Entwicklung eines Evaluations-Frameworks für instanzbasierte Ontologie-Matching-Verfahren

Katrin Zaiß
Institut für Informatik
Heinrich-Heine-Universität Düsseldorf
D-40225 Düsseldorf, Deutschland
zaiss@cs.uni-duesseldorf.de

Zusammenfassung

Ontologien sind ein weit verbreitetes Modell zur Repräsentation von Wissen, und werden u.a. im Semantic Web eingesetzt. In vielen verschiedenen Anwendungsgebieten wie z.B. der Informationsintegration, ist ein Matching der Ontologien notwendig. Es existieren schon einige Matchingsysteme, deren Qualität durch die Einführung von und die Teilnahme an Evaluations-Initiativen wie der OAEI, gut miteinander vergleichen werden kann. Für instanzbasierte Matcher bzw. für Systeme, die hauptsächlich Instanzinformationen zum matchen nutzen, gibt es aber kaum geeignete Test-Ontologien oder -Szenarien. Dieser Beitrag soll die Anforderungen an ein solches Evaluations-Framework spezifizieren und erste Ideen zur Umsetzung eines solchen aufzeigen.

1 Einleitung

Ontologien sind ein weit verbreitetes Modell zur Repräsentation von Wissen, und werden u.a. im Semantic Web eingesetzt. Ziel ist es vor allem das Wissen maschinenlesebar zu machen. So können z.B. Zusammenhänge zwischen Personen und personenbezogenen Daten wie Email o.ä. durch die Verwendung einer Ontologie eindeutig definiert werden. Das Matching von verschiedenen Ontologien (oder auch Schemata) kann in vielen verschiedenen Anwendungen notwendig oder sinnvoll sein und ist daher ein recht gut erforschtes Gebiet. Es existieren einige Matching-Systeme, wie z.B. Coma [DR02], Automatch [BM02], QOM [ES04] oder FCA-Merge [SM01], die verschiedene Matcher verwenden um ein möglichst gutes Mapping zu finden. Die verschiedenen Methoden können grob unterteilt werden in konzept- und instanzbasierte Matcher (siehe [RB01]). Konzeptbasierte Matching-Methoden nutzen die Metainformationen der Konzepte, wie z.B. Label, Kommentare, Datentypen o.ä. , welche mit geeigneter Distanz- oder Ähnlichkeitsfunktionen, wie der Edit-Distanz z-b. , verglichen werden. Dies ist ein sehr sinnvoller Ansatz, falls die Informationen recht ähnlich sind und die Struktur der Ontologien nicht zu unterschiedlich ist. Durch die Verwendung externer Quellen wie Thesauri, z.B. WordNet, können sogar Synonyme oder Homonyme entdeckt werden. Dennoch gibt es einige Konflikte, die nicht immer mit konzeptbasierten Methoden gelöst werden können. Ontologien werden normalerweise manuell erstellt, so dass die Metainformationen immer das subjektive Verständnis des Entwicklers widerspiegeln. Möglicherweise ergeben Metainformationen auch nur in einem sehr speziellen Kontext Sinn oder Konzepte werden unterschiedlich detailliert/strukturiert dargestellt. In diesen Fällen können instanzbasierte Matching-Methoden helfen semantisch äquivalente Klassen trotzdem zu finden. Instanzen liefern in ihrer Gesamtheit sogar mehr Informationen über ein Konzept als die Metainformationen. Die Schwierigkeit besteht darin diese Instanzinformationen in geeigneter Form aufzubereiten, so dass sie vergleichbar sind. Es wurden einige instanzbasierte Matcher entwickelt wie z.B. Glue [DMDH04], Coma++ [EM07], oder [ZSC08]. Das Mapping

wird entweder bestimmt mit Hilfe von Instanz-Duplikaten oder durch die Berechnung von Eigenschaften wie Durchschnittswerte, Wortverteilungen o.ä.; andere dagegen klassifizieren Instanzen mit Hilfe von Bayes-Klassifikatoren.

Um instanzbasierte mit konzeptbasierten Methoden oder auch komplette Systeme untereinander vergleichen oder testen/verbessern zu können, müssen entsprechende Evaluations-Frameworks definiert werden. Innerhalb des Frameworks sollten verschiedene Ontologien und Testszenarien enthalten sein, die einige Anforderungen erfüllen müssen, z.B. verschiedenartig strukturierte Ontologien. Es existieren schon derartige Frameworks, wie z.B. das von der Ontology Alignment Evaluation Initiative entwickelte, die aber vor allem für instanzbasierte Matching-Algorithmen nicht gut geeignet sind.

Der Rest des Papers ist wie folgt organisiert: In Kapitel 2 werden verwandte Arbeiten beschrieben und deren Defizite in Bezug auf instanzbasierte Matching-Verfahren erläutert. Kapitel 3 definiert die Anforderungen an ein Evaluations-Framework, deren Umsetzung in Kapitel 4 beschrieben wird. Abschließend gibt es eine Zusammenfassung und einen Ausblick in Kapitel 5.

2 Verwandte Arbeiten

Es gibt schon einige Frameworks zur Evaluation von Ontologie-Matching-Verfahren. Als wichtigstes ist die Ontology Alignment Evaluation Initiative (kurz: OAEI, [Oae08]) zu nennen. Die OAEI veröffentlicht jährlich verschiedene Testszenarien, die jedes Matching-System durchführen kann. Die Ergebnisse werden bei einem Workshop im Rahmen der International Semantic Web Conference veröffentlicht und diskutiert. Für unsere Zwecke besonders interessant sind die Benchmark-Tests, weil dort viele verschiedene Ontologien vorhanden und sehr viele verschiedene Tests durchgeführt werden. Zusätzlich sind die korrekten Alignments angegeben, was die Evaluation, d.h. vor allem die Bestimmung von Precision und Recall, vereinfacht. Die Referenz-Ontologie besteht aus 33 Konzepten, 64 Attributen (40 Objekt-Eigenschaften und 24 Datentyp-Eigenschaften), 56 Instanzen und 20 anonymen Instanzen. Insgesamt gibt es zusätzlich zur Referenz-Ontologie noch 50 weitere Ontologien, bei denen es sich immer um eine modifizierte Referenz-Ontologie handelt. Die Modifikationen umfassen die folgenden:

- Einbauen von Rechtschreibfehlern,
- Verändern der Schreibweise (Groß-/Kleinschreibung etc.),
- Ersetzen von Konzeptbezeichnungen durch ihre Synonyme oder zufällige Strings,
- Entfernen oder Übersetzen von Kommentaren,
- Übersetzen in eine andere Sprache (z.B. Französisch),
- Generalisierung der Sprache, Erweiterung von Klassen,
- Abflachung, Erweiterung oder Unterdrückung der Hierarchie.

Diese modifizierten Ontologien werden gegen die Referenz-Ontologie gematcht, und Precision und Recall werden mit Hilfe der angegeben korrekten Alignments bestimmt. Es ist wichtig anzumerken, dass hier nur 1:1 Korrespondenzen zwischen Konzepten und Attributen gefunden werden können bzw. sollen, und auch nur diese in den vorgegebenen Alignments zu finden sind. Die Struktur des Benchmarks ist sehr gut geeignet um konzeptbasierte Matcher bzw. Systeme, die vorwiegend solche enthalten, zu testen. Die sehr geringe Anzahl an Instanzen benachteiligt aber eindeutig (vorwiegend) instanzbasierte Matching-Systeme.

Ein weiteres Framework, welches aber eher zum Matchen von Instanzen geeignet ist, ist das ISLAB Instance Matching Benchmark. Die Referenz-Ontologie enthält 5 Klassen, 17 Attribute (4 Objekt-Eigenschaften, 13 Datentyp-Eigenschaften) und 302 Instanzen. Auch hier handelt es

sich bei den weiteren Ontologien um Modifikationen der Referenz-Ontologie, allerdings beziehen sich die Modifikationen ausschließlich auf die Instanzebene. Die Instanzwerte werden verändert, in dem Rechtschreibfehler eingefügt, Werte gelöscht/permutiert oder gleiche Instanzen in unterschiedliche Klassen eingeordnet werden. Die Ontologien dieses Benchmarks enthalten nur wenige Klassen, so dass die entsprechenden Tests für Matchingssysteme, die nicht ausschließlich auf Basis von Instanz-Matching arbeiten, nicht sehr geeignet sind.

Ein kürzlich neu entwickeltes Benchmark ist das STBenchmark [ATV08]. An einer vorgegeben Eingabe-Ontologie (z.B. aus DBLP oder BioWarehouse) werden verschiedene Transformationen durchgeführt, so dass jeweils ein Mapping-Szenario entsteht. Die erforderlichen Instanzen werden mit Hilfe des Datengenerators ToXGene künstlich erzeugt, was ein Nachteil dieses Systems ist. Zusätzlich ist die Anwendung des Systems nicht leicht verständlich (was den Anforderungen an ein Evaluations-Framework widerspricht, siehe nächstes Kapitel), und die korrekten Alignments sind nicht angegeben, was eine Evaluation erschwert.

3 Anforderungen

Wenn man ein Evaluations Framework entwickeln will, dann gibt es einige Prinzipien, die man beachten sollte. Im Folgenden sollen diese, wie in [ES07] beschrieben, kurz erläutert werden.

- Systematik: Die Tests müssen eindeutig und nachvollziehbar sein, und ihre Durchführung, auch zu verschiedenen Zeitpunkten, muss vergleichbar sein.
- Kontinuität: Eine kontinuierliche Wiederholung/Wiederholbarkeit der Tests soll gegeben sein, damit eine Entwicklung/Verbesserung festgestellt werden kann.
- Qualität und Quantität: Die Definition der Bewertungsregeln muss exakt und unmissverständlich formuliert sein. Außerdem sollte die Qualität der Test-Ontologien so gut wie möglich sein und keine der Test-Sets darf eine bestimmte Klasse von Matching-Systemen bevorzugen.
- Verbreitung: Das Benchmark und die Evaluationsresultate sollten frei zugänglich sein.
- Verständlichkeit: Die Resultate sollten analysiert werden können und für alle verständlich sein. Daher sollten nicht nur die allgemeinen Resultate, sondern auch die von den Systemen berechneten Alignments zur Verfügung gestellt werden.

Diese Prinzipien gelten allgemein für die Erstellung von Evaluations-Frameworks. Für unser Framework, welches auch bzw. insbesondere instanzbasierten Matching-Methoden eine Evaluation erlauben soll, definieren wir zusätzlich folgende Anforderungen:

- große Anzahl von Instanzen: Um die Skalierung und auch die Qualität von Matching-Algorithmen in möglichst realitätsnahen Szenarien testen zu können, muss die Anzahl der Instanzen ausreichend sein. Zudem sollten die vorhandenen Werte hinreichend unterschiedlich sein und nicht nur aus einigen wenigen aber oft wiederholten Instanzen bestehen (es sei denn, dies ist eine Eigenschaft des dazugehörigen Attributs).
- An- und Abwesenheit von Duplikaten: Wie oben beschrieben nutzen einige Systeme die Anwesenheit von Duplikaten um ein Mapping zu bestimmen. Da diese Systeme weder benach- noch bevorteiligt werden soll, sollte es Szenarien mit und ohne Instanz-Duplikate geben.
- unterschiedliche Strukturen: Ontologien können unterschiedlich strukturiert sein, d.h. sie sind z.B. unterschiedlich detailliert, obwohl sie semantisch ähnlich sind, oder enthaltenen andere/zusätzliche Relationen oder Attribute.

- unterschiedliche Formatierung: Instanzen können bei gleicher Semantik unterschiedlich formatiert sein, ein gutes Beispiel dafür ist das Datum. Um zu testen, in wie weit Matcher semantisch ähnliche aber unterschiedlich formatierte Instanzen verwenden können, sollen unterschiedliche Variationen in verschiedenen Ontologien vorhanden sein.

- Einbeziehung von Rechtschreibfehlern: (Reale) Ontologien werden größtenteils von Menschen erstellt, so dass die Instanzen natürlicherweise auch Rechtschreibfehler wie eingefügte oder ausgelassene Zeichen oder nicht korrekte Anwendung von Groß- und Kleinschreibung enthalten können.

- 1:n Mappings: Innerhalb des Frameworks soll es auch Ontologiepaare geben, in denen man 1:n Korrespondenzen finden kann, was eine logische Folge der Forderung von verschiedenen strukturierten Ontologien ist.

4 Ideen zur Umsetzung

Im vorherigen Kapitel wurden die allgemeinen und die speziellen Anforderungen an ein Framework zur Evaluation von (instanzbasierten) Matching-Sytemen oder -Algorithmen definiert. In diesem Kapitel sollen einige Umsetzungsmöglichkeiten skizziert und deren Vor-und Nachteile diskutiert werden.

Allgemeines Ziel ist die Erstellung eines Benchmarks, dass eine große Menge an Ontologien enthält. Ähnlich zu dem Benchmark der OAEI soll es eine Referenz-Ontologie geben, die sich auf eine bestimmte Domäne beschränkt. Die Wahl der Domäne hängt von der technischen Umsetzung ab, auf die später näher eingegangen wird. Desweiteren sollen die im vorherigen Kapitel und die von der OAEI beschriebenen Modifikationen der Ontologien (unterschiedliche Struktur/Formatierung etc.) umgesetzt und auch in verschiedenen Kombinationen zusammengesetzt werden. Zusätzlich ist es sinnvoll, die Referenzontologie in einigen Testszenarien auch um einige verwandte Konzepte/Themenbereiche zu erweitern (was auch eher der realen Welt entspricht), so dass das Verhalten der Systeme in diesen Fällen auch bewertet werden kann. Im Allgemeinen ist die Organisation der Testszenarien und die Erstellung unterschiedlicher Ontologien nicht sehr schwierig, wenn man erst einmal eine geeignete Referenzontologie hat.

Die Erstellung einer Referenzontologie ist differenzierter zu betrachten. Wir möchten möglichst realistische Ontologien für unsere Framework verwenden, so dass wir keine künstlichen Datengeneratoren und ausgedachte Ontologien verwenden, sondern auf Webinhalte zurückgreifen wollen. Grundlage dazu bilden ein Webcrawler und ein Parser, die geeignete Webseiten untersuchen und die passenden Inhalte extrahieren. Eine Möglichkeit wäre z.B. die Extraktion der Daten von der DBLP-Website. Die dort angebotenen Informationen sind in einer festen Struktur präsentiert, so dass man relativ leicht eine Ontologie extrahieren kann (manuell oder ggf. automatisch). Die Extraktion der Daten kann in jedem Fall automatisch durchgeführt werden. Ein Vorteil einer so erzeugten Ontologie ist, dass die Daten real und durch die Struktur leicht zu extrahieren sind. Allerdings wird die Bibliographie-Domäne schon von der OAEI genutzt, so dass ein anderes Gebiet für die Evaluation vielleicht etwas aufschlussreicher wäre. Eine andere Möglichkeit wäre die Nutzung der Wikipedia-Seiten (siehe auch [WWA+]). Insbesondere auf den englischen Seiten steht bei zahlreiche Themen eine Infobox zur Verfügung, die eine klare Struktur hat (zu finden auf http://de.wikipedia.org/wiki/Kategorie:Vorlage:Infobox). Diese Struktur könnte manuell oder automatisch in eine Ontologie transformiert werden, die Instanzen könnten automatisch von der entsprechenden Seiten extrahiert und den Konzepten zugeordnet werden. Interessant wäre es auch Ontologien zu einem Thema von verschiedensprachigen Seiten zu extrahieren und zu matchen. Passende Links auf den Seiten kann man auch als Relationen zu anderen Konzepten auffassen. Generell kommen verschiedenartige Websites in Frage, deren Eignung durch genauere Untersuchung und verschiedene Tests festgestellt werden muss.

5 Zusammenfassung und Ausblick

Das Matching von Ontologien ist ein weit verbreitetes Problem für das schon einige Lösungen existieren. Um die verschiedenen Ansätze sinnvoll vergleichen oder auch verbessern zu können ist es wichtig, dass geeignete Frameworks zur Evaluation entwickelt werden. Einige solcher Evaluationsinitiativen existieren schon, aber diese enthalten nur sehr wenige Instanzen und sind deswegen für instanzbasierte Matcher nicht sehr geeignet. Es wurden einige Anforderungen an das zu entwickelte Evaluations-Framework definiert. Die Umsetzung soll hauptsächlich mit Hilfe eines Webcrawlers und eines Parsers erfolgen, die automatisch Ontologien erzeugen bzw. Instanzen für einen vorgegebene Ontologie extrahieren. In naher Zukunft sollen die Erzeugung der Ontologien anhand verschiedener Webseiten getestet werden. Diese Referenzontologien müssen modifiziert werden, so dass ein Benchmark von verschiedenen Ontologien entsteht. Nachdem die erforderlichen Alignments definiert worden sind, sollen die Tests mit frei verfügbaren Systeme und unseren selbst entwickelten Matchern durchgeführt werden. Sobald die Entwicklung des Frameworks abgeschlossen ist, soll es der Allgemeinheit zur Verfügung gestellt werden.

Literatur

[ATV08] Bogdan Alexe, Wang-Chiew Tan, and Yannis Velegrakis. STBenchmark: Towards a benchmark for mapping systems. *Proc. VLDB Endow.*, 1(1):230–244, 2008.

[BM02] Jacob Berlin and Amihai Motro. Database Schema Matching Using Machine Learning with Feature Selection. In *Advanced Information Systems Engineering, 14th International Conference, CAiSE 2002, Toronto, Canada, May 27-31, 2002, Proceedings*, pages 452–466, 2002.

[DMDH04] AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Y. Halevy. Ontology Matching: A Machine Learning Approach. In *Handbook on Ontologies*, pages 385–404. Springer, 2004.

[DR02] Hong Hai Do and Erhard Rahm. Coma - a system for flexible combination of schema matching approaches. In *VLDB 2002, Proceedings of 28th International Conference on Very Large Data Bases, August 20-23, 2002, Hong Kong, China*, pages 610–621, 2002.

[EM07] Daniel Engmann and Sabine Maßmann. Instance Matching with COMA++. In *Datenbanksysteme in Business, Technologie und Web (BTW 2007), Workshop Proceedings, 5.-6. März 2007, Aachen, Germany*, 2007.

[ES04] Marc Ehrig and Steffen Staab. QOM - Quick Ontology Mapping. In *INFORMATIK 2004 - Informatik verbindet, Band 1, Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Ulm, 20.-24. September 2004*, pages 356–361. GI, 2004.

[ES07] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.

[Oae08] http://oaei.ontologymatching.org/2008/, 2008.

[RB01] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB J.*, 10(4):334–350, 2001.

[SM01] Gerd Stumme and Alexander Maedche. FCA-MERGE: Bottom-Up Merging of Ontologies. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*, pages 225–234, 2001.

[WWA+] Daniel S. Weld, Fei Wu, Eytan Adar, Saleema Amershi, James Fogarty, Raphael Hoffmann, Kayur Patel, and Michael Skinner. Intelligence in wikipedia. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 1609–1614.

[ZSC08] Katrin Zaiß, Tim Schlüter, and Stefan Conrad. Instance-Based Ontology Matching using Regular Expressions. In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems: OTM 2008 Workshops, ODBase 2008, LNCS 5333, 9-14. November 2008, Monterrey, Mexico*, pages 40–41. Springer-Verlag, 2008.