

Vasari: Wissensextraktion mittels iterativ entwickelter Dokumentbeschreibungen

Norbert Gövert Norbert Fuhr

<http://ls6-www.cs.uni-dortmund.de/vasari/>

1 Einleitung

In vielen Bereichen liegt der größte Teil des verfügbaren Wissens nur in textueller Form vor. Während z. B. in Unternehmen in der Vergangenheit große Anstrengungen unternommen wurden, um das verfügbare Faktenwissen in umfassenden Datenbanken (*Data Warehouse*) zu sammeln, geht man heute davon aus, dass das in Texten enthaltene Wissen das Faktenwissen bezüglich des Umfangs um Größenordnungen übertrifft.

Wir stellen ein System vor, welches solches Wissen aus unterschiedlich stark strukturierten Texten extrahiert. Grundlage für die Wissensextraktion ist die Beschreibung der Dokumente in einer eigens dazu entwickelten Sprache. Diese basiert im Wesentlichen auf Delimiterregeln und linguistischen Regeln. Für die Erstellung der Dokumentenbeschreibungen steht ein Werkzeug zur Verfügung, welches interaktiv die iterative Verbesserung einer Beschreibung erlaubt und gleichzeitig die Auswirkungen von Änderungen visualisiert.

Als Anwendungskontext für die hier beschriebenen Methoden stellen wir im folgenden Abschnitt das *Vasari*-Projekt vor, welches die Erstellung eines Kunst-Informationssystems zum Ziel hat. In Abschnitt 3 skizzieren wir den Prozess der Wissensextraktion, Abschnitt 4 beschreibt die iterative Erstellung von Dokumentbeschreibungen.

2 Das Giorgio-Vasari-Projekt

Ziel des vom Bochumer Kunstverein *artregister.org* initiierten Giorgio-Vasari¹-Projektes ist es, sowohl einer interessierten Öffentlichkeit, als auch dem Fachpublikum vielfältige Informationen über Kunst und Künstler auf einfache Weise zu erschließen. Die Errichtung einer allgemein zugänglichen Datenbank im World Wide Web soll das Fundament dafür schaffen.

Als Datengrundlage für den Aufbau eines solchen Web-Informationssystems dienen zahlreiche Kunstlexika. Die Einträge in diesen Lexika orientieren sich meist an den Künstlern selbst; die unterschiedlichen Lexika bilden jedoch stets auch unterschiedliche Schwerpunkte; während beispielsweise [Thieme & Becker 92] sich mit den Biographien der Künstler beschäftigen (im genannten Nachschlagewerk sind davon ca. 250 000 erfasst), konzentrieren sich andere Lexika z. B. auf die Werksverzeichnisse der Künstler. Wieder andere Nachschlagewerke widmen sich der bloßen Identifikation von Künstlern (z. B. [Gorenflo 88]) oder enthalten Signaturen, Monogramme und Symbole von Künstlern.

Als Basis für dieses Projekt stellt uns *artregister.org* die eingescannten Texte aus derzeit etwa 20 solcher Nachschlagewerke² zur Verfügung, die mit einer OCR-Software aufbereitet wurden.

¹Vasari, Giorgio, * Arezzo 30. Juli 1511, † Florenz 27. Juni 1574, ital. Maler, Architekt und Kunstschriftsteller. Begründer der Kunstgeschichte. Seine Künstlerbiographien („Die Lebensbeschreibungen der berühmtesten italienischen Architekten, Maler und Bildhauer“, 1546–50, erw. Ausgabe 1568) gehören zu den wichtigsten Grundlagen der Kunstgeschichte.

²Aufgrund des Alters der meisten Texte unterliegen diese keinerlei Copyright-Beschränkungen.

3 VaLa: Beschreibungssprache für schwach strukturierte Dokumente

Grundlage für die Wissensextraktion ist die semantische Anreicherung der Dokumente durch geeignete Annotation. Als Syntax für die Annotierung findet XML Anwendung. Damit die Annotation automatisch erfolgen kann, werden die Quelldokumente zunächst in der eigens dazu entwickelten *Vasari Language* (VaLa) abstrakt beschrieben. VaLa erlaubt die Kombination von informatischen und computerlinguistischen Methoden für eine effektivere Extraktion von Wissen: Neben den Delimiterregeln, wie sie z. B. in den Beschreibungssprachen von Jedi [Fankhauser & Xu 93] und MarkItUp! [Huck et al. 98] verwendet werden, kann auch Computerlinguistisches Wissen (z. B. Wissen über *Part-of-Speech*-Klassen und Phrasen) in den Extraktionsprozess einbezogen werden.

VaLa verwendet *XML Schema* [Fallside 01], dessen Ausdruckskraft in Bezug auf die Definition von Datentypen und Dokument-Strukturen ausgenutzt wird. VaLa-Beschreibungen geben zunächst einmal die Struktur der Dokumente wieder, die aus dem Wissensextraktionsprozess resultieren sollen. Um nun die Abbildung zwischen den Quelldokumenten und den zu füllenden Knoten dieser XML-Struktur herzustellen, wird das XML-Schema durch *Kontextausdrücke* ergänzt. Diese identifizieren durch Matching auf den Quelldokumenten die Teile, die an die entsprechende Stelle der XML-Struktur einzufügen sind. Die Kontextausdrücke können sich folgendes Wissen zu Nutze machen:

- Zeichenketten innerhalb der Quelldokumente können durch reguläre Ausdrücke extrahiert werden. Der Aufbau komplexer Ausdrücke wird durch die Möglichkeit, einfachere Ausdrücke ineinander zu schachteln, erleichtert.
- Linguistische Konstrukte (Wörter, Wortklassen, Entitäten) werden mittels computerlinguistischer Kategorien gematcht. Zur Extraktion dieser Konstrukte aus den Quelldokumenten setzen wir SPPC ein [Neumann & Piskorski 02].
- Weitere Bedingungen an einen zu extrahierenden Teil eines Quelldokumentes können durch Spezifikation entsprechender XML-Schema-Datentypen sowie durch Angabe von Ausdrücken für Rechts-/Linkskontexte.

Kennzeichnend für die Vasari-Anwendung ist, dass wir Quelldokumente aus unterschiedlichen Kunstlexika vorliegen haben, deren Inhalte sich teilweise überschneiden. Dies lässt sich für eine verbesserte Erkennung von Entitäten (z. B. Künstler- oder Ortsnamen) ausnutzen: Die aus den Artikeln eines Lexikons extrahierten Entitäten werden in einem Wörterbuch gesammelt, welches dann zur Erkennung von Entitäten in anderen Lexika verwendet wird.

4 Interaktive, iterative Erstellung von VaLa-Beschreibungen

Die Erstellung von abstrakten Beschreibungen schwach strukturierter Dokumente ist ein schwieriges Unterfangen: Auf der einen Seite ist es bei großen Datenmengen nicht möglich, jedes Dokument einzeln zu berücksichtigen. Auf der anderen Seite möchte man natürlich möglichst viele Dokumente mit der Beschreibung abdecken. Um eine bessere Kontrolle über die Effektivität einer VaLa-Beschreibung zu erreichen, entwickeln wir derzeit ein Werkzeug zur interaktiven, iterativen Erstellung solcher Beschreibungen.

Ähnlich wie bei MarkItUp! werden VaLa-Beschreibungen anhand von Beispielen aus der Menge von Quelldokumenten erstellt. Der Benutzer definiert zunächst die für die Zieldokumente angestrebte Struktur. Die Kontextausdrücke können anhand der Beispieldokumente erstellt werden. Zur Kontrolle der Ausdrücke wird eine so erstellte VaLa-Beschreibung direkt auf beliebige weitere Beispieldokumente angewendet; das Ergebnis des Extraktionsprozesses wird zusammen mit der Beschreibung und dem Ausgangsdokument angezeigt.

Eine so gewonnene Beschreibung kann nun iterativ verbessert werden. Zu jedem Iterationsschritt gibt der Benutzer Feedback zum bislang erreichten Ergebnis ab: Das durch den Extraktionsprozess gewonnene Markup kann als korrekt oder falsch bewertet werden; zusätzlich gewünschtes Markup kann direkt in das Ergebnis eingegeben werden. Nach dem Feedback wird die Beschreibung entsprechend des Feedbacks weiter entwickelt. Immer dann, wenn eine neue Version der Beschreibung erstellt wurde, erfolgt eine Anwendung auf die gewünschten Beispieldokumente. Anhand der zum Zeitpunkt der Anwendung bereits vorliegenden

Feedback-Daten kann das System bereits einen Teil der Bewertung des Ergebnisses übernehmen und diese für den Benutzer visualisieren.

Für die iterativ erstellten Beschreibungen steht eine Versionsverwaltung zur Verfügung, anhand derer die Ergebnisse für jede Version zu jedem Zeitpunkt rekonstruiert werden können.

5 Ausblick

Nachdem mittels der hier vorgestellten Methoden Wissen aus unterschiedlichen Dokumentenquellen gewonnen wurde, kann es in einem darauffolgenden Synthese-Schritt zusammengeführt werden. So können die in den unstrukturierten Quelldokumenten implizit vorhandenen Zusammenhänge explizit hergestellt und in Form mächtiger Such- und Navigationsstrukturen verfügbar gemacht werden. Für die Nutzer ergeben sich somit Wege, das für ihn relevante Wissen effektiver zu erschließen. Für die Erstellung der einheitlichen Wissensrepräsentation soll das *Resource Description Format* [Lassila & Swick 99] zum Einsatz kommen.

Literatur

Fallside, D. (2001). *XML Schema Part 0: Primer*. <http://www.w3.org/TR/xmlschema-0/>.

Fankhauser, P.; Xu, Y. (1993). MarkItUp! An incremental approach to document structure recognition. In: *Proceedings of the 5th International Conference on Electronic Publishing, Document Manipulation, and Typography (EP '94)*, volume 6, pages 447–456. <ftp://ftp.darmstadt.gmd.de/pub/oasys/reports/P-94-07.ps.Z>.

Gorenflo, R. M. (ed.) (1988). *Verzeichnis der bildenden Künstler von 1880 bis heute*. Brün-Verlag, Rüsselsheim. 3 Bände.

Huck, G.; Fankhauser, P.; Aberer, K.; Neuhold, E. J. (1998). Jedi: Extracting and Synthesizing Information from the Web. In: *Proceedings of the 3rd IFICIS International Conference on Cooperative Information Systems, New York, August 20-22, 1998*, pages 32–43. IEEE-CS.

Lassila, O.; Swick, R. (1999). *Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation*. Technical report, World Wide Web Consortium. <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.

Neumann, G.; Piskorski, J. (2002). A Shallow Text Processing Core Engine. *Journal of Computational Intelligence* 18. <http://www.dfki.de/~neumann/publications/new-ps/comp-intell.pdf>.

Thieme, U.; Becker, F. (eds.) (1992). *Allgemeines Lexikon der bildenden Künstler von der Antike bis zur Gegenwart*. Deutscher Taschenbuch Verlag, München. 19 Bände; Nachdruck; Original von 1909/1910.