Wissen dynamisch organisieren

Einsatzfelder einer automatischen semantischen Analyse verteilter Informationen

Dr. Jörg Wurzer, iQser AG[1]
Prof. Dr. Stefan Smolnik, European Business School (EBS)[2]

Zusammenfassung. *Die semantische Analyse von verteilten Informationen macht eine dynamische Organisation von Wissen möglich. Zum einen wird die Notwendigkeit von Suchanfragen reduziert, da wichtige Informationen in einem gegebenen Kontext ermittelt und bereitgestellt werden. Zum anderen wird kodifiziertes Wissen aus Informationen identifiziert und verfügbar gemacht. Der Beitrag reflektiert den Bedarf an automatischer semantischer Analyse sowie entsprechender Anwendungsfelder in der Praxis.*

1. Heutiger Zugang zu Informationen

Es gibt in konventionellen Systemen im Wesentlichen zwei Möglichkeiten, Informationen zu finden und abzurufen: Volltextsuche oder Verzeichnishierarchien. Beide Vorgehensweisen haben sich zwar bewährt, weisen aber auch erhebliche Nachteile auf [WS08].

Vorteil der Volltextsuche ist die Einfachheit bei der Bedienung, die sich beispielsweise bei Google bewährt hat. Schon die Eingabe eines Stichworts oder einer Stichwortkombination führt zu einem Ergebnis. Allerdings weißt die Volltextsuche auch Nachteile auf:

- Die Qualität des Ergebnisses ist abhängig von der Auswahl und Kombination der Suchwörter. Ist das Suchwort ein häufig genutztes Wort, ist das Suchergebnis sehr umfangreich. Dieses wird in Form einer langen Liste präsentiert, deren Einträge der Anwender einzeln bewerten muss. Das kann viel Zeit in Anspruch nehmen.
- Die Volltextsuche berücksichtigt allein das Vorkommen eines oder mehrerer Wörter in einem Text. Die Bedeutung des Worts im Gesamtzusam-

[1] Chlupfgasse 2, 8303 Basserdorf, Schweiz, E-Mail: joerg.wurzer@iqser.net, Twitter: http://twitter.com/jwurzer

[2] Rheingaustr. 1, 65375 Oestrich-Winkel, Email: stefan.smolnik@ebs.edu

menhang des Textes wird jedoch nicht ausgewertet. So kann es auch ein für den Gesamttext völlig nebensächlichen Sachverhalt beschreiben.

Die Bestimmung der Relevanz der Suchtreffer ist eine große Herausforderung für Suchmaschinen. So werden die Absicht des Suchenden und der Zusammenhang des Informationsbedarfs berücksichtigt. Google hat hier eine Sortierung der Suchergebnisse eingeführt, die Ergebnisse in Form einer Webseite daran bewertet, wie oft sie von anderen Webseiten in Form eines Links referenziert wird. Das ist kann als ein Behelf eingestuft werden, denn Vorbild ist die Bedeutungsbewertung einer wissenschaftlichen Arbeit mittels der Häufigkeit, mit der sie zitiert wird.

Vorteil von Verzeichnishierarchien ist, dass sie von einer Person erstellt werden können, welche die Logik an der Bedeutung der klassifizierten Inhalte und an deren Verwendung ausrichten kann. Es handelt sich folglich um eine durchdachte Struktur, die den Interessen des Anwenders Rechnung trägt. Allerdings treten auch hier eine Reihe an Nachteilen auf:

- Die Informationsobjekte, seien es Dokumente, E-Mails oder Webseiten in einem Portal, werden auf eine definierte Weise einer Verzeichnishierarchie zugeordnet. In der Praxis gibt es jedoch oft verschiedene Betrachtungsweisen auf einen Informationsbestand. Mal interessiert zum Beispiel die Kundensicht, dann eine Produktsicht und schließlich eine personenbezogene Sicht auf die Informationen. Ein Verzeichnis muss sich jedoch auf eine konkrete Struktur festlegen. Veränderungen, die sich mit der Zeit in der Organisation oder den Aufgabenstellungen ergeben, sind nur mühsam zu implementieren.
- Die meisten Informationen sind in mehreren Zusammenhängen von Bedeutung. Eine mehrfache Zuordnung von Informationsobjekten in einem Verzeichnisbaum führt aber in der Regel zu einer redundanten Speicherung von Informationsobjekten. Dies führt nicht nur zu erhöhtem Ressourcenbedarf, sondern wirft auch Probleme der Compliance auf und ist hinderlich bei der Aktualisierung der Informationsobjekte.
- Die Qualität des Verzeichnisbaums ist abhängig vom Autor. Sie spiegelt den aktuellen Stand des organisationalen Wissens wider. Es mag jedoch Informationsobjekte geben, die sich nicht in diesem Verzeichnisbaum einordnen lassen. Dafür gibt es dann lediglich einen Behelf mit „Sonstigen Dateien“ oder ähnliches.
- Die Informationsobjekte in einem Verzeichnis beschränken sich in der Regel auf Dateien oder Objekten eines bestimmten Typs, zum Beispiel E-Mails. Es gibt bei den meisten Systemen keine atomare Auflösung der Informationsobjekte und deren Zuordnung. Das hat einen großen Nachteil für den Überblick über und die Organisation von Informationen.

2. Jede Information über verschiedene Kontexte erreichbar

Die nachfolgend vorgestellte Technologie bietet eine Alternative zu den beiden diskutierten Ansätzen (vgl. Abb. 1). Informationsobjekte werden nach inhaltlichen Kriterien vielfach vernetzt. So entsteht ein semantisches Netzwerk von Informationsobjekten, die jeweils durch den Kontext zu finden sind, der im augenblicklichen Interessenfokus des Anwenders steht. Wenn sich beispielsweise ein Anwender für Geschäftsabschlüsse, Verträge, Produkte, Personen, Servicefälle eines bestimmten Kunden interessiert, so sind alle diese Informationen dem Kunden zugeordnet. Dabei können die Informationen aus verschiedenen Systemen stammen und müssen nicht in eine zentrale Datenbank eingespeist werden, in denen die Verknüpfungen fest kodiert wären. Jeder beliebige Kontext ist denkbar: eine Person, ein Projekt, ein ganz bestimmtes Dokument, ein Meeting, eine Aufgabe, eine Nachricht, etc. Die vorgestellte Technologie zeigt in jedem Fall Informationsobjekte an, die in diesem Zusammenhang von Bedeutung sind. Ein Personenprofil kann auf Publikationen im und außerhalb eines Unternehmens hinweisen, auf Projekte, an die eine Person teilgenommen hat, an Kontakten zur Wissenschaft oder zur Wirtschaft. Oder ein Dokument kann auf den Autor verweisen, auf andere Artikel, die das Thema vertiefen, oder zu Themenbereichen, die zum Beispiel in der Forschung und Entwicklung eines Unternehmens definiert wurden [PB06].

Damit ergibt sich ein neues Vorgehen für den Aufruf. An Stelle über Begriffe zu suchen oder einen Verzeichnispfad zu verfolgen, wird der Kontext aufgerufen, der im augenblicklichen Interesse des Anwenders steht. Auch ist eine Kombination von Kontexten möglich, um Informationen noch strikter einzugrenzen. Der Anwender hat so im Überblick, welche Informationen innerhalb eines Kontextes von Bedeutung sind. Damit kann er ohne zu suchen Zusammenhänge im Informationsbestand entdecken und Schlussfolgerungen hin zu neuem Wissen ziehen. So kann er zum Beispiel erkennen, welche Expertisen ein bestimmter Mitarbeiter hat, die sich in Projekterfahrungen, Publikationen und persönlichen Kontakten sowie der Ausbildung auszeichnet. Jeder Kontext ist bei dieser Technologie kein Abstraktum, sondern ein konkretes Informationsobjekt. Damit ist es möglich, durch den Informationsbestand über die Verknüpfungen von Informationsobjekt zu Informationsobjekt zu navigieren. Anders als in einem Verzeichnissystem ist das semantische Netzwerk nicht hierarchisch, sondern führt über die Verknüpfungen immer wieder zu weiteren Informationsobjekten. Dies ist hilfreich sowohl für die Exploration von Informationsbeständen als auch für die gezielte Suche.

Die vorgestellte Technologie stellt einen neuen Ansatz für den Zugang und die Organisation von Informationen dar: Alle relevanten Informationen werden in einem gegebenen Kontext automatisch identifiziert und zugeordnet. Eine Suche wird dadurch überflüssig. Im Unterschied zu anderen Ansätze erfolgt diese Organisation von Informationen vollautomatisch, dynamisch und selbstoptimierend ohne vorausgehende Ontologiedefinitionen oder semantische Annotationen bzw. Aufbereitung eines Datenbestands [HKRS08] [Wic07]. Die Kombination leichtgewichtiger Analyseverfahren sorgt dabei sowohl für eine hohe Präzision als auch einen geringen Ressourcenverbrauch.

3. Der Uniform Information Layer

Um das semantische Informationsnetzwerk zu erstellen, werden zunächst alle Informationsobjekte aus verschiedenen Systemen eingebunden (vgl. Abb. 1). Dies wird durch den Uniform Information Layer realisiert. Mit Hilfe einer ContentProvider-API wird für jede Datenquelle mindestens eine Schnittstelle implementiert, welche die enthaltenen Objekte in ein generisches Objektformat transformiert und umgekehrt. Dabei spielt es keine Rolle, ob das Informationsobjekt ein unstrukturiertes Dokument oder ein strukturierter Datensatz aus einer relationalen Datenbank oder einer Applikation ist. Die Schnittstellenimplementierung ist nur mit einem geringen Aufwand verbunden, da die Analytik in der Middleware integriert ist.

Verteilte Informationen werden mittels des Uniform Information Layer zentral und mit harmonisierter Beschreibung verfügbar. Der Uniform Information Layer und der zugehörige Server ermöglichen folglich sowohl eine übergreifende Volltextsuche als auch eine komplexe Suche mit Hilfe semantischer Beschreibungen wie Metadaten und Informationstyp. Ebenso lassen sich zu jedem Informationsobjekt die Relationen abrufen.

Das semantische Netzwerk für diese Relationen ist Ergebnis eines dreistufigen Analyseverfahrens. Das Netzwerk passt sich dabei immer der aktuellen Informationslage an: Sobald ein neues Informationsobjekt erzeugt, verändert oder gelöscht wurde, wird das semantische Netzwerk in Echtzeit aktualisiert, ohne das der gesamte Informationsbestand neu berechnet werden muss. Das ist bei sehr großen Informationsbeständen eine notwendige Systemeigenschaft. Prozesse können mit der vorgestellten Technologie direkt verbunden werden, da der Server auf einer ereignisgesteuerten Architektur basiert. Folglich können einzelne Prozessschritte überwacht und gesteuert werden. Dabei entstehende Aufgaben werden automatisch mit kontextsensitiven Informationen ergänzt, die für die Bearbeitung benötigt werden.

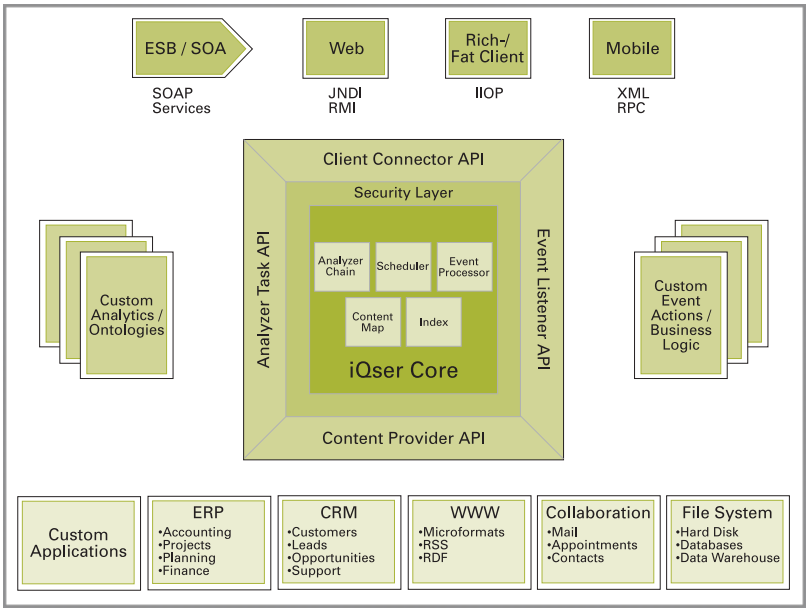

Abb 1: Informationsarchitektur

4. Kombination von drei Analyseverfahren

Für die Erstellung des semantischen Netzwerks werden drei Analyseverfahren kombiniert. Die Analyseprozesse werden von einem Protokoll gesteuert, das jede Veränderung im Informationsbestand aber auch alle Benutzeraktionen mit dem System beinhaltet.

1. Syntaktische Analyse: Darunter wird eine Verknüpfung von Informationsobjekten aufgrund von Metadaten, die für Beziehungen relevant sind, verstanden. Diese Metadaten werden als Schlüsselattribute bezeichnet, die in den ContentProvidern als solche klassifiziert werden.. Das kann im Falle einer E-Mail ein Schlüsselattribut zum Beispiel Absender oder Adressat sein. Zu diesem Attribut wird nach einer Entsprechung im Informationsbestand gesucht. So kann eine E-Mail-Adresse in einem Personenprofil oder einem Dokument vorkommen. Im Ergebnis wird die E-Mail-Korrespondenz zu einer Person gebündelt und ist ohne manuelle Zuordnung abrufbar. Jeder Informationsobjekttyp kann mit beliebig vielen Schlüsselattributen versehen werden. In der Regel sind dies strukturierende Elemente, wie zum Beispiel eine Kundennummer oder eine Projektreferenznummer.

2. Musteranalyse: In diesem Analyseschritt wird ein Mustervergleich zwischen verschiedenen Inhalten durchgeführt. Dafür zieht das Analyseverfahren die bedeutungstragenden Begriffe aus einem Text heraus und transformiert diese Wortmenge in eine Abfrage, deren Ergebnis ähnliche Informationsobjekte auflistet. Jedes Informationsobjekt wird mit einem Ähnlichkeitswert zwischen 0 und 1 versehen. Manuell erzeugte Verknüpfungen erhalten den Wert 1, während der Wert 0 ein Grenzwert für die völlige Verschiedenheit ist. Der Wert 1 ist für manuell erstellte Verknüpfungen reserviert. Die bedeutungstragenden Begriffe werden auf der Basis ihrer Signifikanz in einem Informationsobjekt berechnet und selektiert: Der Ähnlichkeitswert wird dann auf Grund der Anzahl und Häufigkeit übereinstimmender Wörter berechnet. Auch der Wortabstand spielt eine Rolle. Dieses Verfahren untersucht eine thematische Nähe und nutzt dabei die Tatsache, dass durch die signifikanten Begriffe ein Thema komplex definiert ist und eine Themenverwandtschaft entsprechend genau festgestellt werden kann [HQW06].
3. Semantische Analyse: Dieses Verfahren beruht auf der Erkenntnis aus der Sprachphilosophie, dass die Bedeutung von Sprache in ihrem Vollzug entsteht und auf immer wieder neue Weise definiert wird. Die Bedeutung von Informationsobjekten entsteht analog durch ihre Verwendung. Sie verändert sich im Laufe der Zeit und wechselt den Kontext ihrer Verwendung oder die Relevanz bis zur Bedeutungslosigkeit bei einem veralteten Dokument. Aus diesem Grund werden alle Aktionen des Anwenders von der Erstellung eines Informationsobjekts über die Änderung bis zur Löschung protokolliert. Selbst der Aufruf eines Informationsobjekts wird protokolliert. Ein spezieller Algorithmus berechnet nach jedem Protokolleintrag die Beziehung von Informationsobjekten neu. Werden zum Beispiel zwei Informationsobjekte oft in einer Folge aufgerufen oder bearbeitet, kann das System daraus schließen, dass ein Bedeutungszusammenhang besteht. Dieser wächst, wenn sich das Muster wiederholt. Die Relevanz einer Verknüpfung kann auch wieder abnehmen, wenn andere Verknüpfungen häufiger verwendet werden. So erhält jede Verknüpfung eine zugeordnete Relevanz, die es dem Anwender erleichtert, eine Auswahl zu treffen, wenn zum Beispiel mehrere Dokumente einer Person zugeordnet wurden.

Es kann der Fall eintreten, dass zwei Informationsobjekte aufgrund zweier oder aller drei Verfahren miteinander verknüpft werden. Eine weitere Option ist die manuelle Verknüpfung. Es kann auch eintreten, dass es mehrere Gründe gibt, warum zwei Informationsobjekte über die syntaktische Analyse miteinander verbunden sind. Deshalb kann es zwischen zwei Informationsobjekten nicht nur eine Verknüpfung sondern potentiell n Verknüpfungen geben. Die Gesamtgewichtung der Verknüpfung wird aus den Einzelgewichtungen berechnet. Optional kann dem Anwender des Systems angezeigt

werden, warum eine Verknüpfung entstanden ist. Damit bleibt das System nachvollziehbar.

Über eine offene Schnittstelle kann die Analysekette erweitert werden, um kundenspezifische Anforderungen zu berücksichtigen oder andere Verfahren zu kombinieren.

5. Abfrage von verknüpften Informationen

Die vorgestellten Konzepte und Verfahren ermöglichen, dass Informationsobjekte über vielfältige Kontexte erreichbar sind. Tritt der Fall ein, dass es zu einem Informationsobjekte sehr viele verknüpfte Informationsobjekte gibt, zum Beispiel zu einem Projekt, kann der Anwender die Kontextsuche einsetzen. Dabei wird eine Abfrage auf die Informationsobjekte angewendet, die in Beziehung zu einem bestimmten Informationsobjekt stehen. Diese Selektion mit Hilfe von Verknüpfungen kann in beliebiger Komplexität durchgeführt werden [AABDNPT07] [Wur08].

Jede Veränderung im Informationsbestand löst ein Ereignis aus, das im Server zum Anlass für eine Aktion verwendet werden kann. Dazu zählt auch ein neu entdeckter inhaltlicher Zusammenhang. Eine daraufhin ausgelöste Aktion kann die Benachrichtigung eines Anwenders sein oder aber das System verändert oder erstellt neue Informationen. So kann eine Serviceanfrage eines Kunden automatisch beantwortet, dem entsprechenden Sachbearbeiter zugeleitet und ein Ticket für die Beantwortung der Anfrage erstellt werden. Komplexe Prozesse lassen sich auch in Verbindung mit bereits eingesetzten Lösungen des Business Process Management verbinden und steuern. Neben der systemübergreifenden Prozesssteuerung in einem Unternehmen lassen sich auch Forschung, Wettbewerb und Meinungstrends automatisch überwachen. Alle Informationen sind verfügbar, sobald sie veröffentlicht werden, sei es in einem Unternehmensnetz oder im Internet. Ausgangspunkt der Überwachung müssen nicht ein oder mehrere Suchbegriffe sein, sondern zum Beispiel vollständige Dokumente, die ein Thema beschreiben. Letzteres kann Ausgangspunkt einer Recherche sein und relevante Informationen aus einem großen Informationsbestand „wie ein Magnet anziehen“. Denkbar wären die Beschreibung von Forschungsthemen, Produkten oder Märkten. Auch Patentschriften lassen sich so mit dem Stand der Forschung abgleichen.

Literatur

[AABDNPT07] Antoniou, G.; Aßmann, U.; Barglio, C.; Decker, S.; Nenze, N.; Patranjan, P.-L.; Tolksdorf , R. (Hrsg.): Reasoning Web. Third International Summerschool 2007. Springer. 2007.

[HKRS08] Hitzler, P.; Krötsch, M.; Rudolph, S.; Sure, Y.: Semantic Web. Grundlagen. Springer. 2008.

[HQW06] Heyer, G.; Quasthoff, U.; Wittig, T.: Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse. W3L-Verlag. 2006.

[PB06] Pellegrini, T.; Blumauer, A. (Hrsg.): Semantic Web: Wege zur vernetzten Wissensgesellschaft. Springer. 2006.

[Wic07] Wichmann, G.: Entwurf Semantic Web: Entwicklungen, Werkzeuge, Sprachen. VDM Verlag Dr. Müller. 2007.

[WS08] Wurzer, J.; Smolnik, S.: Towards an automatic semantic integration of information. Maicher, L.; Garshol, L. M. (Hrsg.): Subject-centric Computing. 2008.

[Wur08] Wurzer, J.: New Approach for Semantic Web by Automatic Semantics. European Semantic Technology Conference. Vienna. 2008.