Visuelle Analyse von RDF-Daten mittels semantischer Linsen

Steffen Lohmann[1], Philipp Heim[2], Davaadorj Tsendragchaa[2], Thomas Ertl[2]

DEI Laboratory, Universidad Carlos III de Madrid [1]
Institut für Visualisierung und Interaktive Systeme, Universität Stuttgart [2]

Zusammenfassung

Das Semantic Web hat zur Entstehung einer Vielzahl von RDF-Datensätzen beigetragen, die in strukturierter Form umfangreiche Informationen bereitstellen und damit ein enormes Nutzungspotenzial bergen. Um dieses Potenzial voll erschließen zu können, bedarf es Methoden und Werkzeuge, die auch vom Laien einfach zu verstehen und zu bedienen sind. Mit diesem Ziel haben wir einen Ansatz entwickelt, der die Analyse von RDF-Daten mit Hilfe von Streudiagrammen und semantischen Linsen visuell unterstützt. Die Linsen können generisch für beliebige Objekteigenschaften erstellt und über logische Operatoren miteinander kombiniert werden. Eine Umsetzung für den DBpedia-Datensatz illustriert die einfache Anwendbarkeit und leichte Verständlichkeit dieses Ansatzes.

1 Einleitung

Das Semantic Web hat in den letzten Jahren eine Vielzahl von Datensätzen hervorgebracht, die Informationen in standardisierter Form und für den strukturierten Zugriff bereitstellen. Eines der bekanntesten Projekte ist *DBpedia*, bei dem Informationen aus *Wikipedia*-Artikeln extrahiert, semantisch eindeutig repräsentiert und mit weiteren Semantic-Web-Daten verlinkt werden (Auer et al. 2007). Die so aufbereiteten und angereicherten Informationen bergen ein enormes Nutzungspotenzial, da sie komplexe Anfragen und automatisierte Schlussfolgerungen erlauben, die auf den ursprünglichen Quellen (wie beispielsweise den originalen Wikipedia-Artikeln) sehr aufwändig bis unmöglich sind. Die Formulierung solcher Anfragen geschieht in der Regel mittels *SPARQL*, einer formalen Anfragesprache für Datensätze, die auf dem *Resource Description Framework (RDF)* basieren – so wie DBpedia und die meisten anderen Datensätze des Semantic Web.

Da ein Erlernen dieser formalen Sprache vom durchschnittlichen Webnutzer nicht verlangt werden kann, bedarf es alternativer Zugriffsformen, die eine gleichermaßen intuitive wie effiziente Nutzung der Daten ermöglichen. Die spezifischen Eigenschaften von RDF-Daten sollten hierbei nicht nur berücksichtigt, sondern aktiv zur Bereitstellung innovativer Funktio-

nalitäten und einer verbesserten Interaktion eingesetzt werden. Mit diesem Ziel haben wir einen anwenderorientierten Ansatz entwickelt, der die Analyse von RDF-Daten mit Hilfe von *Streudiagrammen* und *semantischen Linsen* visuell unterstützt. Die Streudiagramme geben eine *globale* Übersicht und unterstützen die Entdeckung von Mustern und Korrelationen auch in großen Datenmengen. Die semantischen Linsen ermöglichen die *lokale* Exploration und Filterung von ausgewählten Teilbereichen, wobei sie durch ihre Interaktionsmetapher einfach zu bedienen sind. Im Folgenden beschreiben wir diesen Ansatz ausführlicher und illustrieren dessen Potenziale am Beispiel des DBpedia-Datensatzes und der visuellen Analyse europäischer Staaten anhand verschiedener Faktoren.

2 Semantische Linsen

Unser Konzept der semantischen Linsen basiert auf einer Interaktionsmetapher, die in den neunziger Jahren von Bier et al. (1994) am *Xerox PARC* entwickelt und *magische Linse* genannt wurde. In ihrer ursprünglichen Definition besteht eine magische Linse aus einer Fläche beliebiger Form und Größe, die über der eigentlichen Benutzungsschnittstelle liegt und mit einem Operator versehen ist, der bestimmte Informationen filtert oder farblich kennzeichnet. Ähnlich einer optischen Linse in der realen Welt kann diese virtuelle Linse über verschiedene Stellen der Benutzungsschnittstelle geführt werden, um beispielsweise eine komplexe Visualisierung auf ausgewählte Informationsdimensionen zu reduzieren und somit fokussiert betrachten zu können.

In der Vergangenheit wurden eine Reihe von Variationen und Weiterentwicklungen des Konzepts der magischen Linse vorgeschlagen. Beispiele reichen von einer Variante zur Videoanalyse (Ryall et al. 2005) bis hin zu einer 3D-Umsetzung für Volumendaten (Viega et al. 1996). Unsere Umsetzung der magischen Linse haben wir „semantische Linse“ getauft, da sie die semantische Struktur von RDF-Daten ausnutzt, d.h. beliebige Objekteigenschaften und Datentypen interpretieren kann. Ein verwandter Ansatz sind die Linsen des *Fresnel*-Vokabulars, die eine Selektion von Eigenschaften bei der Anzeige von RDF-Daten ermöglichen (Pietriga et al. 2006). Eine zu Fresnel ähnliche Umsetzung findet sich außerdem im Semantic Web Browser *Haystack*, wo mittels Linsen bestimmte Aspekte bei der Anzeige von RDF-Daten gruppiert werden können (Quan et al. 2003). Darüber hinaus gibt es eine verwandte Arbeit von Rotard et al. (2007), in der mit Linsen markierte Textstellen innerhalb von Webseiten mit RDF-Daten semantisch angereichert werden.

Im Gegensatz zu diesen Arbeiten sind die semantischen Linsen in unserem Ansatz eine unmittelbare Umsetzung der Idee der magischen Linse für semantische Daten. Entsprechend dem Grundkonzept bestehen sie aus einer Fläche mit binärem Filter, den die durch die Linse betrachteten Objekte entweder erfüllen oder nicht. Die Objekte werden weder um Informationen angereichert, noch werden bestimmte Eigenschaften ausgeblendet oder gruppiert. Darüber hinaus erlaubt unsere Umsetzung eine generische Erstellung beliebiger Linsen durch die Nutzer, wohingegen in bisherigen Ansätzen lediglich vordefinierte Linsen verwendet werden konnten.

Eine semantische Linse kann in unserem Ansatz für prinzipiell jede Objekteigenschaft erstellt werden, die im RDF-Datensatz vorhanden ist, wobei der Datentyp der Objekteigenschaft das Filterkriterium bestimmt. Würde man beispielsweise in einer Visualisierung eines sozialen Netzwerkes alle nicht volljährigen Personen mit semantischen Linsen markieren wollen, so wären die Personen die Objekte, das Alter die interessierende Objekteigenschaft und die Volljährigkeit das Filterkriterium, für das man in Deutschland einen Schwellenwert von 18 Jahren festlegen würde. Darüber hinaus lassen sich auch mehrere semantische Linsen über Boolesche Operatoren miteinander kombinieren, so dass Filterketten realisiert werden können, wie sie für magische Linsen bei Fishkin und Stone (1995) beschrieben wurden. Anschaulicher wird der Ansatz an einem konkreten Anwendungsbeispiel, wie wir es im folgenden Abschnitt beschreiben.

3 Anwendungsbeispiel der visuellen Analyse

Um unseren Ansatz zu veranschaulichen und zu testen, haben wir die Webanwendung *SemLens* entwickelt. SemLens greift per SPARQL auf DBpedia zu und benötigt zur Ausführung lediglich einen Webbrowser mit installiertem Flash Player. Nach Auswahl einer geeigneten DBpedia-Klasse (wie z. B. „European Countries“), werden alle Objekte, die dieser Klasse angehören, als Punkte in einem Streudiagramm visualisiert (vgl. Abb. 1, A). Rechts neben dem Streudiagramm sind die Eigenschaften dieser Objekte gelistet (B). Als Diagrammachsen werden automatisch diejenigen Eigenschaften vorausgewählt, die die meisten Objekte miteinander teilen. Diese Vorauswahl kann der Nutzer anpassen, indem er den Achsen andere Eigenschaften aus der Liste zuweist. In Abb. 1 wurden beispielsweise die Objekteigenschaften „longitude“ und „HDI rank“ (HDI = Human Development Index) als Achsendimensionen gewählt (C1, C2), woraufhin sich die Streuung der Objekte im Diagramm sowie der Datentyp und Wertebereich der Achsen anpasst. Das Streudiagramm ermöglicht einen schnellen Überblick über die Verteilung der Objekte und unterstützt die Entdeckung von Mustern und Korrelationen in den Daten. Abb. 1 zeigt beispielsweise einen korrelativen Zusammenhang zwischen dem HDI-Rang eines europäischen Landes und seiner geographischen Länge. Auch wenn ein tendenziell geringerer HDI-Rang in den ehemaligen „Ostblock“-Staaten zu erwarten war, wird dieser Zusammenhang durch das Streudiagramm visuell veranschaulicht und die Identifikation von „Ausreißern“ aus diesem Zusammenhang erleichtert. In diesem Fall weichen beispielsweise die Datenpunkte von Portugal (D1) und Zypern (D2) recht deutlich von der gegebenen Korrelation ab.

Für die genauere Untersuchung dieser Ausreißer und weiterer lokaler Diagrammbereiche eignen sich nun die semantischen Linsen. Im gegebenen Anwendungsfall unterstützen sie das explorative Austesten verschiedener Erklärungsansätze. Hierfür wurden, wie in Abb. 1 dargestellt, zwei semantische Linsen definiert: Die erste markiert alle Staaten, die ein Bruttoinlandsprodukt (BIP) pro Kopf von mehr als 25.000 Euro haben (Linse L1); die zweite alle Staaten mit Euro-Währung (Linse L2). SemLens unterstützt die Erstellung von semantischen Linsen, indem es den Benutzerdialog automatisch an den Datentyp der gewählten Objekteigenschaft anpasst. Für numerische Daten (wie im Fall von L1) werden beispielweise andere Filterkriterien angeboten als für textbasierte Daten (wie in L2).

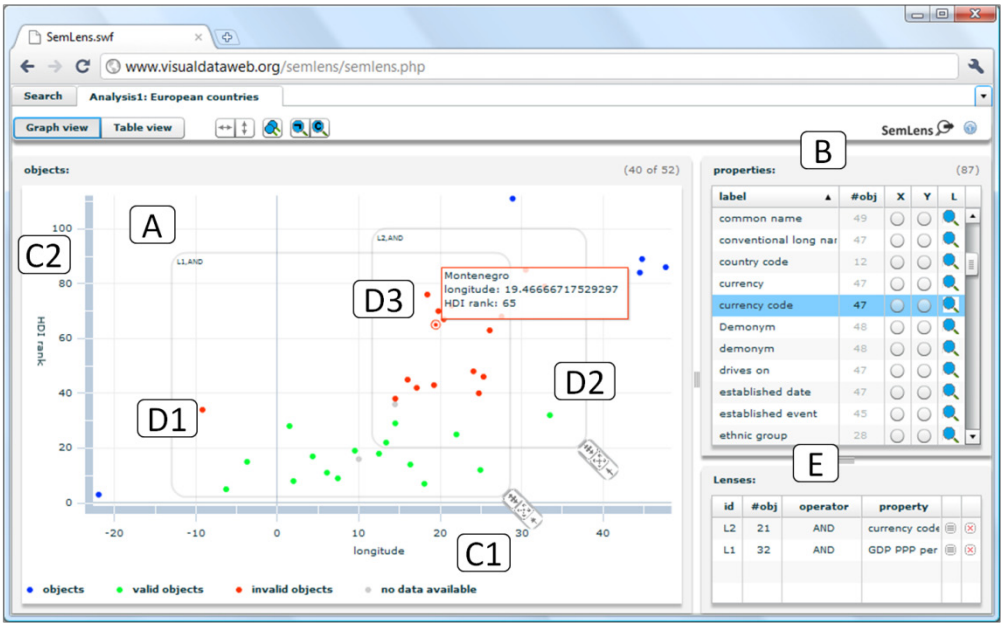

Abbildung 1: Visuelle Analyse von DBpedia-Daten mit SemLens
(Live-Demo und weitere Informationen zu SemLens unter http://semlens.visualdataweb.org)

Die Rotfärbung des Datenpunkts von Portugal nach Positionierung von L1 im Streudiagramm deutet darauf hin, dass der vergleichsweise geringe HDI-Rang mit dem ebenfalls vergleichsweise geringen BIP pro Kopf zusammenhängen könnte. Die Grünfärbung von Zypern nach Positionierung von L2 deutet andererseits auf einen positiven Zusammenhang von Euro-Währung und HDI-Rang hin. Bei der weiteren Exploration dieses Zusammenhangs durch Vergrößern des Linsenbereichs von L2 fällt Montenegro auf, das zwar den Euro als Währung hat, aber nur einen HDI-Rank von 65 aufweist. Vergrößern wir nun jedoch auch den Bereich von L1, überlagern sich die beiden Linsen beim Datenpunkt von Montenegro und führen aufgrund ihrer UND-Verknüpfung zu einer Rotfärbung des Datenpunkts (D3) – ein klares Indiz, dass Montenegro ein zu geringes BIP pro Kopf aufweist, um einen HDI-Rang vergleichbar dem anderer Euro-Staaten zu erzielen. Die Einbeziehung zusätzlicher Faktoren über semantische Linsen erlaubt somit eine multidimensionale Exploration, die in unserem Fall idealerweise zu einer sauberen Abgrenzung von roten und grünen Datenpunkten innerhalb des Streudiagramms führt.

Neben dem standardmäßig vorausgewählten UND-Operator können die Filterkriterien verschiedener semantischer Linsen auch per ODER miteinander verknüpft werden. Darüber hinaus lassen sich NOT-Linsen definieren und mehrere Linsen zu einer aggregieren, so dass die Erstellung beliebiger Boolescher Ausdrücke möglich ist (vgl. Fishkin und Stone 1995). Da die Reihenfolge, in der die Linsen über der Visualisierung angeordnet sind, für die Boolesche Verknüpfung entscheidend ist, kann diese jederzeit im Streudiagramm angezeigt und durch Verschieben von Linsen im sogenannten *Lens Stack* (E) angepasst werden.

4 Fazit

Aufgrund der generischen Struktur von RDF lassen sich semantische Linsen für beliebige Datensätze und Objekteigenschaften definieren. Die semantisch eindeutigen Beschreibungsmöglichkeiten von RDF ermöglichen darüber hinaus eine automatische Anpassung des Dialogs zur Linsenerstellung an den Datentyp der jeweiligen Objekteigenschaft. Da jede Linse ein binäres Filterkriterium samt Booleschen Operator definiert, können durch Übereinanderlegen der Linsen nahezu jegliche semantische Anfragen interaktiv und visuell erstellt werden.

Wie das angeführte Beispiel der Analyse von DBpedia-Daten zeigt, ermöglicht die intuitive Umsetzung der Linsenmetapher in SemLens prinzipiell auch Laien, komplexe Analysen auf RDF-Daten durchzuführen – was vor dem Hintergrund einer steigenden Anzahl an interessanten Datensätzen im Semantic Web ein hohes Nutzungspotenzial verspricht. Letztlich ist der grundlegende Ansatz aber nicht auf RDF beschränkt, sondern kann auch auf strukturell ähnliche Datenmodelle angewandt werden. Voraussetzung ist jedoch, dass eine Objektmenge existiert, die mehrere Eigenschaften teilt, und dass der Datentyp dieser Eigenschaften wohldefiniert ist.

Ebenso müssen zur Visualisierung der Daten nicht notwendigerweise Streudiagramme eingesetzt werden, sondern lassen sich auch andere, stärker auf den Datentyp und das Analyseziel angepasste, Darstellungsformen zur Exploration mit semantischen Linsen verwenden. Beispielsweise können Karten für Geodaten oder Graphvisualisierungen für semantische und soziale Netze durchaus geeignete Visualisierungsalternativen bieten. Der Vorteil von Streudiagrammen ist jedoch ist die generische Anwendbarkeit auf nahezu beliebige Datentypen.

Literaturverzeichnis

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In: *Proceedings of ISWC '07*. Berlin/Heidelberg: Springer, S. 722-735.

Bier, E., Stone, M., Pier, K., Buxton, W. & DeRose, T. (1993). Toolglass and Magic Lenses: The See-Through Interface. In: *Proceedings of SIGGRAPH '93*. New York: ACM, S. 73-80.

Fishkin, K. & Stone, M. C. (1995). Enhanced Dynamic Queries via Movable Filters. In: *Proceedings of CHI '95*. New York: ACM, S. 415-420.

Pietriga, E., Bizer, C., Karger, D. and Lee, R. (2006). Fresnel: A Browser-Independent Presentation Vocabulary for RDF. In: *Proceedings of ISWC '06*. Berlin/Heidelberg: Springer, S. 158-171.

Quan, D., Huynh, D. and Karger, D. (2003). Haystack: A Platform for Authoring End User Semantic Web Applications. In: *Proceedings of ISWC '03*. Berlin/Heidelberg: Springer, S. 738-753.

Rotard, M., Giereth, M., & Ertl, T. (2007). Semantic Lenses: Seamless Augmentation of Web Pages with Context Information from Implicit Queries. In: *Comput. Graph.* 31(3), S. 361-369.

Ryall, K., Li, Q. & Esenther, A. (2005). Temporal Magic Lens: Combined Spatial and Temporal Query and Presentation. In: *Proceedings of INTERACT '05*. Berlin/Heidelberg: Springer, S. 809-822.

Viega, J., Conway, M. J., Williams, G. & Pausch, R. (1995): 3D Magic Lenses. In: *Proceedings of UIST '95*. New York: ACM, S. 51-58.