Wikis als Mittel zur Ontologieverfeinerung

Lars Bröcker
Fraunhofer-Institut für Medienkommunikation (IMK)
lars.broecker@imk.fraunhofer.de

Abstract: Forschungsgruppen auf der ganzen Welt sind damit beschäftigt, die Vision des Semantic Web Wirklichkeit werden zu lassen. Dafür ist ein wesentlich höherer Grad der Datenstrukturierung und -annotierung nötig, als er im heutigen WWW anzutreffen ist. Es gibt viele verschiedene Ansätze zur Ontologiegewinnung aus bestehenden Datensammlungen, die zur Strukturierung dieser Sammlungen beitragen. Neben den formal (z.B. in Datenbanken) erfassten Kerndaten gibt es aber in vielen Fällen zusätzliche Informationen, die nicht von solchen Verfahren abgedeckt werden können, ihrerseits aber wertvolle Kontextinformationen zur Verfügung stellen können.
Dieser Artikel skizziert einen Ansatz, mit dem dieses nicht formal abgelegte Wissen erfasst, ausgewertet und zur Verbesserung einer maschinell erzeugten Ontologie eines Themengebiets herangezogen werden kann. Als Grundlage dient ein Wiki-System, das auf einer oder mehreren Wissensbasen, z.B. digitalen Archiven, aufsetzt. Die aus der Nutzergemeinschaft einfließenden Informationen werden analysiert und zur Verfeinerung der maschinell erstellten Ontologie verwendet. Ein Relevance-Feedback-Mechanismus dient zur Qualitätskontrolle.

1 Einleitung

Das Semantic Web ist das bisher ambitionierteste Vorhaben des World Wide Web Consortiums (W3C) [BHL01]. Forscher auf der ganzen Welt sind dabei, die nötigen Voraussetzungen zu schaffen, um das äußerst erfolgreiche WWW [1] heutiger Prägung zu ersetzen. Das Semantic Web soll es Menschen und Software-Agenten ermöglichen, den Inhalt von Webseiten auszuwerten und zu nutzen. Eine Kernaufgabe auf dem Weg zum Semantic Web ist daher die Entwicklung von Methoden, um Metadaten zu den Inhalten einer Webseite in einer maschinell auswertbaren Form zur Verfügung stellen zu können. Mit der Web Ontology Language (OWL) und dem Resource Description Framework (RDF) hat das W3C zwei aufeinander aufbauende Sprachen erarbeitet, die zur Definition von Ontologien und dazu konformer Annotation von Daten genutzt werden können. Die Erzeugung domänenspezifischer Ontologien ist allerdings ein äußerst zeit- und kostenintensives Unterfangen. So beziffert ein Report der University of Strathclyde die Kosten pro Konzept auf 40 britische Pfund[Shi03] - detaillierte Ontologien können allerdings mehrere tausend Konzepte beinhalten. Damit das Semantic Web nicht an Datenmangel scheitert,

[1] Der jüngste Bericht des Internet Software Consortiums[ISC05] beziffert die Steigerungsrate der im Netz erreichbaren Webserver gegenüber dem Vorjahr um 36 Prozent. Seit Beginn der Berichterstattung 1993 ist die Anzahl kontinuierlich gestiegen.

gibt es verschiedene Ansätze zur (semi-) automatischen Generierung von Ontologien auf der Basis bestehender Datensammlungen. Eine ausführliche Übersicht über diese Ansätze geben Ding und Foo in einem zweiteiligen Artikel im Journal on Information Science [DF02a, DF02b].

2 Motivation

Sollen bereits im WWW verfügbare digitale Archive an die Begebenheiten des Semantic Web angepasst werden, so bietet sich die Verwendung des Datenmodells des Archivs an, bei Bedarf ergänzt durch die Integration passender Konzepte bereits etablierter Ontologien. Die Annotation der Daten bzgl. der geschaffenen Ontologie lässt sich weitgehend automatisieren, denn es handelt sich um eine andere Ausgabeform bekannter Daten. Der Mehrwert gegenüber der ursprünglichen Fassung besteht in der besseren maschinellen Lesbarkeit des digitalen Archivs. Allerdings wird lediglich der Status Quo in ein neues Rahmenwerk übertragen, das Archiv selbst ändert sich nicht.
Dabei lässt sich der Wert auch für menschliche Nutzer deutlich steigern, indem die Techniken des Semantic Web zur detaillierteren Erschließung des Materials genutzt werden. Insbesondere in zur Beschreibung verwendeten Freitextfeldern stecken große Möglichkeiten: Hier finden sich oftmals weiterführende Informationen über ein bestimmtes Digitalisat, die nicht zu den formal abgelegten Stammdaten gehören, aber wertvoll für die Einordnung und das Verstehen der Inhalte des Archivs sind. So werden z.B. Orte oder Personen erwähnt, bzw. Zuordnungen zu Ereignissen oder Epochen getätigt, jedoch nicht näher erklärt. Aufgrund der Art der Ablage können sie auch nicht gezielt zugegriffen werden. Dass dies so gehandhabt wird, liegt oftmals an dem Selbstverständnis der Archivare, aber insbesondere auch an den Aufwänden, die für die semantische Aufarbeitung der Beschreibungen benötigt werden würde. Wünschenswert ist demnach ein System, das die Vorteile der Verwendung des Semantic Web für ein digitales Archiv auch auf nicht strukturierte Daten ausweitet, und dabei Mittel zur Verfügung stellt, dies in gewissem Rahmen zu automatisieren. Ein Ansatz für solch ein System wird nachfolgend skizziert.

3 Ansatz

Ausgegangen wird von einem digitalen Archiv, dessen Datenbasis bekannt ist und sich in vollständigem Zugriff befindet. Nur so lässt sich die weitgehend automatische Übertragung der enthaltenen Daten in Sprachen des Semantic Web gewährleisten. Darüber hinaus sollte es beschreibende Texte geben, zumindest für einen Teil des Bestands.
Die gewünschte Plattform muss in der Lage sein, Inhalte des digitalen Archivs zu analysieren und Vorschläge für die semantische Auszeichnung zu generieren. Ein menschlicher Supervisor bewertet die Vorschläge und ermöglicht so Feedback für evtl. notwendige Iterationen des Prozesses. Der Prozess sollte schrittweise durchgeführt werden können, damit auf Änderungen im Archiv reagiert werden kann. Je nach betreibender Institution gibt es

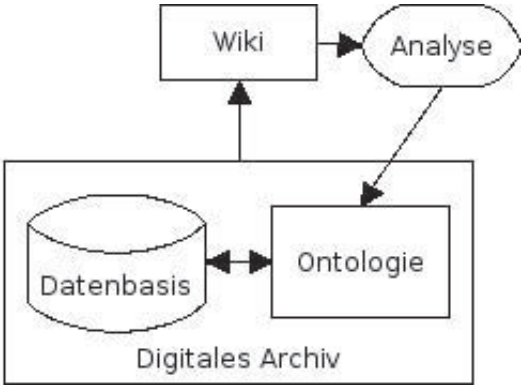

Abbildung 1: Grundaufbau des Systems

evtl. unterschiedliche Personen, die als Supervisor oder Erzeuger beschreibender Texte dienen können. Diese sollten, um die Beteiligungshürde niedrig zu halten, möglichst einfachen Zugang zu der Plattform haben, da sie im Regelfall andere Aufgaben haben.

3.1 Systemaufbau

Diese und weitere Überlegungen führen zu einer Systemarchitektur wie in Abbildung 1. Als Basis dient das digitale Archiv, bereits ergänzt um die Ontologie, die ihm zugrunde liegt. Die Daten des Archivs dienen als Input in ein Wiki, mit dem sich beschreibende Texte der Inhalte des Archivs auch von Laien einfach ändern lassen. Diese Texte werden von einem Analysemodul ausgewertet, das Vorschläge für Ontologieerweiterungen generiert. Angenommene Vorschläge werden in die Ontologie des Archivs integriert, wodurch sie sowohl für Anfragen als auch für zukünftige Analysen des Materials zur Verfügung stehen.

3.1.1 Wiki-System

Wikis ermöglichen die Erstellung von Angeboten im WWW, deren Inhalte von potentiell jedem Nutzer verändert und ergänzt werden können. Sie werden traditionell im Umfeld der Open-Source-Programmierung zum Sammeln von Informationen über Softwareprojekte eingesetzt. Das scheint intuitiv nur für einen klar definierten, möglichst kleinen Nutzerkreis zu funktionieren, das Beispiel der Wikipedia[2] zeigt jedoch, dass die Idee der Wikis auch im freien WWW funktionieren kann. Für den hier intendierten Anwendungszweck ist wesentlich, dass Wikis eine äußerst einfache Änderung enthaltener Inhalte erlauben, so dass keine längere Schulung der Anwender erforderlich ist. Ferner führen praktisch alle Wikis eine Änderungshistorie mit, mit der sich versehentliche Löschungen leicht kor-

[2]Siehe `www.wikipedia.org`

rigieren lassen. Ist eine Öffnung des Wikis nach außen nicht gewünscht, so können die Änderungsrechte eingeschränkt, bzw. kann das Wiki direkt im Intranet installiert werden. Ein weiterer Vorteil der Verwendung eines Wikis in diesem Szenario liegt darin, dass Autoren händisch Verweise auf andere Seiten des Wiki einpflegen und bei Bedarf sogar Seiten für neue Themen auf einfachste Art und Weise neu erzeugen können. Durch diese Vereinfachung der manuellen Pflege gewinnt das Analysemodul wertvollen Input für die Ontologieverfeinerung.

3.1.2 Analysemodul

Das Analysemodul hat mehrere Aufgaben. In den im System enthaltenen Texten wird zwecks besserer Vernetzung sowohl nach dem Vorkommen von Konzepten gesucht, die bereits modelliert sind, als auch nach Kandidaten für neue Konzepte. Diese können zum Beispiel mittels Named Entity Recognition aufgespürt und anschließend zur Entscheidung an den menschlichen Supervisor gemeldet werden. Eine Negativliste enthält dabei bereits abgelehnte Konzeptkandidaten. Textanalyse auf Satzbasis dient zur Generierung benannter Assoziationen zwischen Konzepten. Bereits bekannte Assoziationstypen werden dabei als Schablone verwendet. Das Feedback der Supervisoren dient als Input der iterativen Erzeugung besserer Hypothesen. Im Rahmen der Analyse werden sich voraussichtlich Anknüpfungspunkte zu externen Angeboten ergeben, die in den Analyseprozess einbezogen werden können. So bietet zum Beispiel das Umweltbundesamt unter der Adresse `www.semantic-network.de` eine Sammlung von Web Services an, die neben umweltrelevanten Daten auch Informationen zu deutschen geographischen Bezeichnern (z.B. Orts-, Gemeindenamen) als Topic Map im XTM-Format[ISO03] zur Verfügung stellen. Diese lassen sich gut in eine semantische Repräsentation der Daten integrieren.

3.2 Einsatzgebiete

Grundsätzlich ist das beschriebene System für jedes digitale Archiv mit Fließtextdaten einsetzbar. Jedoch erscheint momentan der Einsatz bei geschichtlichen oder kunsthistorischen Archiven am geeignetsten. Solche Archive enthalten üblicherweise reichhaltige Beschreibungen, die sich für eine semantische Auswertung gut verwenden lassen. Auch die Erweiterung der Texte aus einer interessierten Nutzergemeinschaft heraus ist in diesen Sparten eher zu erwarten. Bis ein solches System allerdings eine gewisse Reife und Größe erreicht hat, ist von einer völlig unreglementierten Beteiligung der Öffentlichkeit eher abzuraten. Sobald eine Gemeinschaft die Plattform angenommen hat, können mutwillige erzeugte Störungen Außenstehender so schnell entfernt werden, dass diese keinen negativen Einfluss auf den gemeinsamen Inhalt haben.

4 Thematisch verwandte Arbeiten

Zum Thema des Einsatzes von Techniken des Semantic Web in Wikis gibt es nur wenige Projekte. Am bekanntesten sind Platypus Wiki [TC04], Peri Peri (`http://www.srcf.ucam.org/~cjp39/Peri/`) und Rhizome (`http://rhizome.liminalzone.org`). Diese Systeme verwenden RDF zur Organisation des Wikis, d.h. alle Eingaben sind in RDF abgelegt, Verweise als URI realisiert. So können grundlegende Daten über die Seiten in strukturierter Weise abgefragt werden. Eine Analyse der Eingaben findet nicht statt, Vorschläge für neue Konzepte oder Verknüpfungen werden nicht generiert. Im Gegensatz zu den anderen Systemen, die Formularfelder anbieten, verwendet Rhizome eine eigene Sprache zur Erfassung der Metadaten, die angeblich verlustfrei nach RDF konvertierbar ist und eine detaillierte und an die Domäne angepasste Eingabe der Daten ermöglichen soll. Dadurch dürfte sich allerdings die Einstiegshürde für neue Autoren drastisch erhöhen.

5 Zusammenfassung

In diesem Beitrag wurde ein System skizziert, das zur Verfeinerung einer weitestgehend maschinell erzeugten Ontologie eines digitalen Archivs eingesetzt werden kann. Ein semantisch aufgewertetes Wiki vereinfacht die kollaborative Wissensgenerierung aufgrund der niedrigen Einstiegshürde ungemein, zugleich erlaubt es räumlich getrenntes Arbeiten, da es webgestützt arbeitet. Das Zusammenspiel eines bestehenden Archivs mit Beiträgen aus der Community ermöglicht eine stärkere inhaltliche Vernetzung des Angebots bei gleichzeitiger Verfeinerung der Ontologie.

Literatur

[BHL01] T. Berners-Lee, J. Hendler und O. Lassila. The Semantic Web. *Scientific American*, 2001.

[DF02a] Ying Ding und Schubert Foo. Ontology research and development. Part 1 - a review of ontology generation. *Journal of Information Science*, 28(2):123–137, 2002.

[DF02b] Ying Ding und Schubert Foo. Ontology research and development. Part 2 - a review of ontology mapping and evolving. *Journal of Information Science*, 28(5):123–137, 2002.

[ISC05] ISC. Internet Domain Survey 2005. Survey, Internet Software Consortium, 2005. Siehe `http://www.isc.org/ds/WWW-200207/index.html`.

[ISO03] ISO 13250:2003 - SGML Applications - Topic Maps. International standard, International Organization for Standardization, 2003.

[Shi03] Ali Shiri. Schemas and Ontologies, Building a Semantic Infrastructure for the Grid and Digital Libraries. *Library Hi Tech News*, 20(7), 2003.

[TC04] Roberto Tazzoli und Paolo Castagna Stefano Emilio Campanini. Towards a semantic Wiki Wiki Web, Poster Track at The 3rd International Semantic Web Conference (ISWC 2004), 2004.