

Mit Big-Data-Suche zu ungeahnten Ergebnissen

Expertennetzwerke zusammenstellen und in Echtzeit Millionen von Dokumenten durchsuchen – dafür nutzen Pharmakonzerne heute Software für Big-Data-Suche und -Analyse.

Frank Zscheile, München

In forschungsintensiven Pharmaunternehmen erstellen interne wie externe Spezialisten tagtäglich eine immense Anzahl hochtechnischer Dokumente: Forschungsunterlagen, Einträge in medizinische Datenbanken, Versuchsberichte, Patenteinreichungen u.v.m. Hinzu kommt die gesammelte Kommunikation zwischen Forschern aus verschiedenen Fachgebieten. Auf der anderen Seite stehen die Experten, die in den verschiedenen Geschäftsfeldern eines biopharmazeutischen Unternehmens tätig sind: Auch hier kann in einem weltweit tätigen Unternehmen wie

AstraZeneca die Anzahl schnell bei 10.000 und mehr Personen alleine in Forschung und Entwicklung liegen.

Im Rahmen der Forschungs- und Entwicklungsarbeit kommt es täglich vor, dass jemand einen Experten bzw. Informationen im Konzern zu einem dezidierten Thema sucht. Wer kennt sich z.B. mit „Arteriosklerose“ besonders gut aus, wer weiß über Wirkstoffe und aktive Moleküle in den Medikamenten Bescheid, welche Dokumente zu Nebenwirkungen und Patenten gibt es?

Inhaltsanalyse mit computerlinguistischen Methoden

Der Pharmakonzern experimentierte zunächst mit einfachen Suchmaschinen, um Experten anhand der Spuren zu lokalisieren, die sie in Dokumenten hinterlassen. Erst die Einführung einer spezialisierten Big Data Such & Analyse-Software (des Herstellers Sinequa) half jedoch letztlich, aus der schier Vielfalt an Daten verschiedenster Quellen die richtigen Informationen herauszufiltern. Die Inhaltsanalyse basiert dabei auf einer patentierten semantischen sowie linguistischen Analyse und ist einsetzbar für beliebige unstrukturierte und strukturierte Daten.



So erhält der Anwender auch Fundstellen, in denen der eigentliche Suchbegriff gar nicht vorkommt, hingegen Synonyme oder inhaltlich ähnliche Begriffe.

In einem initialen Prozess analysierte AstraZeneca mit der Software zunächst rund 200 Mio. interne und externe Dokumente aus dem Bereich F&E. Dabei wurden auch fachspezifische Relationen zwischen Begriffen (synonyme und semantisch verwandte Begriffe) ermittelt.

Der Index ergab ein Datenreservoir, aus dem relevante Informationen in weniger als zwei Sek. zusammengestellt werden können. Bei Eingabe eines Begriffes in die Suchmaske erstellt das System aus all dem, was zu diesem Thema geschrieben wurde, die besten Experten bzw. Forschungsgruppen zusammen. Der hohe Wettbewerbsdruck in der Pharmabranche zwingt die Hersteller, Innovationen möglichst schnell zur Marktreife zu

bringen. Wer hier in kürzester Zeit also interdisziplinäre Expertenteams konstituieren kann, die standortübergreifend an einem gemeinsamen Thema arbeiten, hat dabei definitiv einen Wettbewerbsvorteil.

Ähnlich nutzt der internationale Pharmakonzern UCB die Sinequa-Plattform. Dieser filtert damit relevante Informationen aus strukturierten und unstrukturierten Daten von jahrelangen klinischen Versuchen. Das biopharmazeutische

Unternehmen mit Hauptsitz in Brüssel und über 8.600 Beschäftigten weltweit erforscht und entwickelt innovative Behandlungsmöglichkeiten in den Bereichen Zentrales Nervensystem und Immunkrankheiten.

SAS-Datensätze gehen in die Millionen

Im Rahmen der Entwicklung medikamentöser Therapien führt UCB umfangreiche klinische Studien durch. Auf die dabei entstehenden Big-Data-Inhalte müssen die Forscher bei der Medikamentenentwicklung permanent zugreifen, um relevante Daten für weiterführende statistische Analysen zu finden. Rund zehn Mio. Files haben sich dabei in den letzten 15 Jahren angesammelt – ein Datenberg von 20 Terabyte, der jährlich um derzeit 30 % wächst. Darunter ist mehr als eine Mio. SAS-Datensätze. Mit wachsender Anzahl an Dateitypen und Inhaltsvolumina wird das Auffinden benötigter Informationen zusehends schwieriger. Die Enterprise Search-Lösung hilft UCB, hier Zeit und Kosten einzusparen, weil Analysten und Programmierer relevante klinische Datensätze schneller finden.