlimmsten Fall alle Treffer in $\mathbf{E_1}$ bzw. $\mathbf{E_2}$ in der ausgegebenen Reihenfolge inspizieren, bis er am Ende diesen Treffer findet. Befindet sich der Treffer am Anfang des Ergebnisses, muss er weniger Zeit aufwenden. Liegt der Treffer eher am Ende der linear geordneten Treffermenge, muss er mehr Zeit investieren. Offensichtlich korreliert die Position des Treffers mit dem Zeitaufwand, den der Benutzer für die Trefferinspektion aufbringen muss.[7]

Mathematisch kann dieser Zeitaufwand anhand der Trefferpositionen quantifiziert werden. Für jeden Treffer aus $\mathbf{e \in E}$ kann die Position in $\mathbf{E_1}$ und $\mathbf{E_2}$ ermittelt werden. Wenn gilt $\mathbf{pos(e,E_1) > pos(e,E_2)}$, dann hat sich die Position von **e** verbessert. Gilt $\mathbf{pos(e,E_1) < pos(e,E_2)}$, dann hat sich die Position von **e** verschlechtert. Gilt $\mathbf{pos(e,E_1) = pos(e,E_2)}$, ist bezogen auf diesen Treffer zwischen beiden Suchmaschinen kein Unterschied ermittelbar. Hieraus kann direkt ein Maß abgeleitet werden, welches angibt wie viele Positionsverbesserungen/-verschlechterungen $\mathbf{S_2}$ gegenüber $\mathbf{S_1}$ bzgl. einer Anfrage über der Dokumentenmenge produziert. Relati-

[6] Zu mindestens der Autor einer Information kann als dieser Benutzer angesehen werden.

[7] Wobei wir einige Idealisierungen vornehmen: 1) dass der Benutzer alle Treffer bis zum gesuchten ansieht, 2) dass der Benutzer nur sequentiell vorgeht und 3) dass der Benutzer für die Inspektion jeden Treffers durchschnittlich die gleiche Zeit aufwendet.

viert anhand der jeweils schlechteren Position kann so ein prozentualer Wert für die vom Benutzer erzielbaren Zeitgewinne/-verluste ermittelt werden.

Durch Ermittlung der Positionsverbesserungen über alle Anfragen, durch deren Gewichtung anhand des Umfangs von **E**, durch deren Mittelung über alle Anfragen und durch Gewichtung anhand der Häufigkeit der Anfragen kann ein realistisches Bild der durchschnittlichen, relativen Zeitersparnis ermittelt werden. Darüber hinaus kann dieses Modell auch die unterschiedlichen Wahrscheinlichkeiten berücksichtigen, mit denen Treffer auf den ersten drei relevanten SERPs gefunden werden.

Die Mengen **A** und **B** geben darüber hinaus Auskunft, wie viele Treffer vermieden, bzw. wie viele Treffer zusätzlich gefunden werden können, und welche Zeitersparnis sich hieraus für den Benutzer ergibt.

Zu beachten ist, dass dieses Modell *nicht* über die Relevanz der Ergebnisse der Suchmaschinen argumentiert, sondern lediglich über die Verbesserungen, Verschlechterungen von Treffermengen und deren Positionen innerhalb von Suchergebnissen. Damit aber ist es unabhängig von konkreten Bewertungen von Benutzern und kann zum direkten, objektiven, reproduzierbaren, automatischen und wiederholten Vergleich von Suchergebnissen und Suchmaschinen auf großen Anfragemengen eingesetzt werden.

Vergleich einer semantischen Suche mit zwei Volltextsuchen

Experiment für ein Internetportal

Im Jahr 2009 hatten wir die Gelegenheit die semantische Suche Ontonyms mit der Volltextsuche eines – aus gutem Grund ungenannt bleiben wollenden – Internetportals anhand von rd. 6.450 Dokumenten des Portals und rd. 4.000 realen Benutzeranfragen, die 1.300 unterschiedliche Suchanfragen beinhalteten, vergleichen zu können. Eine erste Zusammenfassung der Ergebnisse dieses Vergleichs wurde in [7] publiziert.

Die Volltextsuche des Portalbetreibers basierte – und basiert noch heute (!) – auf Substring-Matching und einer kleineren Menge von manuell gepflegten Synonymen. Die semantische Suche Ontonyms umfasste zum damaligen Zeitpunkt einen Thesaurus mit rd. 8.330 Bezeichnungen, Wordstammbildung und thesaurusbasierter Ausnahmeerkennung, Erkennung zusammengesetzter Bezeichnungen und einfacher Rechtschreibfehlerkorrektur.

Tabelle Y.1 fasst die anhand der Anfragehäufigkeit gewichteten und über alle Suchanfragen gemittelten Ergebnisse des Vergleichs zusammen. Die Ergebnisse

werden separat für die ersten drei SERPs[8], die Seiten der Suchergebnisse, die vom Großteil der Benutzer noch betrachtet werden, und für alle Treffer gezeigt.

	1-3 SERP	Alle SERPs
∅ Zeitersparnis durch weniger Treffer	10,5%	14%
∅ Zeitersparnis durch verändertes Ranking	4,5%	9%
∅ Anzahl Treffer der Volltextsuche	38,3	187,8
∅ Anzahl Treffer der semantischen Suche	34,5	138,5
∅ Anzahl unpassender Treffer der Volltextsuche	23,2	97,3
∅ Anzahl passender Treffer, die die Volltextsuche nicht fand	19,4	48

Tabelle Y.1: Vergleichsergebnisse Volltext- vs. Semantische Suche für ein Internetportal

Generell lässt sich feststellen, dass bei der Verwendung einer semantischen Suche durch ein verändertes Ranking und durch weniger Treffer, die aus der Vermeidung der Substring-Suche und die Berücksichtigung des Termkontextes resultieren, auf den ersten drei SERPs durchschnittlich 15% der Zeit zur Trefferinspektion (bzw. 23% auf allen SERPs) vom Benutzer eingespart werden können.

Obwohl die semantische Suche zusätzliche Treffer durch die Berücksichtigung von Synonymen, verwandten und ähnlichen Bezeichnungen findet, erzielt sie auf den ersten drei SERPs durchschnittlich 10% weniger Treffer (bzw. insgesamt 27% auf allen SERPs). Die rd. 60% unpassenden Treffer auf den ersten drei SERPs (74% auf allen) werden durch die semantische Suche durch rd. 50% passgenauere Treffer ersetzt (25% auf allen SERPs).

Die Ergebnisse zeigen, dass die Nutzer dieses Internetportals von einer besseren Suchfunktion profitieren könnten. Insbesondere die eingesetzte „Substring-Suche“ führt zu suboptimalen Suchergebnissen in diesem Portal[9].

Der Nutzen für die Benutzer besteht eindeutig in

1) präziseren Ergebnissen durch Vermeidung der Substring-Suche und Berücksichtigung des Termkontextes,
2) zusätzlichen Treffern durch Berücksichtigung verwandter und ähnlicher Begriffe, damit
3) Verringerung der vom Benutzer insgesamt zu stellenden Anfragen und
4) Reduktion des Inspektionsaufwands durch Verbesserung der Ergebnisanordnung.

[8] Den SERPs lag bei diesem Experiment eine Paginierung von 25 Treffern pro SERP zugrunde.

[9] Interessanterweise sind die Benutzer mit diesen Ergebnissen – lt. Aussage des Portalbetreibers – zufrieden. Unklar ist, ob die Benutzer bzgl. der schlechten Ergebnisse bereits kapituliert und sich damit arrangiert haben, oder ob andere Marktmechanismen (wie z.B. hohe page impressions zur Steigerung des Marktwerts oder geringe Erfolgsrate, um zahlende Kunden zu binden, etc.) den Betreiber zu dieser Aussage bewegen.

Experiment für das WDB Suchportal für Weiterbildung in Berlin-Brandenburg

Wie bereits oben dargestellt, hat Ontonym gemeinsam mit EUROPUBLIC die Suche im WDB Suchportal Berlin-Brandenburg im Jahr 2013 auf die semantische Suche Ontonyms umgestellt. Im Rahmen dieser Umstellung ergab sich erneut die Möglichkeit, einen Vergleich zwischen der Volltextsuche des WDB Suchportals, die unter anderem eine implizite „Prefix-Suche“ nutzt, und der semantischen Suche Ontonyms durchzuführen.[10]

Zum damaligen Zeitpunkt Bestand der Thesaurus der semantischen Suche aus rd. 11.150 Bezeichnungen. Der Versuch wurde anhand von annähernd 28.000 Beschreibungen von Weiterbildungsangeboten und 205.550 über 2 Monate aufgezeichneten, realen Benutzeranfragen, die öfter als 4 mal gestellt wurden, mit rd. 7.200 unterschiedlichen Anfragen durchgeführt.

Bevor wir auf die quantitativen Ergebnisse dieses Vergleichs eingehen, wollen wir anhand von exemplarischen Anfragen und Ergebnissen einige charakteristische Klassen von Vergleichsergebnissen präsentieren, um darzustellen, warum der Nutzen semantischer Suche nicht allein durch qualitative Aussagen wie „mehr Ergebnisse“, „korrekte Ergebnisse“ etc. zusammengefasst werden kann.

Klassen von Vergleichsergebnissen

Die folgenden Grafiken stellen die Suchergebnisse der Volltextsuche (oben) und der semantischen Suche (unten) von links nach rechts in absteigender Reihenfolge geordnet dar. Treffer der Suchmaschinen werden durch Kreise dargestellt. Linien verbinden Dokumente die von beiden Suchmaschinen gefunden wurden (Kreise). Kreise ohne verbindende Linien wurden nur von einer der Suchmaschinen erzielt.

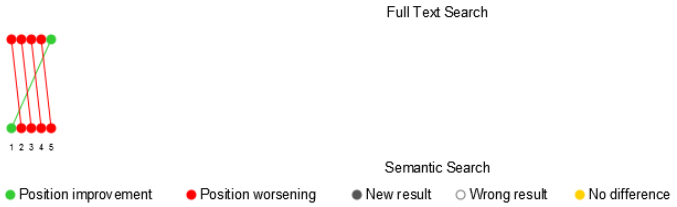

Abb. Y.2 Nahezu identische Ergebnisse

Bei der obigen Anfrage nach „haushandwerker“ erzielten beide Suchmaschinen die gleichen Treffer in leicht veränderter Reihenfolge (Abb. Y.2).

[10] Bei diesem Vergleich war die Anzahl der maximal zurückgelieferten Treffer bei beiden Suchen durch Vorgabe des WDB Suchportals auf 1.000 Treffer eingeschränkt.

Abb. Y.3 Nahezu identische Reihenfolge

Bei der Anfrage nach „pflegebasiskurs“ (Abb. Y.3) kommt es zu einer leicht veränderten Reihenfolge der Ergebnisse. Einerseits werden einige Dokumente als relevanter betrachtet, da die matchende Bezeichnung in ihrem Titel stärker gewichtet wird, andererseits da neue Dokumente, die Schreibweisen wie „Pflege-Basiskurs“ oder Synonyme wie „Basispflegekurs“ enthalten, als passgenauer eingestuft werden.

Abb. Y.4 Zusätzliche Treffer

Wie am Beispiel der Ausgabe für die Anfrage „QM“ ersichtlich wird (Abb. Y.4), können zusätzliche Treffer über die Berücksichtigung von Abkürzungen, Synonymen, Schreibvarianten und fremdsprachlichen Bezeichnungen erkannt werden. Durch die Erkennung unterschiedlicher Bezeichnungen im Text kann der gesuchte Begriff eine höhere Bewertung erhalten, so dass viele Treffer als passgenauer identifiziert werden können.

Während der erste Treffer bei der Anfrage noch gleich bewertet wird, kann die semantische Suche durch die Berücksichtigung verwandter und ähnlicher Bezeichnungen zusätzliche Treffer finden. Einige Fehltreffer können ausgeschlossen

werden, da die semantische Suche kein Prefix-Matching durchführt.

Abb. Y.5 Vermeidung fehlerhafter Erkennung zusammengesetzter Ausdrücke

Neben der Vermeidung von Prefix-Matching-Fehlern und der Erkennung synonymer Bezeichnungen, können Interpretationsfehler bei Anfragen wie „kreatives schreiben“ vermieden werden (Abb. Y.5). Eine konventionelle Volltextsuche betrachtet beide Terme als UND oder UND/ODER verknüpft, so dass auch Treffer gefunden werden, in denen beide Begriff an unterschiedlichen Positionen im Text auftreten. Durch die Nutzung eines Thesaurus kann eine solche Bezeichnung – wie auch der „Assistent der Geschäftsführung“ – als zusammenhängend identifiziert werden, so dass viele Fehlinterpretationen vermieden werden können.

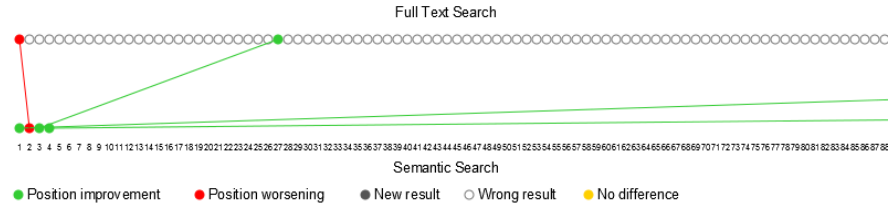

Abb. Y.6 Drastische Reduktion der Treffermenge

Im Fall der Anfrage „Reise Kauffrau“ kann dies soweit gehen (Abb. Y.6), dass die Benutzerin lediglich einen Bruchteil der Treffer inspizieren muss.

Abb. Y.7 Zusätzliche Treffer durch Synonyme und verwandte Begriffe

Es kann jedoch auch ein anderer Fall eintreten, wie z.B. bei der Anfrage nach „radio“ welches mehrere Synonyme, wie „Rundfunk“ oder „Hörfunk“, besitzt und das auch mehrdeutig interpretierbar ist als „Radiogerät“ oder „Rundfunkempfänger“. Für eine solche isolierte Anfrage kann keine Interpretation bevorzugt werden, so dass alle Bedeutungen und deren Bezeichnungen berücksichtigt werden müssen, was zu einer starken Erhöhung der Treffermenge durch eine semantische Suche führen kann. Andererseits kann eine semantische Suche Prefix-Matching-Fehler ausschließen und irrelevante Treffer zu „Radiologie“ oder „Radiologe“ vermeiden (Abb. Y.7).

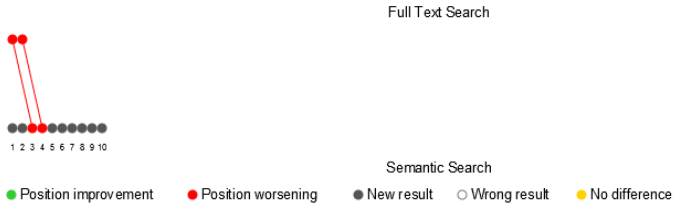

Abb. Y.8 Zusätzliche Treffer durch Erkennung umgangssprachlicher Bezeichnungen

Dies kann so weit gehen, dass bei der Verwendung umgangssprachlicher Bezeichnung, die selten beim Schreiben von Textdokumenten genutzt werden (wie der Anfrage „lokführer“, Abb Y.8), und durch die Berücksichtigung von Benutzer-erfundenen Bezeichnungen, wie der Umschreibung der Bezeichnung „Pflegebasiskurs“ durch die vom Benutzer verwendete Bezeichnung „Basispflegekurs“ (Abb Y.9), eine sehr viel größere Treffermenge identifiziert werden kann.

Abb. Y.9 Zusätzliche Treffer durch Erkennung benutzer-erfundener Bezeichnungen

Quantitative Ergebnisse

Diese Breite der möglichen Vergleichsergebnisse macht es, wie eingangs beschrieben, schwierig zu beurteilen, ob denn nun semantische Suche durch präzisiere Ergebnisse einen Vorteil gegenüber einer Volltextsuche darstellt oder ob die zusätzlichen Treffer nicht doch den benutzerseitigen Aufwand erhöhen und damit eher von Nachteil sind. Aufschluss gibt wiederum die Darstellung der quantitativen Ergebnisse des Vergleichs in Tabelle Y.2.

Auch für dieses Experiment lässt sich generell feststellen, dass durch ein verändertes Ranking und durch weniger Treffer, die bei der semantischen Suche das Resultat der Vermeidung des impliziten Prefix-Matchings, der Berücksichtigung des Termkontextes und der Ersetzung der impliziten ODER- durch eine UND/ODER-Verknüpfung sind, durchschnittlich 20% der Zeit zur Trefferinspektion auf den ersten drei SERPs (bzw. 50% auf allen SERPs) vom Benutzer eingespart werden können.

	1-3 SERP	Alle 100 SERPs
∅ Zeitersparnis durch weniger Treffer	10,4%	13,3%
∅ Zeitersparnis durch verändertes Ranking	9,6%	37,2%
∅ Anzahl Treffer der Volltextsuche	27,7	474,7
∅ Anzahl Treffer der semantischen Suche	24,5	346,1
∅ Anzahl unpassender Treffer der Volltextsuche	15,2	257,5
∅ Anzahl passender Treffer, die die Volltextsuche nicht fand	12,1	129,0

Tabelle Y.2: Vergleichsergebnisse: Volltextsuche und Semantische Suche des WDB Suchportals für Weiterbildung in Berlin-Brandenburg

Obwohl die semantische Suche zusätzliche Treffer durch die Berücksichtigung von Synonymen, verwandten und ähnlichen Bezeichnungen findet, erzielt sie durchschnittlich 12% weniger Treffer auf den ersten drei SERP (bzw. insgesamt 27% auf allen SERPs). Die rd. 55% unpassenden Treffer auf den ersten drei SERP (54% auf allen) werden durch die semantische Suche durch rd. 44% passgenauerer Treffer ersetzt (27% auf allen SERPs).

Auch diese Ergebnisse zeigen, dass die Nutzer des WDB Suchportals von einer besseren Suchfunktion profitieren. Auch hier wird der Nutzen für die Benutzer eindeutig durch 1) präzisere Ergebnisse, 2) zusätzliche Treffer durch Berücksichtigung verwandter und ähnlicher Begriffe, 3) damit einhergehend die Verringerung der vom Benutzer insgesamt zu stellenden Anfragen und 4) die Reduktion des Inspektionsaufwands durch Verbesserung der Ergebnisanordnung erzielt.

Vergleich der Ergebnisse

Obwohl sich die Ergebnisse beider Experimente in den Absolutwerten unterscheiden, liegen die relativen Verbesserungen – bis auf eine Ausnahme – in der gleichen Größenordnung.

Die Zeitersparnis durch besseres Ranking weicht beim zweiten Experiment gegenüber dem ersten Experiment signifikant ab. Wir denken, dass diese Abweichung das Resultat der insgesamt auf 1.000 Treffer begrenzten Menge der zurück-

gelieferten Ergebnisse und der veränderten Interpretation der Verknüpfung nicht-zusammenhängender Anfragebegriffe ist.

Resümee

Wir haben in diesem Artikel ein automatisiertes Vergleichsverfahren für Suchmaschinen beschrieben, mit dem objektive, reproduzierbare Aussagen über die Veränderungen in den Ergebnismengen ermittelt werden können. Durch den Vergleich einer semantischen Suche mit zwei unterschiedlichen, herkömmlichen Volltextsuchen, haben wir, anhand von zwei unterschiedlichen Dokumenten- und Anfragemengen und unterschiedlichen Ausbaustufen eines Thesaurus, der beide Anwendungsbereiche gleichermaßen abdeckt, einerseits gezeigt, dass die durch semantische Suche erzeilbare Verbesserung in beiden Fällen in derselben Größenordnung liegt. Andererseits konnten wir, durch diese sich gegenseitig bestätigenden Werte, unserem Wissen nach erstmals fundierte und reproduzierbare Aussagen über die Wirtschaftlichkeit semantischer Suche ermitteln.

Literatur

1. van Rijsbergen, C. J. (1979) *Information Retrieval*, Butterworth.
2. Salton, G., McGill, M.J. (1983) *Introduction to modern information retrieval*, McGraw-Hill, New York.
3. Xu, J. (1999) „Internet search engines: real world IR issues and challenges“, Presentation at *CIKM 99*, Kansas City.
4. Jansen, M., B. J., Spink, A. and Saracevic, T. (2000) „Real life, real users and real needs: A study and analysis of users queries on the Web“, *Information Processing and Management*, 36(2), pp. 207-227.
5. iProspect (2008) „Blended Search Results Study“, April. Accessed 26/5/2009 via http://www.iprospect.com/premiumPDFs/researchstudy_apr2008_blendedsearchresults.pdf, last access 3/1/2014 via http://www.herramientas-seo.com/pdf/estudio-buscadores-iprospect.pdf.
6. Günther, J.S. (2008) „Erfolgreiches Onlinemarketing mit Google“, Kapitel 2.6 Nutzerverhalten beim Suchen, vwh Verlag Werner Hülsbusch.
7. Tolksdorf, R. Hoppe, T. (2009), „Quantitative Analyse von Ergebnissen Semantischer Suche“, KnowTech 2009, 11. Kongress zum IT-gestützten Wissensmanagement, 6.-7. Oktober 2009, Bad Homburg.