Die Suche nach Wissen statt nach Webseiten

Fabian Suchanek und Gerhard Weikum
Max-Planck-Institut fuer Informatik, Saarbrücken
Abteilung: Datenbanken und Informationssysteme
suchanek@mpii.mpg.de, weikum@mpii.mpg.de

January 18, 2008

Abstract

Internet-Suchmaschinen können heutzutage bereits Webseiten finden, die bestimmte Wörter enthalten. Will man aber Informationen finden, die auf verschiedenen Seiten verstreut sind, oder ist man gar an einem logischen Zusammenhang zwischen diesen Informationen interessiert, so helfen die heutigen Suchmaschinen nicht viel. Unser Ansatz ist es deshalb, Information gezielt aus Webseiten zu sammeln und in einer großen Wissensstruktur, einer Ontologie, anzuordnen. Durch strukturelle Analyse, Musterextraktion und statistisches Lernen können wir eine solche Ontologie aus der Enzyklopädie Wikipedia (und anderen Internet-Quellen) automatisch aufbauen. Das Ergebnis, die Ontologie Yago, enthält rund eine Million Entitäten und über 6 Millionen Fakten. Yago hat auch ein Web-Interface, das komplexe Wissensanfragen online beantworten kann. Yago könnte der Ausgangspunkt für eine neue Generation von Suchmaschinen werden, sowohl fuer die Suche im Internet, als auch für die Suche in digitalen Bibliotheken oder E-Science-Repositories. Dieser Artikel beschreibt den Stand von Yago im Jahre 2007.

Today's Web search engines can find Web pages that contain certain keywords. Up to now, however, any advanced information demands that concern facts from multiple Web pages, let alone a logical connection between them, are inherently beyond the answering capabilities of search engines. This is why our approach is to collect information from Web sites and to organize it in a huge knowledge structure, an ontology. Using pattern extraction, structural analysis and statistical learning methods, we have developed tools that can automatically build and maintain such an ontology from the contents of the Wikipedia encyclopedia and other Internet sources. Our ontology, coined Yago, contains about one million entities and concepts, and knows more than six million facts about them. Yago also has a Web interface for answering knowledge queries online. Yago could be the starting point for a new generation of search engines, for searching the Web as well as digital libraries and e-science repositories. This article describes the state of Yago in the year 2007.

Die Suche im Internet mit Suchmaschinen

Im letzten Jahrzehnt hat sich das Internet zu einer bedeutenden Informationsquelle entwickelt. Bahnfahrpläne, Nachrichten, wissenschaftliche Artikel, Unternehmensdaten, ja sogar ganze Enzyklopädien sind inzwischen online verfügbar. Der Großteil dieser Internetseiten wird von Suchmaschinen erfasst. Google beispielsweise erlaubt uns, Milliarden von Internetseiten innerhalb von Sekundenbruchteilen nach Suchwörtern zu durchforsten. Fuer viele Suchanfragen ist diese Technologie völlig ausreichend. Meistens finden wir nach kurzem Durchstöbern der angezeigten Ergebnisse das Gesuchte. Ist man beispielsweise an dem Physiker Max Planck interessiert, so liefert die Suche nach "Max Planck" direkt nach dem erstplatzierten Link auf die Max-Planck-Gesellschaft mehrere Biographien des Physikers. Selbst Fragen wie "Wann wurde Max Planck geboren?" lassen sich durch einen kurzen Blick auf die Ergebnisseite direkt beantworten. Die englische Version von Google antwortet auf die Frage "When was Max Planck born?" gar wie aus der Pistole geschossen: "Max Planck - Date of Birth: 23 April 1858".

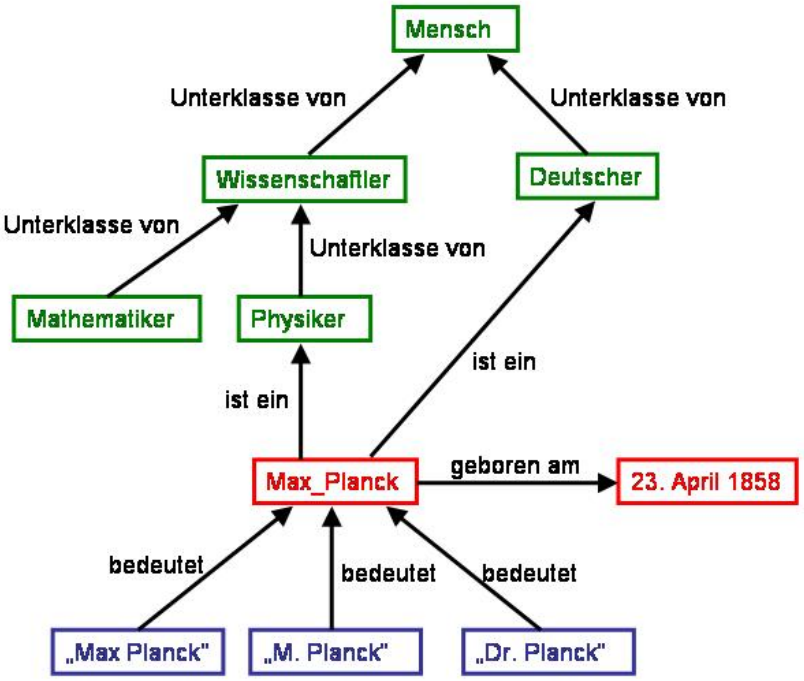
Trotzdem stößt man als Internetbenutzer gelegentlich an die Grenzen dieser Technologie. Möchte man beispielsweise wissen, welche Physiker noch im selben Jahr wie Max Planck geboren wurden, so lässt sich diese Frage kaum Google-tauglich formulieren. Alle Anfragen nach "Physiker, geboren, Jahr, Max Planck" geben lediglich Max Planck selbst zurück. Man ist also gezwungen, zunächst das Geburtsdatum von Max Planck zu ergoogeln und sodann nach Physikern zu fragen, die ebenfalls in diesem Jahr geboren wurden. Wenn man dann komplexere Dinge wissen möchte (Welche dieser Physiker waren zudem auch politisch aktiv?), so führt kein Weg daran vorbei, die entsprechenden Internetseiten durchzulesen.

Der Grund fuer diese Unbequemlichkeit ist, dass Google nicht Wissen durchsucht, sondern Webseiten. Google kann nur diejenigen Informationswünsche befriedigen, zu denen es bereits eine vorgefertigte Antwort auf einer Webseite gibt. Findet sich die Antwort auf mehreren Seiten verstreut oder ist gar eine Antwort nur durch logische Schlussfolgerung zu erlangen, so ist man mit Google an der falschen Adresse. Abstrakt betrachtet ist das Problem, dass heutige Computer lediglich über Texte verfügen, nicht aber über Wissen. Dieser Mangel an Allgemeinwissen ist beispielsweise ebenfalls dafür verantwortlich, dass maschinelle Übersetzung teilweise erheiternde Ergebnisse produziert. Wenn es gelänge, das Wissen dieser Welt dem Computer in einer großen Wissensstruktur zur Verfügung zu stellen, so könnte er diese Aufgaben sehr viel leichter bewältigen.

Wissensrepräsentation in Ontologien

Eine solche strukturierte Wissenssammlung heißt "Ontologie". Im einfachsten Fall ist eine Ontologie ein gerichteter Graph, dessen Knoten so genannte

"Entitäten" sind und dessen Kanten "Relationen" sind. Beispielsweise steht die Entität "Max Planck" mit der Entität "23. April 1858" in der Relation "geboren am", denn Max Planck wurde am 23. April 1858 geboren. Obwohl dieses Modell gewissen Einschränkungen unterliegt, lassen sich damit bereits viele Wissensbausteine auf einfache Weise ausdrücken.



Entitäten, die viele gemeinsame Eigenschaften haben, fassen wir in so genannten Klassen zusammen. Max Planck beispielsweise, ebenso wie seine physikbegeisterten Kollegen, gehört der Klasse "Physiker" an. In der Ontologie ist die Klasse "Physiker" nichts anderes als eine weitere Entität, die mit der Relation "ist ein" mit den Physikern verbunden wird. Jeder Physiker ist ein Wissenschaftler, sodass die Klassen "Physiker" und "Wissenschaftler" in der "Unterklasse von"-Relation stehen. Dadurch ergibt sich eine Hierarchie von Klassen, in der jeweils die obere (allgemeinere) die unteren (spezielleren) einschließt.

Als nächsten Abstraktionsschritt führen wir eine Unterscheidung ein zwischen Wörtern und ihren Bedeutungen. Wir unterscheiden also zwischen "Max Planck" (dem Wort) und Max Planck (dem Physiker). Dies ist sinnvoll, da verschiedene Wörter sich auf dasselbe Individuum beziehen können (beispielsweise "Dr. Planck" oder "M. Planck"). Umgekehrt kann dasselbe Wort sich auch auf unterschiedliche Individuen beziehen (es gibt z.B. mehrere Personen mit dem Namen "Planck"). Darüber hinaus abstrahieren wir durch diese Un-

terscheidung über die Wahl der Sprache hinweg. So können sich dann, vereinfacht, die Wörter "Physiker", "physicist" und "physicien" allesamt auf die Klasse "Physiker" beziehen. In der Ontologie sind die Wörter nichts anderes als weitere Entitäten.

Oft ergänzt man eine solche Ontologie um Regeln zur logischen Schlussfolgerung (Axiome). Eines der grundlegendsten Axiome besagt beispielsweise, dass eine Entität allen Oberklassen ihrer Klasse angehört. Wenn also bekannt ist, dass Max Planck ein Physiker war, so folgt aus der Unterklassenbeziehung von Physiker und Wissenschaftler, dass Max Planck ein Wissenschaftler war. Wenn jeder Physiker ein Wissenschaftler ist und jeder Wissenschaftler ein Mensch, so ist jeder Physiker ein Mensch (Transitivität der Unterklassenbeziehung). Ein Axiomensystem kann auch ausdrücken, dass zwei Relationen invers zueinander sind, dass bestimmte Relationen einander kausal bedingen oder dass Zeitintervalle einander einschließen. Auf diese Weise kann der Computer aus Max Plancks Geburtsort Kiel, seinem Geburtsdatum und seinem Todestag beispielsweise den logischen Schluss ziehen, dass Max Planck ein deutscher Wissenschaftler war, der beiden Weltkriege durchlebt hat.

Weitergehende Wissensrepräsentationen verwenden logische Formeln zur Darstellung von Zusammenhängen wie etwa: Jeder Mensch hat zwei Eltern unterschiedlichen Geschlechts, ein promovierter Wissenschaftler hat einen Doktorvater, ein Professor muss publiziert haben, usw. Unter Umständen wollen wir auch spekulatives Wissen repräsentieren oder aufgrund von Mehrdeutigkeiten Spielräume für unterschiedliche Interpretationen lassen, die wir dann mit Wahrscheinlichkeiten versehen können. Beispielsweise kann mit "Paris" die französische Hauptstadt oder die Figur aus der Ilias gemeint sein, oder wir wollen konkurrierende Hypothesen zu den Ursachen einer bestimmten Krankheit in die Ontologie aufnehmen, oder wir wollen die Messungenauigkeit für die Sonnenumlaufzeit des Mars festhalten. In diesen Fällen müssen wir zur Wissensrepräsentation logische und probabilistische Methoden kombinieren.

Ontologien spielen eine zentrale Rolle in der Vision des "Semantic Web", das vom WWW-Erfinder Tim Berners-Lee als Nachfolgegeneration der derzeitigen Web2.0-Welle gesehen wird. Es soll dann möglich werden, Webseiten direkt in Verbindung mit Entitäten und den dahinter stehenden kognitiven Konzepten zu bringen und über Logikkalküle intelligente Schlussfolgerungen zu ziehen, um beispielsweise den besten Klarinettenlehrer zu finden, den die Tochter des Hauses in weniger als einer halben Stunde von ihrem Gymnasium aus erreichen kann. Dazu müssen jedoch auch alle Webseiten entsprechend explizit mit ontologischen Konzepten annotiert und in einem Logikformalismus repräsentiert werden. Heute erfordert ein solches Unterfangen noch sehr hohen und fehleranfälligen manuellen Aufwand für jede einzelne Webseite, so dass fundamentale Skalierbarkeitsprobleme einer schnellen Realisierung der Vision (noch) im Wege stehen. Unsere aktuellen Forschungsarbeiten zur intelligenten Wissenssuche haben mit der Semantic-Web-Vision das Ziel gemeinsam, verwenden aber Meth-

oden, die von heute direkt verfügbaren Datenquellen ausgehen und daraus automatisch umfangreiche Wissenssammlungen aufbauen.

Automatische Konstruktion und Pflege von Ontologien

Die entscheidende Frage ist nun, wie man eine Ontologie mit Wissen füllt. Dazu gibt es mehrere Ansätze. Eine Möglichkeit ist, die Entitäten und Relationen alle von Hand einzufügen. In der Tat sind die am weitesten verbreiteten Ontologien heutzutage in manueller Kleinarbeit erstellt worden: WordNet ist ein Lexikon der englischen Sprache mit rund 200'000 Begriffen in einer ontologischen Struktur. SUMO ist eine Ontologie mit Hunderttausenden von Entitäten, und die kommerzielle Ontologie Cyc enthält gar zwei Millionen Fakten und Axiome. Trotz dieser Wissensmengen wird eine von Hand erstellte Ontologie stets der aktuellen Entwicklung hinterherhinken. Keine der genannten Ontologien kennt beispielsweise das neueste Windows System oder die Fußballstars der letzten WM.

Am Max-Planck-Institut für Informatik verfolgen wir daher andere Ansätze zur Konstruktion und Pflege der Ontologie. Ein Ansatz nutzt dazu die große Online-Enzyklopädie Wikipedia. Wikipedia enthält Artikel zu Abertausenden von Persönlichkeiten, Produkten, Begriffen und Organisationen. Jeder dieser Artikel ist bestimmten Kategorien zugeordnet. So befindet sich beispielsweise der Artikel über Max Planck in den Kategorien "Deutscher", "Physiker" und "Geboren 1858". Diese Information machen wir uns zunutze, um Klassenzugehörigkeit und Geburtsdatum der Entität "Max Planck" in der Ontologie zu vermerken. Wikipedia kennt zwar eine große Anzahl von Individuen, stellt aber keine gut strukturierte Hierarchie von Klassen zur Verfügung. Die Information, dass "Physiker" "Wissenschaftler" sind und "Wissenschaftler" "Menschen" ist sehr schwer in Wikipedia zu finden. Wir kombinieren daher die Daten aus Wikipedia durch automatisierte Verfahren mit den Daten aus der oben genannten WordNet-Ontologie. Dadurch erhalten wir bereits eine sehr große Wissensstruktur, in der alle in Wikipedia bekannten Entitäten ihren Platz haben. Auch andere strukturierte Wissensquellen (wie etwa die Filmdatenbank IMDB) nutzen wir aus.

Leider ist jedoch nicht alles Wissen bereits in strukturierter Form verfügbar. Die häufigste Form von Internetseiten ist unstrukturierter, natuerlichsprachlicher Text. Beispiele sind Biographien, Lexikoneinträge oder Nachrichtentexte. Um auch diese Information einzusammeln, nutzt man einen Ansatz namens "Pattern Matching". Will man beispielsweise neue Geburtsdaten in die Ontologie einfügen, so findet man zunächst anhand bekannter Geburtsdaten heraus, nach welchem Muster Geburtsdaten häufig auf Webseiten genannt werden. Ein sehr gängiges Muster fuer Geburtsdaten ist z.B. "X wurde am Y geboren" ("Max Planck wurde am 23. April 1858 geboren"). Durchsucht man nun das Internet nach weiteren Vorkommnissen dieses Musters, so werden andere Paare aus Person und Geburtsdatum zu Tage gefördert. Diese kann man

dann in die Ontologie eintragen. Dieser Ansatz leidet darunter, dass bereits leichte Veränderungen im Satzbau das Muster zerstören können. Beispielsweise passt das Muster "X wurde am Y geboren" nicht auf den Satz "Max Planck, der große Physiker, wurde am 23. April 1958 geboren". Wir haben deshalb den Pattern-Matching-Ansatz so verfeinert, dass er die grammatikalische Struktur der Sätze mit einbezieht. Das Muster fordert dann lediglich, dass X das Subjekt des Prädikats "wurde geboren" sein muss, welches über die Präposition "am" mit Y verbunden ist. Dieses Muster passt nun auch auf den Satz "Max Planck, der große Physiker, ...".

Die hinter diesem Lernprozess stehenden Musterextraktionen sind in dem am Institut entwickelten Softwarewerkzeug Leila implementiert. Damit Leila nicht durch die Vielfalt und Unschärfe der natürlichen Sprache in die Irre geführt wird und allzu schnell falsche Hypothesen für Muster generiert, werden Musterkandidaten durch ein statistisches Lernverfahren auf ihre Robustheit getestet. Auf diese Weise extrahiert Leila überwiegend korrekte Fakten. Beispielsweise kann Leila aus der Gesamtheit aller Wikipedia-Artikel mit hoher Konfidenz lernen, dass Weltschmerz ein Gefühl ist, dass Kalkutta am Ganges und Paris an der Seine liegt - und zwar aus Sätzen wie "Kalkutta liegt im Delta des Ganges" und "Paris hat viele Museen am Ufer der Seine", und dass Saarländer eine Volksgruppe sind und Hamburger (die Sandwichs) keine.

Yago

Durch die Kombination dieser Techniken ist es uns gelungen, eine sehr große Ontologie herzustellen: Yago (Yet another Great Ontology). Yago kennt momentan fast eine Million Entitäten und weiß rund 6 Millionen Fakten über diese Entitäten. Der Kern von Yago enthält - wie uns eine empirische Evaluation zeigte - nahezu ausschließlich korrekte Fakten, die wir mit unseren robustesten Methoden aus Wikipedia-Artikeln und deren Verknüpfung mit WordNet extrahiert und organisiert haben. Weiteres Wissen können wir durch die Analyse von Webseiten und Datenbanken mit Werkzeugen wie Leila automatisch hinzufügen. Wenn dabei statistische Lernverfahren und Heuristiken zum Tragen kommen, so würde man erwarten, dass die Korrektheitsrate des neuen Wissens abnimmt. Wenn man aber bereits, wie in unserem Fall, eine hochwertige Ontologie als Ausgangspunkt zur Verfügung hat, so kann man neue Hypothesen hinsichtlich ihrer Konsistenz mit dieser Ontologie abgleichen. Man fügt dann also nur diejenigen neuen Fakten hinzu, die nicht mit den bereits vorhandenen im Widerspruch stehen. Auf diese Weise wird die Ontologie um neue, hochwertige Fakten erweitert, die dann wiederum zum Beurteilen weiterer Hypothesen zur Verfügung stehen. In gewisser Weise ist der Lernprozess also selbstregulierend: Je mehr Wissen Yago enthält, desto robuster und einfacher wird das Erwerben von zusätzlichem Wissen.

Wissenssuche

Unsere Wissenssammlung Yago ist online verfügbar[1] und kann über eine spezielle Abfragesprache Anfragen beantworten. So lässt sich die eingangs gestellte Frage "Welche Physiker wurden in selben Jahr geboren wie Max Planck?" fuer Yago wie folgt formulieren:

```
"Max Planck" bornInYear $year
```
(die Variable $year wird nun das Geburtsdatum von Max Planck enthalten)
```
$otherPhysicist bornInYear $year
```
(wir fragen nach einem anderen Menschen, der ebenfalls in $year geboren wurde...)
```
$otherPhysicist isa physicist
```
(...und stellen die Bedingung, dass er ebenfalls Physiker sein muss)

Yago antwortet prompt mit mehreren Dutzend anderen Physikern. Will man wissen, wer von ihnen zudem noch politisch aktiv war, so fügt man die Bedingung "$otherPhysicist isa politician" hinzu. Yago antwortet mit dem Neuseeländer Thomas King, der neben seiner Tätigkeit als Astronom auch im Parlament arbeitete. (Anmerkung: Yago wurde inzwischen erweitert und überholt, sodass man diese Anfrage nun anders formulieren müsste. Das generelle Prinzip gilt aber weiterhin.)

Diese Methoden zur ontologiegestützten Wissenssuche können auch in künftige Suchmaschinen integriert werden und zu einer mächtigeren Form der Wissenssuche und -vernetzung auf den größten Korpora unseres Planeten führen. Am Max-Planck-Institut für Informatik arbeiten wir an Methoden für eine intelligenten Suchmaschine, die die Inhalte aller Webseiten, digitalen Bibliotheken und e-Science-Datenbanken als explizites Wissen - mit Konzepten (z.B.

[1] `http://www.mpi-inf.mpg.de/yago`

Enzyme, Quasare, Dichter, etc.) und Entitäten (z.B. Steapsin, 3C 273, Bertolt Brecht, etc.) und den Relationen dazwischen - repräsentiert und mit hoher Präzision auffindbar macht. Eine solche Suchmaschine wäre ein Durchbruch für den Schritt von der fortgeschrittenen Informationsgesellschaft zu einer modernen Wissensgesellschaft, in der alles Wissen der Menschheit nicht nur im Internet verfügbar ist, sondern auch effektiv genutzt werden kann.

Literatur

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum:
"Yago - A Core of Semantic Knowledge".
16th international World Wide Web conference (WWW 2007).

Fabian M. Suchanek, Georgiana Ifrim and Gerhard Weikum
"Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents"
Knowledge Discovery and Data Mining (KDD), Philadelphia, USA, 2006

Peter Baumgartner and Fabian M. Suchanek:
"Automated Reasoning Support for First Order Ontologies",
Fourth Workshop on Principles and Practice of Semantic Web Reasoning. Lecture Notes in Computer Science, Springer-Verlag, 2006.

Steffen Staab and Rudi Studer:
"Handbook on Ontologies".
Springer-Verlag, 2004.

Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates:
"Unsupervised named-entity extraction from the Web: An experimental study".
Artificial Intelligence 165(1), pp. 91-134, 2005.