Datenqualität durch inhaltsbezogene Referenzierung

Franz Weitl, Burkhard Freitag

{weitl,freitag}@fmi.uni-passau.de

Abstract: Ein Aspekt der Datenqualität bei der Integration verschiedener Dokumentfragmente zu einem neuen Dokument ist die Korrektheit der inhaltlichen Bezüge zwischen den Fragmenten. Wir zeigen, wie mit Hilfe von ontologisch repräsentiertem Wissen über Struktur und Inhalt von Dokumenten inhaltliche Bezüge so spezifiziert werden können, dass ihre Korrektheit in dynamisch zusammengestellten Dokumenten automatisch geprüft werden kann. Im Unterschied zu existierenden Ansätzen wird eine beliebig skalierbare Präzision des Referenzierungsmechanismus bei gleichzeitiger Abstraktion von Implementierungsaspekten erzielt.

1 Einleitung und Problemstellung

Ein großer Teil der Informationen im Web liegt in Form von Dokumenten vor. Unter *Web-Dokument* verstehen wir eine inhaltlich zusammengehörige Sammlung von Web-Seiten, die Information oder Wissen zu einem thematisch eingrenzbaren Bereich für bestimmte Zielgruppen strukturiert und zugänglich macht. Dokumente unterscheiden sich dadurch von anderen Daten, dass die in ihnen enthaltene Information *kohärent* [Se03] und zum großen Teil *implizit*, d.h. einer maschinellen Verarbeitung nicht direkt zugänglich, ist.

Die Erstellung von Web-Dokumenten ist oft aufwendig und kostenintensiv. Deswegen ist beispielsweise im eLearning die Wiederverwendung und automatisierte, bedarfsgetriebene Zusammenstellung von Dokumenten eine zentrale Anforderung [HC01, HN00, Da01, Se03]. Bei der Fusion von Dokumentfragmenten aus verschiedenen Quellen muss die *Dokumentkohärenz*, eine Form der *Datenqualität*, sichergestellt werden [Se03].

Ein Aspekt der Dokumentkohärenz ist die Existenz *inhaltlicher Bezüge* wie Ausblick, Zusammenfassung, Rückblick, Wiederholung, Querverweis, Problembeschreibung und Motivation. Inhaltliche Bezüge sind in Fragmenten enthalten, deren inhaltliche Korrektheit oder Verständlichkeit davon abhängt, dass innerhalb des Web-Dokuments andere Fragmente mit bestimmten inhaltlich-strukturellen Eigenschaften vorhanden sind (vgl. [Se03]).

Im Folgenden wird ein Ansatz zur automatischen Überprüfung der *Korrektheit* der inhaltlichen Bezüge in Dokumenten, welche Fremdressourcen einbinden, und zur automatischen Generierung *inhaltsbasierter* Navigationsstrukturen vorgestellt. Die Existenz globaler Ontologien im Web [NP01] erlaubt die Übertragung des hier vorgestellten Ansatzes auf andere Teilprobleme der Informationsfusion.

Zunächst wird der Stand der Forschung in der Spezifikation von inhaltlichen Bezügen kurz

skizziert, dann der eigene Lösungsansatz motiviert und vorgestellt. Der Artikel schließt mit einer Zusammenfassung und einem Ausblick auf weitere Forschungsfragestellungen.

2 Spezifikation von Bezügen in Dokumenten: Stand der Forschung

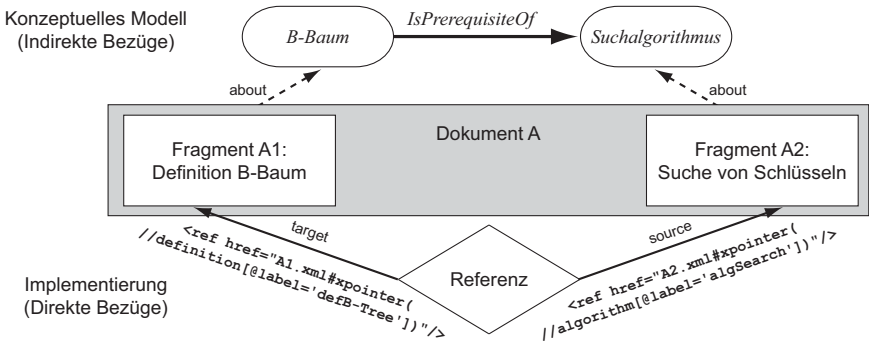

Abbildung 1: Trennung von Inhalt und Vernetzung, Spezifikation von Verknüpfungsstrukturen entweder auf der Ebene der Implementierung oder auf einem konzeptuellen Modell der Diskursdomäne

Als Hauptursache für fehlerhafte technische Bezüge („dangling links“) in Web-Dokumenten wurde die Vermischung von Inhalt und Vernetzungsstrukturen (z.B. in HTML) ausgemacht. Hypermedia-Referenzmodelle [HS90, HBvRG94] betonen daher die Separierung von Informationsstrukturen und Verlinkungsstrukturen. Links werden als „first class entities“ modelliert, die separat von den Inhalten spezifiziert, vorgehalten und gepflegt werden. Dabei kann man zwei verschiedene Konzepte für die Spezifikation von Verknüpfungsstrukturen in Dokumenten unterscheiden (siehe Abb. 1). Bezüge können zum einen direkt auf der Ebene der Implementierung eines hypermedialen Dokuments spezifiziert werden. Im Falle von XML-Dokumenten bietet sich hierfür etwa der XLink-Standard des W3C an [XLi01]. Zum anderen können Vernetzungsstrukturen indirekt auf der Basis eines abstrakten konzeptuellen Modells der Diskursdomäne spezifiziert werden [HN00, Se03]. Die Inhalte selbst werden über die behandelten Themen in Beziehung mit dem konzeptuellen Modell gesetzt. Zur Spezifikation von Themen-basierten Bezügen bieten sich Topic-Maps [PM01] oder auch der W3C-Standard RDF [W3C99] an.

Unabhängig von der technischen Separierung besteht in der Regel eine enge Korrespondenz zwischen dem Inhalt und der Hypertextstruktur eines Web-Dokuments. Die Themenbasierte Vernetzung von Inhalten reicht nicht immer aus, diese Korrespondenz abzubilden, da neben der Struktur der Diskursdomäne andere Aspekte für Beziehungen in Dokumenten eine Rolle spielen können (z.B. didaktische Überlegungen im Falle von Lerndokumenten). Die direkte Spezifizierung von Verknüpfungsstrukturen auf der Implementierungsebene andererseits ermöglicht nicht, dass *inhaltlich* adäquate Verknüpfungspartner in neuen Verwendungskontexten automatisch gefunden werden können. Unser Ansatz besteht darin, Beziehungen als eine Eigenschaft des konkreten Inhalts eines Dokuments (nicht nur als

Eigenschaft der Diskursdomäne) zu modellieren. Bezugsziele werden über Anfragen an „semantische“ Eigenschaften eines Dokuments spezifiziert, ohne auf Aspekte der konkreten Implementierung (z.B. XML-Struktur) Bezug zu nehmen.

3 Lösungsansatz

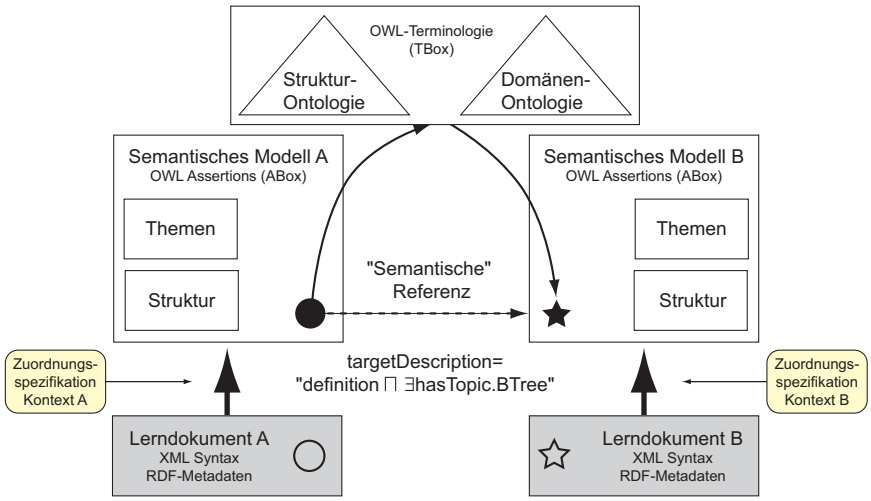

Abbildung 2: Repräsentation von Wissen über Lerndokumente in OWL und Spezifikation von Referenzzielen über beschreibungslogische Aussagen

Ausgangspunkt für unseren Ansatz sind feingranulare Dokumentenformate auf Basis von XML. Weiterhin kann externe Information über das Dokument in Form eines RDF-Metadatensatzes vorliegen, welcher entweder manuell oder mit Hilfe eines Textanalysewerkzeugs generiert wurde [HSC02]. Eine Bedeutung im Sinne von eindeutiger, einheitlich interpretierbarer Semantik erhält die in der XML-Auszeichnung oder in den RDF-Metadaten enthaltenen Information erst dadurch, dass man sie in Beziehung zu einer (oder mehreren) Ontologien setzt, welche als formale, über Systemgrenzen hinweg gültige Repräsentation einer Wissensdomäne dienen [Gr98].

Die in der XML-Auszeichnung und externen Metadaten enthaltene Information wird mit Hilfe einer *Zuordnungsspezifikation* in ein *semantisches Modell* abgebildet, welches die für die Referenzierung von Fragmenten benötigte Information über *Struktur* und *Themen* in Form von *OWL-Zusicherungen* [vHHH$^+$] enthält. Diese Zusicherungen beziehen sich auf zwei globale terminologische Ontologien. Die *Strukturontologie* definiert, in welcher Weise Informationen in Dokumenten gegliedert werden können (Module, Kapitel, Abschnitte, Lektionen) und welche Arten der Informationsvermittlung und -darstellung man unterscheiden kann (Definitionen, Beispiele, Tabellen, Abbildungen). Die *Domänenontologie* modelliert die Diskursdomäne eines Dokuments. Bezüge können durch *Kon-*

zeptbeschreibungen repräsentiert werden. Über den Inferenzdienst *Instance Retrieval*, den beschreibungslogische Systeme (DL-Systeme) wie Racer [HM01] typischerweise anbieten, können automatisch alle Instanzen einer Konzeptbeschreibung in der Wissensbasis (bestehend aus der Strukturontologie, Domänenontologie und dem semantischen Modell) gefunden werden und damit die Referenz aufgelöst werden.

Beispiel In dem semantischen Modell zu Lerndokument B seien folgende beschreibungslogische Zusicherungen enthalten:

$$definition(defBBaum) \tag{1}$$
$$hasTopic(defBBaum, BayerBaum) \tag{2}$$

„Das Fragment $defBBaum$ ist eine Definition und hat $BayerBaum$ als zentrales Thema." In der *Zuordnungsspezifikation* von Kontext B sei definiert, dass $BayerBaum$ ein $BTree$ Thema ist:

$$BTree(BayerBaum) \tag{3}$$

Im Dokument A wird auf die Definition von „BayerBaum" Bezug genommen. Dieser Bezug kann durch folgende Konzeptbeschreibung repräsentiert werden:

$$definition \sqcap \exists hasTopic.BTree \tag{4}$$

Ein Dokument C könnte einen Querverweis der Form „mehr zu Bäumen" enthalten:

$$\exists isAbout.Tree \tag{5}$$

Unter der Voraussetzung, dass sich aus der Strukturontologie ableiten lässt, dass $hasTopic$ eine Unterrolle von $isAbout$ ist ($hasTopic \sqsubseteq isAbout$), und dass sich aus der Domänenontologie ableiten lässt, dass $BTree$ ein Unterkonzept von $Tree$ ist ($BTree \sqsubseteq Tree$), ist die Definition $defBBaum$ sowohl eine Instanz der Konzeptbeschreibung (4) als auch (5).

4 Zusammenfassung

Eine besondere Herausforderung bei der Fusion von Dokumenten sind inhaltliche Bezüge, die in neuen Kontexten überprüft und angepasst werden müssen. Ziel ist die Automatisierung dieser Integrationsaufgaben. Es wurde motiviert, warum bestehende Techniken wie XLink oder TopicMaps zur Spezifikation von inhaltlichen Beziehungen in Dokumenten nicht ausreichen. Der vorgestellte Lösungsansatz beruht auf der Repräsentation des notwendigen expliziten Wissens über Dokumente in einem formalen semantischen Modell, das auf Ontologien für Dokumentstrukturen und Diskursdomänen basiert. Dadurch werden Bezüge in Dokumenten in einer implementierungsunabhängigen, systemübergreifend einheitlich interpretierbaren Weise repräsentiert und verschiedene Abstraktionsgrade bei der Spezifikation von Referenzzielen ermöglicht.

Die nächsten Arbeitsschritte konzentrieren sich darauf, Anforderungen an die Korrektheit von inhaltlichen Bezügen präzise zu charakterisieren und einen effizienten Zuordnungsmechanismus von lokalen Metadaten zu einem Ontologie-basierten semantischen Modell zu definieren. Für die Informationsfusion von allgemeinem Interesse ist die Frage, ob die auf DL-Reasoning basierenden Retrievaltechniken skalieren. Dazu soll untersucht werden, ob bzw. inwieweit Fragmente der Beschreibungslogik auf eine relationale Sprache wie DATALOG abgebildet und mit relationalen Mitteln ausgewertet werden können.

Literatur

[Da01] Dahn, I.: Automatic textbook construction and web delivery in the 21st century. *Journal of Structural Learning and Intelligent Systems*. 14(4):401–413. 2001.

[Gr98] Gruber, T. R.: A translation approach to portable ontology specifications. *Knowledge Acquisition*. 5(2):21–66. 1998.

[HBvRG94] Hardman, L., Bulterman, C., und van Rossum G., G.: The Amsterdam Hypermedia Model. *Communications of the ACM*. 37(2):50–62. 1994.

[HC01] Hollfelder, S. und Caumanns, J.: IT-Weiterbildung nach Maß - das Projekt "Teachware on Demand". In: Deiters, W. und Lienemann, C. (Hrsg.), *Report Informationslogistik - Informationen just-in-time*. S. 101–110. Symposion Publishing. Düsseldorf. 2001.

[HM01] Haarslev, V. und Möller, R.: RACER system description. In: *Proceedings of the International Joint Conference on Automated Reasoning (IJCAR-01), Lecture Notes In Artificial Intelligence*. volume 2083. Springer-Verlag. 2001.

[HN00] Henze, N. und Nejdl, W.: Extendible adaptive hypermedia courseware: Integrating different courses and web material. In: *Proceedings of the Intern. Conf. on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2000)*. Trento, Italy. May 2000.

[HS90] Halasz, F. und Schwartz, M. The dexter hypertext reference model. 1990. NIST Hypertext Standardization Workshop.

[HSC02] Handschuh, S., Staab, S., und Ciravegna, F.: S-CREAM - Semi-automatic CREAtion of Metadata. In: *Proceedings of the European Conference on Knowledge Acquisition and Management - EKAW-2002, Lecture Notes in Computer Science*. Madrid, Spain. October 2002. Springer-Verlag.

[NP01] Niles, I. und Pease, A.: Towards a standard upper ontology. In: Welty, C. und Smith, B. (Hrsg.), *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*. Ogunquit, Maine. 2001.

[PM01] Pepper, S. und Moore, G. XML Topic Maps (XTM) 1.0, TopicMaps.Org Specification. http://www.topicmaps.org/xtm/1.0/. 2001. zuletzt besucht Okt. 2003.

[Se03] Seeberg, C.: *Life long Learning - Modulare Wissensbasen für elektronische Lernumgebungen*. Springer-Verlag. Berlin, Heidelberg, New York. 2003.

[vHHH$^+$] van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., und Stein, L. A. OWL Web Ontology Language Reference, W3C Candidate Recommendation 18 August 2003. http://www.w3.org/TR/owl-ref/. zul. bes. Okt. 2003.

[W3C99] Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999. http://www.w3.org/TR/REC-rdf-syntax/. 1999. zuletzt besucht Okt. 2003.

[XLi01] XML Linking Language (XLink) Version 1.0 W3C Recommendation 27 June 2001. http://www.w3.org/TR/xlink/. 2001. zuletzt besucht Apr. 2004.