Paradigmenwechsel bei Recherchesystemen durch Semantik und künstliche Intelligenz?

Yewno: ein semantischer Discovery Service im Pilotversuch an der Bayerischen Staatsbibliothek

Von Berthold Gillitzer

Zuletzt waren es die sogenannten Discovery Systeme, wie z. B. Primo Central von Exlibris, das auch im OPAC der Bayerischen Staatsbibliothek zum Einsatz kommt, die schon mit ihrer Bezeichnung als Instrumente zum „Entdecken" einen Paradigmenwechsel im Zugang zu Informationen beanspruchten. Dokumente, die bislang oftmals nur über unterschiedliche Systeme und Suchwege für den Nutzer zugänglich waren, werden hier integriert auffindbar: Aufsätze, Bücher, Zeitungen, Zeitschriften, egal ob als digitale Dokumente oder im Print.

In einem recht fundamentalen Sinn ist aber auch noch bei den Discovery Systemen etwas ganz beim Alten geblieben: Immer basiert die Suche auf dem Vergleich von Zeichenfolgen. So gesehen gab es einen echten Paradigmenwechsel, als nicht mehr primär ein Bibliothekar als Mittler zwischen dem Nutzer und seinem Literaturwunsch stand. Mit dem Bibliothekar konnte der Nutzer einen echten Dialog in menschlicher Sprache führen. Es wird noch eine Weile dauern, bis Computer zu einem derartig komplexen Vorgang in der Lage sein werden.

Von der Schriftkultur zur digitalen Kultur

Damit ist aber auch schon ein anderer Paradigmenwechsel angesprochen, der bislang in Bibliotheken erst in Ansätzen mitvollzogen wurde und der beim Blick auf Bibliothekskataloge und Recherchetechnologien eine nicht unerhebliche Rolle spielt. Lobin[1] reklamiert einen solchen Wandel für den Übergang von der Schriftkultur zur digitalen Kultur.

Die Ablösung der Schriftkultur bedeutet aber nicht einfach den Ersatz von etwas Altem – der schriftlichen Information – durch etwas vollkommen Anderes. Schreiben und Lesen spielen immer noch eine kaum zu überschätzende Rolle in unserer Gesellschaft, gerade auch das Schreiben und Lesen am Computer und an mobilen Endgeräten. Lobin macht eindringlich darauf aufmerksam, dass erst mit dem Hypertext, der Möglichkeit der Verlinkung von Textteilen und noch mehr von Texten untereinander, ein neues Lesen entsteht, in dem der Leser nicht mehr linear einem Text folgt, sondern selbst entscheidet, welchem Link er folgt und welche Informationen er damit rezipiert.[2] Das „Aufbrechen dieser Linearität"[3], wie Dr. Klaus Ceynowa diese Veränderung bezeichnet, hat dabei nicht nur die Dimension, dass Textteile und Texte verknüpft werden und ein Springen zwischen diesen Elementen möglich wird. Vielmehr werden auf diese Weise Texte auch mit nichttextuellen Elementen wie Bildern, (interaktiven) Graphiken, Forschungsdaten, Tabellen usw. verknüpft, die ihrerseits Information und Wissen vermitteln, aber eben auf eine andere Weise als linear zu lesende Texte.[4]

Fragmentierung als Merkmal der digitalen Kultur und ihre Konsequenzen

Wichtig scheint mir an der Stelle Folgendes: Wenn wir in einem solchen vernetzten Wissensraum Texte als eine Art von Knotenpunkten zwischen einer Vielzahl von diversen Informationseinheiten betrachten, so sprechen wir nicht von den Einheiten, die wir z. B. derzeit in unseren Bibliothekskatalogen als Werke verzeichnet haben. Ein wesentliches Element unserer digitalen Kultur, das mit dem zuvor Dargestellten einhergeht, scheint mir eine Fragmentierung zu sein, bei der Information und Wissen in wesentlich kleineren Einheiten wahrgenommen und verarbeitet wird, als das früher der Fall war.[5] Viele traditionell größere Texteinheiten werden entweder gleich in kleineren Teilen verfügbar oder auch situativ „aufgebrochen", wie Ceynowa dies ausdrückt,[6] und dann in ihren einzelnen Fragmenten rezipiert und in neue Zusammenhänge eingebettet.

Gegenüber der durch Fragmentierung und zugleich durch stärkere Verknüpfung geprägten digitalen Lebenswelt bleibt auch der inzwischen hybride Korpus aus digita-

len Medien und Printdokumenten einer Bibliothek zu einem gewissen Ausmaß fremd. Die in den Bibliotheken vorhandenen Informationen werden nicht in der gleichen Weise verfügbar, wie es viele Informationen im Web sind, weil nur die großen Texte, die Bücher und Zeitschriftenartikel als Ganze zugänglich sind, nicht aber direkt die in ihnen enthaltenen viel kleinteiligeren Informationen, auf die es in der vernetzten digitalen Welt ankommt. In diesem Sinn hat m. E. bislang noch kein Paradigmenwandel vom Katalogsystem zum Discovery Service stattgefunden.

Semantik als Lösungsansatz – Yewno, ein semantischer Discovery Service

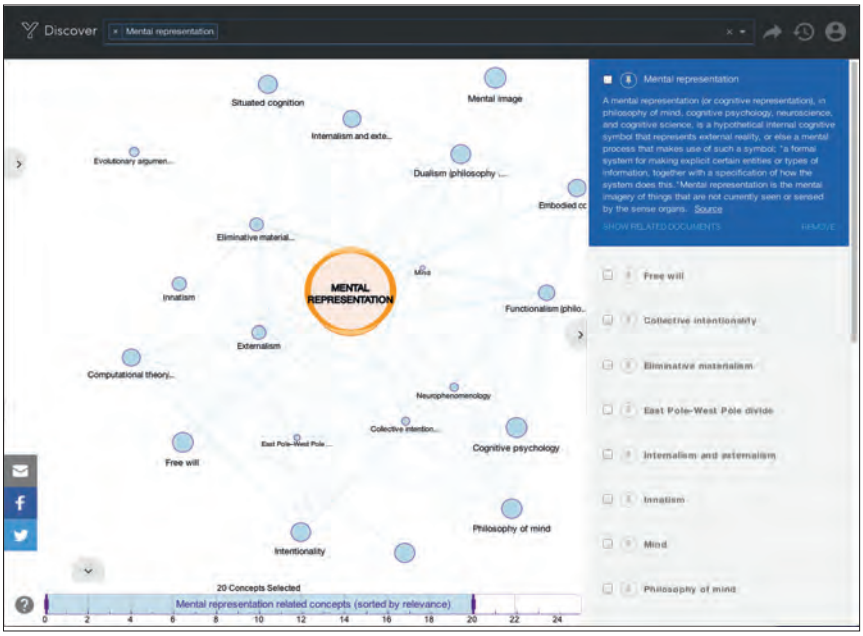
Erst wenn Discovery Services semantisch werden, also wissen, an welcher Stelle es in einem Werk um ein bestimmtes Thema geht, können diese Defizite ausgeglichen werden. Dieser fundamental neue Schritt wird nun mit dem „semantischen Discovery Service Yewno" des gleichnamigen kalifornischen Startup-Unternehmens versucht.[7]

Yewno hat zwei Grundpfeiler, die den angesprochenen Paradigmenwechsel markieren. Der erste ist die Extraktion von Konzepten aus elektronischen Volltexten. Ganz grundsätzlich gibt es dazu auch an anderer Stelle schon Ansätze, als Datamining bezeichnet[8], die meist auf Ontologien, Thesauri oder Wörterbüchern beruhen, einer statischen Datenbasis also, deren Begriffe dann mit computerlinguistischen Verfahren bestimmten Werken zugeordnet werden.[9] Im Gegensatz dazu basiert Yewno auf linguistischer Analyse der Texte mittels künstlicher Intelligenz und maschinellem Lernen.[10]

Yewno arbeitet mit sogenannten Konzepten, die definiert sind als Mengen von Wörtern mit gleicher Bedeutung, gewissermaßen also reine Bedeutungsentitäten im Unterschied zu ihren verbalen Expressionen. Diese Konzepte haben so betrachtet Ähnlichkeit mit klassischen Schlagworten, denen ja auch unterschiedliche Schreibweisen oder Synonyme zugeordnet sein können. Anders als Schlagworte sind sie aber gerade nicht in einer Normdatei oder einer Datenbasis vorgegeben. Konzepte werden vielmehr durch Verfahren der statistischen Semantik extrahiert und bestimmten Textstellen in digitalen Dokumenten, in denen es um diese Konzepte geht, zugeordnet. Im Gegensatz zu den bisher bekannten Verfahren kann die Analysemethode von Yewno nicht mit der Datenbasis veralten. Wird das Verfahren auf Texte über neue Theorien mit entsprechend neuen Konzepten angewandt, werden diese automatisch Teil des Konzeptnetzwerks. Auf dieser technischen Basis gewinnt Yewno nicht nur die Konzepte in den verarbeiteten elektronischen Dokumenten, sondern auch Kenntnisse über die semantischen Relationen zwischen den Konzepten. In der Verknüpfung der Konzepte zeigt sich aber nochmals ein wichtiger Unterschied zu anderen Verfahren, da sich diese Konzeptbeziehungen nicht hierarchisch darstellen, sondern gewissermaßen assoziativ, multidimensional vernetzt, wobei auch die Stärke der Beziehung zwischen zwei Konzepten analysiert werden kann.

Von der Suchmaschine zum inferentiellen Service

Der zweite Grundpfeiler ist die spezielle Weise, wie Yewno die Konzepte und ihre Relationen visualisiert. Die Konzepte werden zwar aus den elektronischen Texten gewonnen, aber die Recherche und ihre Darstellung in einem semantischen Netz erfolgen zunächst unabhängig davon. Als Ergebnis der Recherche werden dem Nutzer die Konzepte mit ihren semantischen Eigenschaften direkt visualisiert: Ein gefundenes Konzept wird im Netz mit verknüpften Konzepten als großer farblich orange markierter Punkt im Zentrum dargestellt, mit einer Erläuterung des Inhalts. Alle Konzepte sind mit ihren verknüpften Konzepten, die wiederum als Punkte dargestellt werden, über Linien verbunden. Die Größe der Punkte gibt dabei Aufschluss darüber, wie stark das Konzept mit dem Ausgangskonzept verbunden ist.



Ein gesuchtes Konzept im Netz der damit verknüpften Konzepte mit Erklärung zur Bedeutung des Konzepts

Die bekannte Art der thematischen Suche wird damit grundsätzlich verändert: Es wird nicht eine lange Ergebnisliste primär präsentiert, sondern das gesuchte Thema vernetzt in seinen vielfältigen sachlichen Bezügen dargestellt. Die multidimensionale Vernetzung und die Möglichkeit, gerade auch gezielt die Verknüpfung zu „weiter entfernten" Begriffen aufzusuchen, ermöglichen ein deutlich intuitiveres Navigieren und können zum Entdecken von Zusammenhängen führen, die dem Nutzer so nicht bekannt waren. Eigentlich kann erst ein System mit dieser Architektur Anspruch auf die Bezeichnung Discovery Service erheben. Nach Ruggero Gramatica sollte deshalb die Anwendung nicht als Suchmaschine betrachtet und bezeichnet werden, sondern als schlussfolgernde Anwendung, als „inference engine"[11], die komplementär zu traditionellen Suchmaschinen und Bibliothekskatalogen konzipiert ist.

Zu guter Letzt enthält Yewno noch einen mehr oder weniger konventionellen Teil, den Weg von den angezeigten Konzepten zu den damit verknüpften Dokumenten. Über den Link „show related documents" werden zunächst Teaser mit den tatsächlich relevanten Textteilen angezeigt, mit ihnen wird dann auch der Zugang zum ganzen Text über einen Link zum Anbieter bereitgestellt. Um sicherzustellen, dass es sich wirklich um ein von der Bibliothek lizenziertes Dokument handelt, findet ein Abgleich mit den Lizenzinformationen der jeweiligen Bibliothek statt, z. B. über deren SFX Knowledgebase.

Pilotbetrieb an der Bayerischen Staatsbibliothek

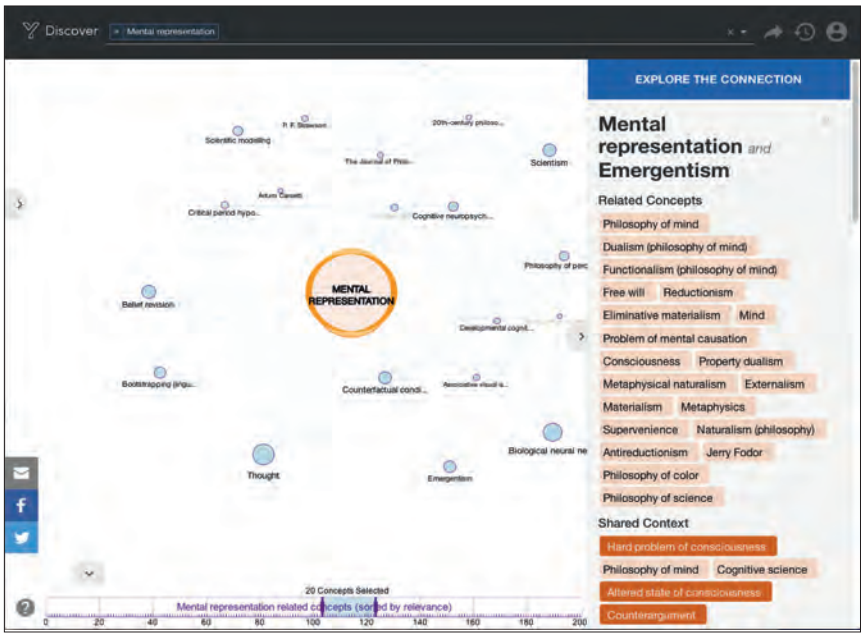
An der Bayerischen Staatsbibliothek wird die Anwendung den Nutzern zunächst in einem dreimonatigen Betatest zur Verfügung gestellt, im Standard, der momentan vorhanden ist. Durchsucht werden also die elektronischen Dokumente, die von Yewno bislang indexiert und verarbeitet wurden, eine große Menge von wissenschaftlichen Zeitschriftenartikeln, die im Openaccess vorliegen, aber auch elektronische Volltexte aus lizenzpflichtigen E-Books und Zeitschriftenartikeln, für die Yewno eine entsprechende Vereinbarung mit den jeweiligen Verlagen abgeschlossen hat. Eigener Bestand an elektronischen Volltexten der Bibliotheken z. B. aus Retrodigitalisierungsprojekten kann von Yewno zusätzlich einbezogen werden. Bislang sind es allein englischsprachige Dokumente, die von Yewno verarbeitet wurden und im System angeboten werden können. Weitere west- und osteuropäische Sprachen sollen aber folgen.

Um die Einhaltung des Datenschutzes nach deutschem Recht zu gewährleisten, aber auch um den externen Zugriff auf lizenzierte Dokumente für eingeschriebene Nutzer der Bayerischen Staatsbibliothek sicherzustellen, läuft der Service über den HAN Proxy[12] der Bibliothek.

Ziel dieses Beta-Tests ist es, ein erstes Feedback von den Nutzern über diese neue Form von Recherche, dieses vollkommen andere Herangehen an thematische Suche zu bekommen. Zugleich ist es notwendig, eine Vorstellung zu entwickeln, wie diese neue Technologie zukünftig vielleicht einmal ein Teil des Standardservice der Bibliothek werden könnte.

Anwendungsperspektiven

Abschließend seien für die weitere Entwicklung und die Perspektiven eines produktiven Einsatzes drei Aspekte erwähnt: Oben wurde schon angemerkt, dass Yewno bislang nur auf englischsprachige Texte Anwendung findet. Für europäische Bibliotheken ist damit die Datenbasis zu schmal, aber letztlich nicht nur für europäische Bibliotheken. Forschung ist eigentlich schon immer international angelegt und multilingual. Deshalb ist bei Yewno die Integration weiterer Sprachen in konkreter Planung.



Gezielt können im semantischen Netz Konzepte ausgewählt werden, deren Beziehung zum Ausgangskonzept schwächer ist, dafür aber vielleicht neue und unerwartete Zusammenhänge offenbart.

Der zweite Aspekt künftiger Weiterentwicklung ist die Verknüpfung zum konventionellen Printangebot einer Bibliothek. Zum einen geht es hier um den Zugang zu gedruckten Zeitschriften und Büchern, zu denen Yewno die paral-

lelen digitalen Texte zwar verarbeitet hat, die Bibliothek aber keine Lizenz besitzt. Zum anderen steht die wichtige Frage im Raum, wie sich Yewnos Technik auch auf den reinen Printbestand anwenden lässt. Von der Gewinnung von Volltextdaten allein zur Indexierung und Analyse bis zur Nutzung vorhandener Sacherschließung ist vieles denkbar. Aber noch ist das ein Feld kommender Forschung.

Eine letzte Option ist in Yewno schon angelegt: Die Technik wurde bereits erfolgreich im biomedizinischen Bereich eingesetzt und derzeit ist auch ein spezialisierter Einsatz im Finanzsektor in Vorbereitung.[13] Yewno könnte auch eines der Tools werden, das Bibliotheken als Partner von Forschung und Wissenschaft in spezialisierten Bereichen z. B. in den Fachinformationsdiensten zum Einsatz bringen.

Werden zwei Konzepte ausgewählt, wird der „semantische Raum" zwischen ihnen angezeigt, mit Themen, die beide Konzepte verbinden.

Der Ausblick auf Entwicklungsperspektiven zeigt, dass vermutlich noch ein guter Weg zu gehen ist, bis der Einsatz einer „inference engine" wie Yewno zum Standardservice der Bibliotheken werden wird. Skeptiker können in Frage stellen, ob sich ein solcher Aufwand lohnen wird. Im Hinblick auf die eingangs angestellten Überlegungen zum Paradigmenwechsel von der Schriftkultur zur digitalen Kultur scheint es für Bibliotheken meines Erachtens eher angeraten, diese Entwicklung aktiv mitzuverfolgen und mitzugestalten. Semantische und inferentielle Rechercheangebote, ob sie nun von Yewno kommen oder eines Tages auch von anderen Anbietern bereitgestellt werden, mögen den Bibliotheken nicht alleine in der sich schnell wandelnden digitalen Welt die Zukunftsfähigkeit garantieren, aber sie könnten ein entscheidender Baustein dafür sein.

Fußnoten

1. Vgl. Lobin, Henning, Engelbarts Traum, Frankfurt a.M. 2014, Kap. 4.4 und Kap. 4.5.
2. Vgl. Lobin 2014, S. 99.
3. Ceynowa, Klaus, Der Text ist tot – es lebe das Wissen, in: Hohe Luft: Philosophie Zeitschrift 2014 (1), S. 53 - 57, S. 54.
4. Vgl. Ceynowa 2014 S. 54 f.
5. Schon früher wurden oft nur Teile von Werken be- und verarbeitet, gerade auch im wissenschaftlichen Kontext, doch waren diese Teile nicht separat verfügbar.
6. Ceynowa 2014, S. 55.
7. http://yewno.com/about/
8. Vgl. Lobin, 2014, S. 159.
9. Als Beispiel dafür kann SLUB Semantics betrachtet werden, welches auf der Basis von Wikipedia Konzepten Katalogaufnahmen um diese Konzepte anreichert. Vgl. dazu auch Bonte, A. et al.: „Brillante Erweiterung des Horizonts: Eine multilinguale semantische Suche für den SLUB-Katalog", in: BIS 4, 4 (2011): 210–213.
10. Vgl. dazu Gilson, Tom und Strauch, Katina, „ATG Interviews Ruggero Gramatica", in: Against the Grain, 2016, S. 64 f.; weitere Informationen zum technischen Hintergrund von Yewno entstammen einem noch nicht veröffentlichten Papier von Ruggero Gramatica und direkter Nachfrage bei Yewno.
11. Vgl. Gilson und Strauch, 2016, S. 66 f.
12. HAN steht für Hidden Automatic Navigator. Dabei handelt es sich um einen Reverse Proxy Server der Firma H+H Software GmbH, der den kontrollierten externen Zugriff auf lizenzierte elektronische Medien ermöglicht.
13. Vgl. Gilson und Strauch, 2016, S. 66 f.

DER AUTOR:
Dr. Berthold Gillitzer ist stellvertretender Leiter der Abteilung Benutzungsdienste in der Bayerischen Staatsbibliothek.