Exposé zur Studienarbeit

Methoden zur automatischen Erstellung von Term-Hierarchien aus Phänotypdaten

Christoph Böhm

24. September 2007

Betreuer: Philip Groth und Prof. Dr. Ulf Leser
Institut für Informatik
Lehrstuhl für Wissensmanagement in der Bioinformatik

1 Hintergrund

In dieser Studienarbeit werden Methoden des Information Retrieval beschrieben, mit denen strukturierte Termlisten aus biomedizinischen Publikationen extrahiert werden können. Terme oder Phrasen werden Konzepte der Biologie repräsentieren. Diese Konzepte sollen auf automatischem Wege aus Publikationen, die Phänotypen beschreiben, gewonnen werden. Bestehen Beziehungen (z.B. [ist-ein] oder [Teil-von]) zwischen diesem Phänotyp-Konzepten, kann dies als eine Phänotyp-Ontologie verstanden werden.

1.1 Ontologien

Ontologien werden in den verschiedensten Bereichen verwendet: als gemeinsames Vokabular von Experten einer Domäne; zur eindeutigen Klassifikation von Dingen; zur Repräsentation von Wissen; für die Kommunikation von Softwareagenten; zur Erweiterung von Anfragen oder zur Anreicherung des WWW mit Semantik. Aufgrund dieser sehr unterschiedlichen Anwendungsgebiete in diversen Disziplinen ist es schwer, eine einheitliche Definition von Ontologien zu finden.

In der Informatik hat sich aber folgende Definition durchgesetzt: Eine Ontologie ist eine explizite, formale Spezifikation von gemeinsamen Konzepten[1] einer Domäne (vgl. [Staab and Studer, 2004]). Diese Definition impliziert die Verwendung von formalen Sprachen zur Repräsentation der Konzepte sowie deren Beziehungen untereinander. Des weiteren sind Mechanismen vorstellbar, die diese Verknüpfungen, bzw. Bedingungen denen die Verknüpfungen unterliegen, verifizieren. Diese Definition ist für unseren Anwendungsfall zu streng.

Im Folgenden wird eine Ontologie als eine Menge von Konzepten (d.h. Terme oder Phrasen) verstanden, die ggf. durch [ist-ein]-Beziehungen (engl. is-a) in Verbindung stehen. Diese Einschränkung kann getroffen werden, da unser Ergebnis einfach strukturiert und somit gut verständlich sein soll. Des weiteren genügt die vereinfachte Struktur den anvisierten Anwendungen (siehe folgender Abschnitt).

Interpretiert man die Konzepte als Knoten und die [ist-ein]-Beziehungen als Kanten, ergibt sich eine Baum oder ein gerichteter azyklischer Graph.

1.2 Phänotypen

Nach [Smith et al., 2005] versteht man unter einem Phänotyp das Erscheinungsbild oder die Manifestation einer Menge von Eigenschaften eines Individuums als Resultat eines Genotyps bzw. dessen Interaktion mit der Umwelt. Einzelne Merkmale dieser Effekte[2] von Genotypen werden in der Biologie von Forschern in Publikationen mit natürlicher Sprache beschrieben: Haarausfall, Fellbeschaffenheit oder -veränderung u.v.m. Natürlichsprachliche Darstellungen ohne ein definiertes Vokabular haben den Nachteil, dass sie

[1] WordNet, ebenfalls eine Ontologie, definiert Konzepte als abstrakte Ideen, abgeleitet aus konkreten Instanzen der realen Welt. Siehe: wordnet.princeton.edu/perl/webwn?s=concept

[2] Effekte von Genotypen sind meist als Abweichung einer Norm, d.h. vom 'Wildtyp', beschrieben.

einen Intepretationsspielraum der Texte lassen. Außerdem ist die rechentechnische Verarbeitbarkeit unstrukturierter Daten schwierig. Um einer solchen Sammlung von Texten (d.h. Phänotypbeschreibungen) ein einheitliches Vokabular zu verleihen, kann eine wie oben beschriebene Ontologie verwendet werden.

Die Mammalian Phenotype Ontology [Smith et al., 2005] ist ein Beispiel für eine derartige Struktur[3]. Sie beinhaltet allerdings lediglich ein begrenztes Vokabular für Phänotypen von Säugetieren[4], was die Anwendung auf andere Modellorganismen erschwert. Eine speziesübergreifende Ontologie, die von Biologen akzeptiert wird, steigert neben der eindeutigen Beschreibbarkeit auch die maschinelle Verarbeitbarkeit. Eine Anwendung einer speziesübergreifenden Ontologie von Phänotypkonzepten ist das Clustern von Phänotypen: das Bestimmen von Ähnlichkeiten von Phänotypen (bzw. deren Beschreibungen) auf der Basis von Annotationen mittels Termen einer derartigen Ontologie.

1.3 PhenomicDB

PhenomicDB [Groth et al., 2007] ist eine speziesübergreifende Genotyp/Phänotyp-Datenbank[5]. Sie enthält Phänotypen und deren assoziierten Gene sowie Genindizes und Orthologiebeziehungen. Die Genindizes beziehen sich auf den Entrez Gene Index des National Center of Biotechnology Information (NCBI) [Maglott et al., 2005]. Die Orthologiebeziehungen wurden mittels HomoloGene des NCBI [Wheeler et al., 2007] ermittelt. Weiterhin wurden Annotationen der Gene mittels Konzepten aus der Gene Ontology[6] (GO) [GeneOntologyConsortium, 2006] integriert. Die GO ist ein gemeinschaftlich erstelltes strukturiertes Vokabular zur Annotation von Genen und deren Produkten. Mit diesen integrierten Informationen ermöglicht PhenomicDB Vergleiche von Phänotypen im großen Maßstab über mehrere Spezies hinweg.

PhenomicDB enthält ca. 400.000 Phänotypen für unterschiedlichste Organismen wie z.B. Drosophila melanogaster, Saccharomyces cerevisiae oder Homo sapiens. Version 2.3 beinhaltet Phänotypen für ca. 10.500 humane Gene.

2 Problem

Die Mammalian Phenotype Ontology wurde ausschließlich von Hand erstellt. Dies hat eine hohe Qualität zur Folge, erfordert aber einen großen Arbeitsaufwand.

Als übergeordnete Vision dieser Arbeit soll aus der speziesübergreifenden Genotyp/Phänotyp Datenbank PhenomicDB auf automatischem Wege eine Phänotyp-Ontologie erstellt werden. Die Extraktion von Konzepten und deren Hierarchisierung sollen mittels Information-Retrieval-Verfahren geschehen. Als Besonderheit der Domäne sollen

[3] MPO → [nervous system phenotype] → [abnormal nervous system physiology] → [abnormal dopamine level] → [decreased dopamine level] ist ein Beispiel aus der Mammalian Phenotype Ontology. Siehe www.informatics.jax.org/javawi2/servlet/WIFetch?page=mpAnnotSummary&id=MP:0005643

[4] Die Mammalian Phenotype Ontology enthält ca. 5600 Terme sowie 1880 Synonyme.

[5] PhenomicDB ist öffentlich verfügbar: www.phenomicdb.de

[6] Die Gene Ontology ist ebenfalls öffentlich verfügbar: www.geneontology.org

die erwähnten Links zu den entsprechenden Genen sowie deren Annotationen mit GO-Termen ausgenutzt werden.

3 Ziel

Ziel dieser Studienarbeit soll es sein, Methoden zur automatischen Erzeugung von Konzept-Hierarchien im obigen Sinne zusammenzutragen. Weiterhin soll ein ausgewählter Ansatz prototypisch implementiert werden.

4 Vorgehensweise

In der Zusammenstellung der Methoden zur Term-Hierarchisierung wird versucht den aktuellen Stand der Forschung abzudecken sowie ältere Publikationen einzubeziehen. Der Übersicht wegen sollen die verschiedenen zur Verfügung stehenden Ansätze kategorisiert werden: z.B. Methoden zur Erstellung von Synonymklassen oder Methoden zur Entdeckung von [ist-ein]-Beziehungen. Eine eher technische Unterscheidung, z.B. auf Lernen oder Syntax basierende Verfahren, ist ebenfalls denkbar. Exemplarisch seien hier die Arbeit von Sanderson und Croft [Sanderson and Croft, 1999] sowie die Ausführungen von Grossmann und Frieder [Grossman and Frieder, 2004] genannt.

Weiterhin wird herausgestellt, welche Verfahren zur Evaluation von Ontologien (hier Termhierarchien) die Forschung bisher erarbeitet hat. Dies soll die Erstellung eines Maßes zur Optimierung bzw. die Vergleichbarkeit mit anderen Ontologien für die weitere Arbeit ermöglichen.

Um einen Eindruck von den in PhenomicDB vorliegenden Daten sowie den Arbeitsweisen und Problemen der betrachteten Methoden zu bekommen, wird eines der Verfahren zur Generierung von Konzept-Hierarchien prototypisch implementiert. Es werden gegebene Terme in Beziehung zueinander gesetzt. Eine möglich Methode hierzu ist die Subsumption, wobei Term A einen Term B subsumiert, falls in (fast) allen Dokumenten in denen Term B auftaucht ebenfalls der Term A erscheint [Sanderson and Croft, 1999].

Die Daten sowie die in eine Hierarchie zu bringenden Terme werden vom Lehrstuhl WBI[7] zur Verfügung gestellt. Zur Implementation wird Java[8] verwendet.

[7] Wissensmanagement in der Bioinformatik, www.informatik.hu-berlin.de/forschung/gebiete/wbi
[8] java.sun.com

Literatur

[GeneOntologyConsortium, 2006] GeneOntologyConsortium (2006). The gene ontology (go) project in 2006. *Nucleic Acids Research*, 34(Database-Issue):322–326.

[Grossman and Frieder, 2004] Grossman, D. A. and Frieder, O. (2004). *Information Retrieval Algorithms and Heuristics*. Springer.

[Groth et al., 2007] Groth, P., Pavlova, N., Kalev, I., Tonov, S., Georgiev, G., Pohlenz, H.-D., and Weiss, B. (2007). Phenomicdb: a new cross-species genotype/phenotype resource. *Nucleic Acids Research*, 35(Database-Issue):696–699.

[Maglott et al., 2005] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res*, 33(Database issue).

[Sanderson and Croft, 1999] Sanderson, M. and Croft, B. (1999). Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213, New York, NY, USA. ACM Press.

[Smith et al., 2005] Smith, C. L., Goldsmith, C. A., and Eppig, J. T. (2005). The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1).

[Staab and Studer, 2004] Staab, S. and Studer, R. (2004). *Handbook on Ontologies (International Handbooks on Information Systems)*. International Handbooks on Information Systems. Springer.

[Wheeler et al., 2007] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2007). Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 35(Database issue).