September 2009

Ausgabe #3-09 Deutschland Euro 12,00 ISSN: 1864-8398

DOK.

Technologien, Strategien & Services für das digitale Dokument

Microblogging goes Business
Wissensnetze als Basis für Enterprise 2.0
Datenschutz für Kollaborationsprozesse

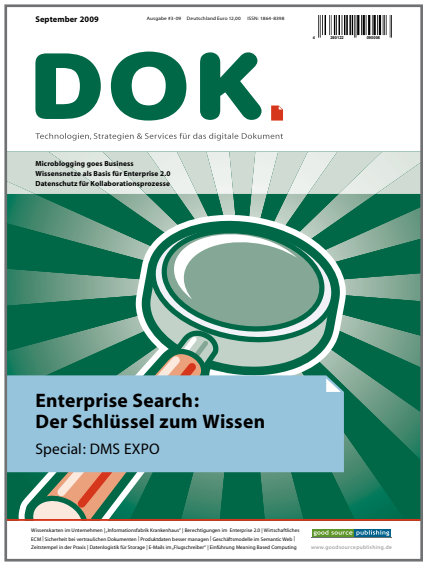
Enterprise Search: Der Schlüssel zum Wissen

Special: DMS EXPO

Wissenskarten im Unternehmen | „Informationsfabrik Krankenhaus“ | Berechtigungen im Enterprise 2.0 | Wirtschaftliches ECM | Sicherheit bei vertraulichen Dokumenten | Produktdaten besser managen | Geschäftsmodelle im Semantic Web | Zeitstempel in der Praxis | Datenlogistik für Storage | E-Mails im „Flugschreiber“ | Einführung Meaning Based Computing

good source publishing

www.goodsourcepublishing.de



Such-Trends im World Wide Web

Das „Netz“ entwickelt sich zu einer Art globalem Gedächtnis. Schon lange ist kein Mensch mehr in der Lage, diese Informationsfülle im Blick zu behalten. Suchmaschinen wie Google dienen uns als Wegweiser und Orientierungshilfe bei unserer Suche nach Information. Doch die Web-Suchmaschinen, wie wir sie heute kennen, entwickeln sich weiter, selbst wenn für den Anwender viele Entwicklungen nicht direkt auf den ersten Blick erkennbar sind. Waren vormals nur Textdokumente hinreichend gut erschlossen, gewinnt die Suche in multimedialen Daten wie Bildern, Audio- oder Videoclips zunehmend an Bedeutung. Die Suche richtet sich zunehmend an den Bedürfnissen des einzelnen Anwenders aus und entwickelt sich hin zu einem persönlichen Informationsassistenten. Nicht nur Informationen, auch Datenbestände rücken in den Fokus der Suchmaschinen. Diese Daten werden inhaltlich zum Bestandteil eines globalen semantischen Netzwerks, in dem heterogene Datenquellen automatisch miteinander verknüpft und in Bezug zueinander gesetzt werden. Auf diese Weise entwickelt sich die traditionelle Web-Suche hin zu einer explorativen Sucherfahrung, bei der das Finden und nicht nur wie bislang die oft erfolglose Suche im Vordergrund steht.

Am 13. März 2009 feierte das World Wide Web am europäischen Forschungszentrum für Teilchenphysik CERN seinen zwanzigsten Geburtstag. Der britische Physiker Sir Tim Berners-Lee, der als „Vater“ des Web gilt, schrieb irgendwann im März 1989 ein Memorandum mit dem Titel „Information Management: A Proposal “ an seinen Vorgesetzten, das als Startsignal für das wenige Jahre später bereits weltumspannende multimediale Informationssystem angesehen wird. Der unglaubliche Erfolg, der dem Web seit seinem Start beschieden war, leitet sich aus der prinzipiellen Offenheit seiner Architektur her. Kein Hersteller kann im Web einen Standard diktieren oder den Zugang zu dessen Quellen rechtlich beschränken. Unabhängig von Computerarchitekturen, Betriebsystemen und Netzwerkstrukturen erlaubt

www.hpi-web.de

Dr. Harald Sack ist Senior Researcher mit Schwerpunkt Multimedia Retrieval, Semantic Web-Wissensrepräsentation und Semantic Enabled Retrieval am **Hasso Plattner-Institut für Softwaresystemtechnik GmbH** in Potsdam. Das von Hasso Plattner, einem der SAP-Gründer, gegründete Institut arbeitet als universitäres Exzellenzzentrum eng mit der Wirtschaft zusammen und hat sich über die Grenzen Deutschlands hinaus innerhalb weniger Jahre einen erstklassigen Ruf erarbeitet.

das Web die Einbindung der unterschiedlichsten Dokumenten-, Daten- und Informationssammlungen. Im Gegensatz zu traditionellen Informationssystemen gestattet das Web eine beliebige Verknüpfung der gespeicherten Ressourcen miteinander über Hyperlinks, die eine Navigation über Dokumentengrenzen hinweg ermöglichen und so die Linearität der ursprünglichen Textdokumente durchbrechen.

Dabei hält das rasante Wachstum des Web ungebrochen an. Im Juli 2008 vermeldete Google's Weblog, dass der Google Web-Crawler, dessen Aufgabe darin besteht, möglichst alle erreichbaren Web-Dokumente für den Index der Suchmaschine zusammenzusammeln, mehr als eine Billion (1012) Dokumente besucht habe. Diese Zahl betrifft auch nur einen Teil der insgesamt über das WWW erreichbaren Dokumente und lässt dabei zugriffsbeschränkte Informationen, die meisten dynamisch generierten Inhalte und die Informationsbestände der zahlreichen Intranets außen vor. Hinter dem erreichbaren „Surface Web" verbirgt sich die noch als weitaus größer geschätzte Zahl von Informationsressourcen des sogenannten „Dark Web", das noch auf seine Erschließung wartet. Wären wir alleine auf handrecherchierte Web-Portale und Hyperlinkverknüpfungen angewiesen, keiner fände sich mehr im Web zurecht. Suchmaschinen wie Google sind es, die mit ihren automatisierten Web-Crawlern das Web durchkämmen, Dokumente erfassen und auswerten und uns damit überhaupt erst den Zugriff auf die Information ermöglichen.

Die grundlegende Aufgabe einer Web-Suchmaschine besteht in der Suche nach Dokumenten, die bezogen auf einen Suchbegriff als relevant eingestuft werden. Dabei kommen zunächst Methoden des klassischen Information Retrievals zum Einsatz, deren Grundlagen bereits in den 1960er-Jahren gelegt wurden. Die vom Web-Crawler erfassten (Text-)Dokumente durchlaufen dabei eine Prozesskette unterschiedlicher Analyseverfahren,

an deren Ende für jedes Dokument eine Liste aussagefähiger Deskriptoren mit zugehöriger Relevanzgewichtung ermittelt wird, nach denen dann der gesamte Dokumentenbestand gezielt durchsucht werden kann. Die Vernetzung der Web-Dokumente untereinander dient dabei als Indikator für die Relevanz der einzelnen Dokumente: Je mehr Hyperlinks aus unterschiedlichen Quellen auf ein bestimmtes Dokument verweisen, desto höher ist die Wahrscheinlichkeit, dass es sich um ein relevantes, d. h. von zahlreichen Web-Autoren „empfohlenes" Dokument handelt. Dabei zählt ein Link umso mehr, wenn er selbst von einem relevanten Dokument stammt. Dieses aus dem Bereich des wissenschaftlichen Zitierens bekannte Prinzip wurde durch Google's PageRank -Algorithmus auf den Bereich der Web-Suche übertragen und trug durch seine qualitativ hochwertigen Suchergebnisse zum Erfolg des heute weltweit größten Suchmaschinenbetreibers bei.

Doch die mediale Vielfalt des Web beschränkt sich nicht alleine auf Text. Bilder, Audiodateien und insbesondere Videos bestimmen mehr und mehr das Erscheinungsbild. Dabei sind die bekannten Suchmaschinen bei der Suche in multimedialen Daten immer noch auf Textdaten angewiesen, die direkt über Analyseverfahren aus den Originaldaten gewonnen oder zusammen mit den Originaldaten abgespeichert vorliegen. Diese beschreibenden Metadaten können dann wieder mithilfe traditioneller Information-Retrieval-Verfahren durchsucht werden und liefern als Ergebnis die damit verbundenen Originaldaten zurück. Genutzt wird dabei etwa der sogenannte „Link-Kontext" multimedialer Daten, die über Hyperlinks in WWW-Dokumente eingebettet werden, d. h. Text, mit dem ein Verweis auf eine Bild-, Audio- oder Videodatei beschrieben wird. Daneben existieren zahlreiche Analyseverfahren, mit denen sich zusätzliche Metadaten aus den multimedialen Originaldaten gewinnen lassen. In Bilddaten lassen sich zunächst bildbestimmende Parameter wie beispielsweise dominante Farben oder Farb- und Helligkeitsverteilung bestimmen, auf deren Basis einander ähnliche Bilddokumente bestimmt werden können. Die Suche erfolgt dabei meist anhand eines Bildbeispiels, zu dem ähnliche Dokumente gefunden werden sollen (Query by Example). Zusätzliches Benutzerfeedback über die Qualität der Suchergebnisse hilft, die Suchverfahren weiter zu verbessern, wobei Methoden des maschinellen Lernens zum Einsatz gelangen. Ebenso lassen sich aus Audiodaten inhaltliche Parameter und Charakteristika gewinnen, die als Grundlage einer ähnlichkeitsbasierten Suche dienen können.

Videodaten, bei denen ebenso wie bei Audiodaten zusätzlich noch die Zeit als eigene Dimension und damit verbunden die Änderung des Inhalts über die Zeit von erheblicher Bedeutung ist, stellen die inhaltliche Analyse der Daten vor neue Herausforderungen und erhöhen die Komplexität des Analysevorgangs. Ein wichtiges Problem ist dabei die Identifikation konkreter Objekte in multimedialen Daten. So lässt sich beispielsweise das Erkennen von Personen im Sinne der Unterscheidung von Bild- und Videodokumenten mit oder ohne Personen bereits heute zufriedenstellend bewältigen. Dabei werden bildbestimmende Merkmale aus den zu untersuchenden Dokumenten analysiert und einem mithilfe von Beispielgesichtern trainierten Klassifikationssystem zur Beurteilung vorgelegt. Dieses entscheidet, ob in der zu klassifizierenden Bilddatei das Gesicht einer Person zu sehen ist oder nicht. Videodokumente werden zu diesem Zweck hierarchisch in einzelne zeitliche Abschnitte bis hin zu Einzelbildern zerlegt. Soll dagegen eine Person identifiziert werden, müssen die als Gesicht erkannten Bildanteile in zum Teil noch sehr fehler- und störungsanfälligen Verfahren mit Referenzdatenbanken abgeglichen werden. Eine weitere vielversprechende Informationsquelle ist die Erkennung natürlicher Sprache in Audio- und Videodateien und deren Übersetzung (Transkribierung) in Textdaten. Hier kommen zunehmend Spracherkennungssysteme zum Einsatz, bei denen auf ein aufwendiges Training des Systems mit dem jeweiligen Sprecher vor der eigentlichen Analyse verzichtet werden kann. Als problematisch erweisen sich jedoch oft unzureichende Aufnahmequalität von Amateuraufnahmen und fehlende Sprecherausbildung. Während ein Spracherkennungssystem den ausgebildeten Nachrichtensprecher bei einer Tonaufnahme in Studioqualität hinreichend gut erkennen kann, versagt es hingegen in Alltagssituationen mit zahlreichen Störgeräuschen.

Die automatische inhaltliche Analyse von multimedialen Dokumenten stößt heute schnell an ihre Grenzen. Noch immer ist das bild- und informationsverarbeitende System der menschlichen Wahrnehmung jeder automatisierten Analyse überlegen, insbesondere dann, wenn komplexe Transfer- und Abstraktionsleistungen anfallen. Daher werden die zur Informationssuche notwendigen Metadaten heute immer noch mit manuellen Verfahren gewonnen. Man unterscheidet prinzipiell zwischen autoritativen Metadaten, d. h. Metadaten, die vom Autor einer Informationsressource selbst bzw. von einem Experten erstellt wurden und deren Qualität daher über jeden Zweifel erhaben ist, und nicht-autoritativen Metadaten, die beispielsweise von den Benutzern einer Informationsressource beigesteuert werden, deren Qualität aber nicht sicher beurteilt werden kann. Dennoch treten benutzergenerierte Metadaten immer mehr in den Vordergrund, da nur durch die immense Zahl an potenziellen Benutzern auch eine Chance besteht, signifikante Teile der Informationsressourcen im WWW einer manuellen Analyse und Auszeichnung mit Metadaten (Annotation) unterziehen zu können. Web-2.0-Dienste und Anwendungen des Social Web erfreuen sich seit einigen Jahren einer enormen Popularität und verhalfen dem Web auf einer breiten Basis zum Erfolg. Diese Dienste erlauben auch dem Nichtfachmann die Produktion und Bereitstellung eigener Inhalte und Daten im Web. In diesem Zusammenhang sind für die Informationssuche im WWW

die sogenannten Bookmarking-Dienste (Collaborative Tagging Systems) von großer Bedeutung, da sie ihren Benutzern die Möglichkeit geben, eigene Schlüsselworte, Kommentare und Beiträge mit vorhandenen Web-Dokumenten zu verknüpfen, um diese anschließend mithilfe der eigenen Anmerkungen wiederzufinden. Diese benutzergenerierten Metadaten erlauben in Verbindung mit dem sozialen Netzwerk der Benutzer eine neue Qualität der inhaltlichen Analyse sowie der damit erzielten Suchergebnisse. Wird das soziale Netzwerk des Benutzers in die Suche mit eingebunden, können Anmerkungen und Metadaten eines Freundes ein höheres Gewicht erzielen als die Anmerkungen eines Fremden.

Personalisierung ist ein weiterer wichtiger Trend im Bereich der Suchmaschinen. Suchergebnisse, die für den einen Benutzer relevant sind, müssen dies für den anderen noch lange nicht sein. Jeder besitzt seine eigenen, persönlichen Informationspräferenzen und -bedürfnisse, die von einer Suchmaschine berücksichtigt werden können. Ein Weg, dieses Problem anzugehen, liegt in der Auswertung von Benutzerinteraktionen aus den vorhandenen Log-Dateien. So können Benutzer anhand der von ihnen gesuchten Begriffe in unterschiedliche Gruppen eingeordnet werden, denen die Suchergebnisse mit unterschiedlicher Relevanzgewichtung präsentiert werden. Idealerweise lernt die Suchmaschine bei jeder Benutzung noch dazu, da das vom Benutzer tatsächlich aus der Ergebnisliste ausgewählte Suchergebnis bei einer zukünftigen Suche höher gewichtet und damit weiter vorne in der Ergebnisliste platziert werden kann. Dazu muss der Benutzer allerdings von der Suchmaschine über geeignete Mechanismen identifiziert werden können.

Google als prominentester Vertreter der Web-Suchmaschinen steht in der Kritik, da als Antwort auf eine Suchanfrage oft zu viele nicht relevante Ergebnisse präsentiert werden. Auf der

anderen Seite kann heute kein Mensch mehr beurteilen, ob sich auch tatsächlich alle relevanten Ergebnisse in der Fülle der angebotenen Suchtreffer befinden. Verantwortlich dafür sind unter anderem sprachliche Mehrdeutigkeiten auf unterschiedlichen semantischen Abstraktionsebenen. Ein Wort selbst kann unterschiedliche Bedeutungen haben (Homonymie), so bezeichnet das Wort „Golf" sowohl eine Sportart, eine Automarke als auch einen Meeresarm. Eine klassische schlüsselwortbasierte Suche nach dem Begriff „Golf" führt also auf Dokumente, in denen das Wort in unterschiedlichen Bedeutungen verwendet wird. Aber selbst wenn auf eine einheitliche Bedeutung geschlossen werden könnte, so kann ein Wort auch in unterschiedlichem Kontext und mit unterschiedlicher Absicht (Pragmatik) vom Autor verwendet werden, die mit den Absichten des suchenden Benutzers nicht übereinstimmen muss. Auf der anderen Seite werden Dokumente mit synonymen Wörtern, die denselben Begriff mit anderen Worten beschreiben, über eine schlüsselwortbasierte Suche nicht gefunden. Neben solchen Synonymen können Begriffe auch mit Metaphern und anderen sprachlichen Ausdrucksmitteln umschrieben werden, ohne dass der inhaltlich damit bezeichnete Suchbegriff in einem an sich relevanten Dokument auftauchen muss. Daher liegt der Schluss nahe, dass vielmehr die inhaltliche Bedeutung (Semantik) eines Dokuments und nicht nur die darin verwendeten Zeichenketten im Vordergrund einer inhaltsbasierten Suche stehen müssen.

Während im Bereich des Information Retrievals die Technologien der Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) im Vordergrund stehen, um die Bedeutung von Dokumenten zu erschließen, geht die Initiative des Semantic Web einen anderen Weg. Im Semantic Web werden Informationsinhalte mit maschinenlesbaren, semantischen Metadaten ausgezeichnet, die deren Bedeutung formalisieren. Bei diesen semantischen Metadaten handelt es sich um sogenannte Wissensrepräsentationen (Ontologien), die maschinell gelesen und verstanden werden können. Im Rahmen des Semantic Web werden standardisierte, aufeinander aufbauende Wissensrepräsentationssprachen von unterschiedlicher semantischer Ausdruckskraft definiert, mit denen die Bedeutung der im Web vorliegenden Informationen formalisiert werden kann. Dadurch ergeben sich neue, intelligentere Möglichkeiten der Informationsverabeitung. So lässt sich z. B. durch semantische Analyse auf implizit in einem Text verborgenes Wissen schließen oder neues Wissen aus vorhandenem Wissen herleiten und auf Korrektheit überprüfen. Auf diese Art wird nicht nur eine zielgenauere Suche ermöglicht, sondern vielmehr auch eine assoziativ motivierte Suche, die anhand impliziter Zusammenhänge Naheliegendes erschließt und aufdeckt und so dem Suchenden einen Einblick in vorhandene Informationen gewährt, die er über eine traditionelle Informationssuche nie gefunden hätte.

Diese Art der explorativen Informationssuche ist uns aus unserem täglichen Leben, abseits des Internets, wohl bekannt. Sucht man in einer Bibliothek ein bestimmtes Buch, erhält man im Katalog Auskunft darüber, wo genau sich das gesuchte Buch befindet. Geht man daraufhin zum betreffenden Regal, in dem sich das gesuchte Buch befindet, fallen einem natürlich auch Bücher ins Auge, die sich in der Umgebung dieses Buches, daneben oder auch darüber befinden. Entsprechend einer bestimmten Aufstellungssystematik wurden miteinander vergleichbare Bücher in diesem Regal eingeordnet und vielleicht findet der Benutzer beim Stöbern ein noch viel interessanteres Buch, von dessen Existenz er bislang noch gar nichts wusste. Dieses „zufällige Finden" wird im Englischen auch als „Serendipity" bezeichnet und steht im Mittelpunkt der explorativen Informationssuche. So wird die Informationssuche selbst zu einem Sucherlebnis, angetrieben durch die Neugierde des Benutzers, die sie auf der anderen Seite aber auch zu beflügeln vermag.