Sonderdruck aus
BI-SPEKTRUM

Bild: Shutterstock

Die relationale Suche

Wieso Search kein Feature ist

Ein Beitrag von
Christian Werling

Die Online-Suche nach Informationen ist für die meisten Menschen genauso selbstverständlich wie das Schreiben einer E-Mail. Kein Wunder also, dass Search auch im BI-Umfeld immer mehr an Bedeutung gewinnt und sich mehrere Analytikanbieter darum bemühen, Suchfunktionen zu ihren Lösungen hinzuzufügen. Dabei ist Search keine Nebensache, die sich mal eben so einfach auf bestehende Lösungen aufdocken lässt. Das wird deutlich, wenn man sich Google und Yahoo anschaut. Google hat Search von Grund auf neu aufgebaut, während es sich bei Yahoo eher um einen Anbau handelte. Auf das Resultat braucht man nicht tiefer einzugehen, denn jeder weiß, dass „googeln" heutzutage ein Synonym für die Online-Suche geworden ist.

Suchfunktionen für numerische Daten zu entwickeln ist eine komplexe Aufgabe. Damit die Suche so funktioniert, wie es die Benutzer erwarten, muss sie in die Architektur des Analysesystems eingebunden sein [For14]. Doch wie sieht eine solche von Grund auf entwickelte Suche aus?

Suche ist nicht gleich Suche

Zunächst ein kurzer Überblick über die verschiedenen, vergleichsweise einfachen Search-Ansätze, die beliebte Dienste wie Google, Amazon, LinkedIn und Facebook verwenden.

Google verwendet die sogenannte „Dokumentensuche". Schließlich sind Webseiten im Prinzip nichts anderes als aufgeblähte Textdokumente. Der PageRank-Algorithmus von Google verleiht verlinkten Dokumenten mehr Gewicht und sorgt dafür, dass die relevantesten Seiten am höchsten eingestuft und zuerst angezeigt werden.

Amazons durchsuchbares Universum besteht aus einer bekannten Menge an Objekten (Produkten) und deren Eigenschaften. Die Suchmaschine nutzt die sogenannte „Facetten-" oder „Objektsuche", um relevante Produkte zu finden. LinkedIn funktioniert ähnlich, nur dass die Objekte in diesem Fall Personen, Unternehmen und Jobs sind.

Facebook verwendet noch eine weitere Technik namens „Graph Search", die die Verbindungen zwischen Freunden analysiert. Mit dem Graph Search können Nutzer festlegen, wie weit eine Suche in ihrem Freundesnetzwerk gehen soll.

Die relationale Suche

Die relationale Suche ist ein komplett neuer BI-Ansatz, der für die Suche und Analyse von Unternehmensdaten aus den Bereichen Finanzen, Marketing, Vertrieb, Supply Chain und andere Datenquellen innerhalb eines Unternehmens entwickelt wurde.

Zwar gab es schon in der Vergangenheit Search-Ansätze im BI- und Analytics-Umfeld, wie zum Beispiel BI Search, bei der Nutzer nach Reports oder Abschnitten von Reports suchen konnten, oder Text Analytics, wo es um die Suche unstrukturierter Daten geht (siehe auch [Rus07]). Auch die meisten aktuellen Search-Ansätze setzen, wie eingangs besprochen, auf der bestehenden BI-Architektur auf, können also als eine Art Weiterentwicklung des von Philip Russom beschriebenen BI Search gesehen werden [Rus07]. Hierbei werden Enterprise-Search-Technologien genutzt, um die vorgedachten Reports/Dashboard zu durchsuchen.

Bei der relationalen Suche hingegen geht es darum, die relationalen Daten eines Unternehmens zu durchsuchen. Im Gegensatz zu den besprochenen Search-Ansätzen muss sich die relationale Suche einem völlig anderen und viel schwierigeren Suchproblem stellen: Unternehmensdaten sind nicht vergleichbar mit den Webdokumenten, die Menschen mit Google suchen oder dem Facebook-Freundesnetzwerk. Unternehmensdaten werden in der Regel in relationalen Datenbanken, privaten Clouds, Public Clouds, Hadoop-Clustern und oft sogar noch in Tabellenkalkulationen oder Sheets gespeichert.

Unternehmensdaten sind kompliziert

Die Daten eines Unternehmens erstrecken sich über mehrere Datenbanken, Tabellen, Spalten, Zeilen und Schlüssel, mit einem komplexen Geflecht von Beziehungen untereinander. Bei globalen Unternehmen werden diese Daten in mehreren Rechenzentren gehalten, die über die ganze Welt verteilt sind. Ihre Quellen sind verschiedene Systeme, die ursprünglich für eine begrenzte Gruppe von Benutzern bestimmt waren. Für eine relationale Suche müssen alle diese Datenquellen erfasst, modelliert und in den In-Memory-Cache geladen werden, wobei die Beziehungen korrekt identifiziert werden müssen und nur den jeweils autorisierten Personen die Suche und Analyse dieser Daten ermöglicht werden darf.

Und genau wie die Nutzer es von der Internetsuche gewohnt sind, muss all dies sofort ohne jegliche Wartezeit geschehen (siehe auch [Pra19]).

100 Prozent genaue Ergebnisse

Manchmal liegt auch Google falsch. Wer nach „Städten, die nicht in Bayern liegen“ sucht, bekommt als Erstes Informationen über Städte in Bayern angezeigt. Allerdings sind die Auswirkungen einer schlechten Google-Suche trivial. Als Konsequenz muss man einfach die Frage umformulieren und es erneut versuchen.

CHRISTIAN WERLING, Informatiker, ist Regional Director D-A-CH bei ThoughtSpot. Er hat über 15 Jahre Erfahrung im Bereich Unternehmenssoftware und war unter anderem bei Unternehmen wie QLIK, Questback, BlueYonder und Treasury Intelligence Solutions tätig. Big Data und Analytics sind dabei stets im Fokus gewesen. Christian Werling startete seine IT-Karriere 2002 bei dem globalen IT-Lösungsanbieter Hewlett Packard. Er lebt am Niederrhein und ist in seiner Freizeit begeisterter Ausdauersportler.

E-Mail:
hello@thoughtspot.com

Bei einer Suche in Unternehmensdaten gibt es keinen Spielraum für ungenaue Antworten. Auf die Frage, wie viel Umsatz im letzten Quartal erwirtschaftet wurde, gibt es nur eine einzige richtige Antwort. Fehler kann man sich hier nicht erlauben, denn die führen schnell zu schlechten Geschäftsentscheidungen und Akzeptanzproblemen bei den Nutzern.

Nicht alles ist für alle Augen bestimmt

Bei der Online-Suche ist Sicherheit zwar wichtig, aber nicht entscheidend. In Unternehmen ist das Gegenteil der Fall: Die Sicherheit muss zum Beispiel die Funktion, die Abteilung oder auch den geografischen Standort des Benutzers berücksichtigen. Es ist einfach, eine Spalte wie Gehälter auszublenden, aber was ist, wenn Vertriebsleitern nur Informationen über ihre Regionen angezeigt werden sollen?

Hinzu kommt, dass die meisten Unternehmen in mehreren Regionen tätig sind, jede mit ihren eigenen Datenschutz- und Compliance-Anforderungen. Sicherheit auf Zeilenbasis ist daher ein Muss für Search-BI (siehe auch [Pra19]). Das alles lässt sich nicht einfach mal so eben anbauen.

Faktor Schnelligkeit

Die Reaktionszeit von Suchmaschinen wird oft in Millisekunden gemessen. Für Search-BI bedeutet dies, dass die relationale Suchmaschine im Handumdrehen Folgendes tun muss:

- Lesen des Inputs (Tastenanschläge, Mausklicks oder Spracheingabe)
- Durchsuchen einer riesigen Datenmenge und Interpretation der Beziehungen zwischen den Datenelementen
- Anwendung der Sicherheitsregeln
- Unterbreitung relevanter Vorschläge
- Übersetzung der Suchanfragen in Queries, die auf die Daten der Benutzer anwendbar sind
- Rückgabe genauer Ergebnisse
- Ansprechende Visualisierung der Ergebnisse

Die Mitarbeiter mögen vielleicht bereit sein, zwei Wochen auf einen Bericht zu warten, verlangen

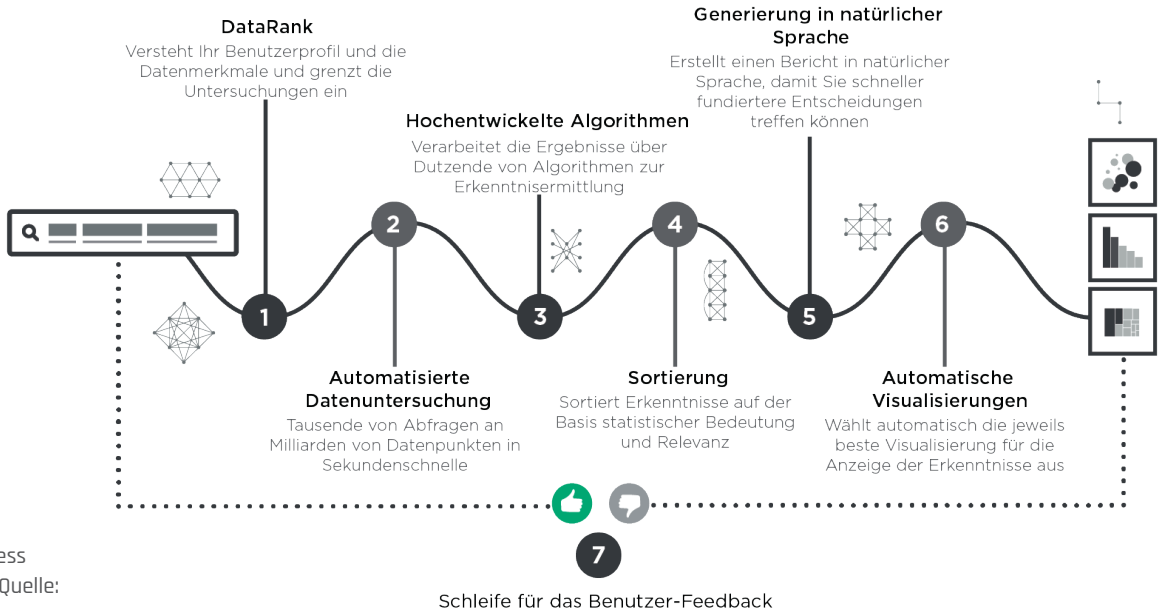

Abb. 1: KI-Prozess für Search-BI (Quelle: ThoughtSpot)

aber, dass jede Suche sofort das entsprechende Resultat anzeigt.

Es gibt viel zu suchen

Die Suche nach Unternehmensdaten ist nicht wie die Suche im Dateisystem auf dem Computer. Es geht hier um Daten im Terabyte-Bereich, die aus vielen verschiedenen Datenquellen stammen.

Bei der Websuche ist das Datenvolumen zwar riesig, aber eine einzige Suche berührt nur einen winzigen Bruchteil dieser Daten. Bei der relationalen Suche muss jedes Datenelement für jeden Suchbegriff berücksichtigt werden, und oft müssen große Datenmengen aggregiert werden, wie zum Beispiel bei der Suche „Umsatz der letzten fünf Jahre“ (siehe auch [Pra19]).

Natürliche Sprachverarbeitung (NLP)

Dank der explosionsartigen Zunahme der sprachgesteuerten Suche über Google, Alexa, Siri und andere steigt die Nachfrage nach sprachgesteuerten analytischen Queries. Das bedeutet, dass die relationale Suche KI-Fähigkeiten benötigt, um die Intention der vom Nutzer gestellten Suchanfragen sowie ein wahrscheinliches Ergebnis zu ermitteln. Noch wichtiger aber ist, dass sie in der Lage sein muss, die analytische Absicht zu erfassen, um die einzig richtige Antwort auf die gestellte Frage liefern zu können. Wenn zum Beispiel ein Nutzer nach „Wie viele McDonalds gibt es in Berlin“ sucht, muss eine NLP-Engine raten, ob mit „McDonalds“ die Restaurantkette, ein Straßenname oder ein Familienname gemeint ist (siehe auch [Hal17]).

Architektur für die relationale Suche

Eine Sucherfahrung zu schaffen, die in der Lage ist, Datenvolumen und Datenkomplexität zu bewältigen, die den Sicherheits- und Compliance-Vorschriften eines Unternehmens genügt, für jeden Anwender einfach zu bedienen und darüber hinaus noch schnell ist, ist kein leichtes Unterfangen.

Die Analyse großer Datenmengen mit vielen Terabyte und Milliarden Datensätzen in Kombination mit komplexen Schemata in Sekundenschnelle lässt sich nicht mit herkömmlichen diskbasierten Lösungen oder mit Speicher-/Disk-Hybridlösungen bewältigen. Dafür bedarf es einer Engine, die von Grund auf dafür entwickelt wurde mit verteilter, massiv paralleler In-Memory-Arbeitsweise [Tho19].

Eine Herausforderung in vielen Unternehmen ist das Missverhältnis zwischen der Anzahl von Reporting-Anfragen aus dem Fachbereich und der Anzahl von Datenanalysten, die diese Berichte erstellen. Lange Wartezeiten und ein hoher Backlog von Anfragen sind die Folge.

Eine Lösung für dieses Problem kann die relationale Suche sein, die den Anwender mit echtem Self-Service für die Datenanalyse und Ad-hoc-Anfragen unterstützt. Wichtig ist hierbei, dass die Search-Plattform dem Anwender eine personalisierte und sichere Sucherfahrung bietet, das heißt sobald etwas in die Suchleiste eingegeben wird, schlägt die Search-Engine automatisch auf Basis der Berechtigungen des Anwenders passende und häufig benutzte Suchbegriffe vor. Dazu können Machine-Learning-Algorithmen genutzt werden, die basierend auf Datencharakteristiken, Nutzerverhalten und rollenbasierten Zugriffsrechten die Suchvorschläge permanent optimieren [Tho19].

Sobald ein Nutzer eine Frage eingibt, muss die Lösung diese Suche korrekt in eine Query übersetzen, um das korrekte Ergebnis berechnen zu können. Wie dieser Prozess funktionieren könnte, zeigt Abbildung 1 (siehe auch [Hal17]).

Eine Search-Lösung sollte nicht nur auf einfache Modelle oder vordefinierte Inhalte beschränkt sein, sondern auch komplexe Schemata verstehen, mehrere Faktentabellen, Formeln, Unterabfragen und alternative Join-Pfade bewältigen können und

so eine stabile und präzise Sucherfahrung liefern. Idealerweise werden auch die häufigsten Queries zwischengespeichert, sodass Berechnungen, für die ältere BI-Tools Stunden benötigen würden, innerhalb von Millisekunden ausgegeben werden können (siehe auch [ErE18]).

Fazit

Es wird deutlich: Hinter der täuschend einfachen Such-Erfahrung verbirgt sich eine Menge Komplexität. Der Aufwand lohnt sich vor allem für jene Unternehmen, die es ihren „normalen" Mitarbeitern ermöglichen wollen, schnelle Antworten auf selbst gestellte Datenfragen zu erhalten, und die wissen, dass diese Fragen jedes Mal anders ausfallen können. Nutzt man für dieses Szenario klassische BI-Tools, erfordert jede Variation die Erstellung einer komplett neuen Anfrage – und das ist arbeits- und zeitaufwendig. Für eine relationale Search-Lösung, deren Architektur darauf ausgelegt ist, schnelle Antworten zu den Unternehmensdaten zu liefern, ist es jedoch das ideale Einsatzszenario.

Für Szenarien, bei denen identische Datenabfragen wiederholt und im gleichen Format präsentiert werden müssen, wie beispielsweise Compliance-Berichte, die von den Aufsichtsbehörden zur Einhaltung der Vorschriften verlangt werden, oder quartalsweise veröffentlichte Finanzberichte, eignen sich traditionelle BI-Tools hingegen besser.

Unternehmen, die Search im BI-Umfeld einführen, sollten sich aber darüber im Klaren sein, dass es hier im Grunde genommen um mehr geht, als jedem im Unternehmen Zugang zu schnellen Datenantworten zu ermöglichen. Mit Search bahnt sich ein großer kultureller Wandel an, der ein hohes Maß an Transparenz und Vertrauen in die Mitarbeiter erfordert, dass diese auf der Grundlage von Datenerkenntnissen angemessen handeln. Dieser Wandel muss von oben begleitet werden. Hier sind die Führungsteams gefragt.

Literatur

[ErE18] Ereth, J. / Eckerson, W.: AI: The New BI – How Algorithms Are Transforming Business Intelligence and Analytics. 2018

[For14] Forrester Research Whitepaper: Boost The Value Of Your Enterprise BI With Advanced Search. 2014

[Hal17] Halper, F.: Advanced Analytics: Moving Toward AI, Machine Learning, and Natural Language Processing. TDWI Best Practices Report, 2017

[Pra19] Prakash, A.: Why Relational Search Is Hard(er) ... And Why We Like It. ThoughtSpot, https://www.thoughtspot.com/codex/why-relational-search-harderand-why-we-it, abgerufen am 19.9.2019

[Rus07] Russom, P.: BI Search and Text Analytics. 2007

[Tho19] White Paper Relational Search: A New Paradigm for Data Analytics. 2019, https://go.thoughtspot.com/white-paper-thoughtspot-relational-search-2nd-edition.htm, abgerufen am 19.9.2019