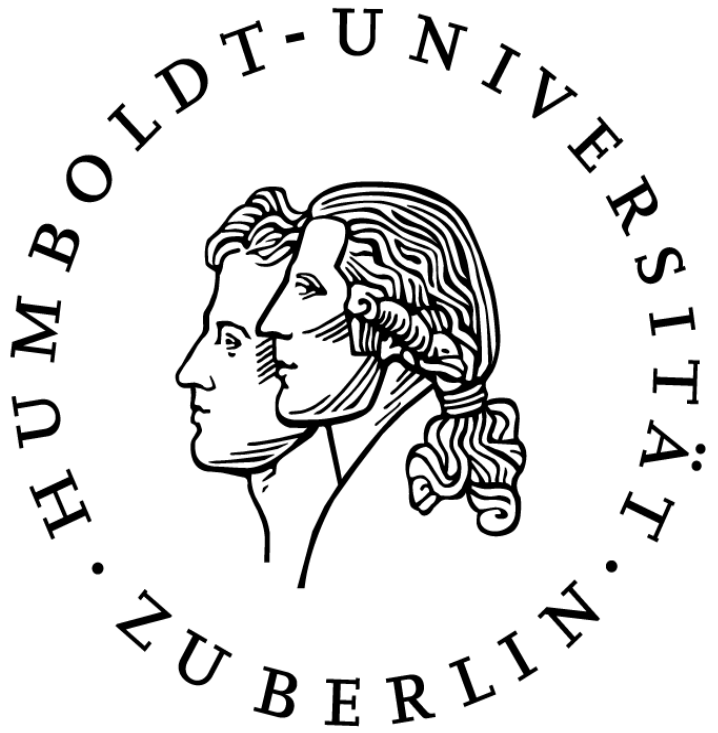

Exposé zur Studienarbeit

Klassifikationsverfahren zur systematischen Fremdschlüsselbestimmung aus Inklusionsconstraints

5. Juni 2008

Autorin : Alexandra Rostin
Email : alros@gmx.de
Betreuer : Prof. Dr. Ulf Leser

1. Hintergrund

Die Integration von unstrukturierten biowissenschaftlichen Datenbanken stellt einen sehr ressourcenintensiven Prozess dar. Diese Datenbanken unterscheiden sich teilweise erheblich in ihrem Aufbau (Datenstruktur) und ihrem Schwerpunkt, z.B. durch Verwendung andersartiger Identifikatoren oder unterschiedlicher Beschreibungen für das gleiche Objekt[1]. Das Ziel der Integration ist es, diese heterogenen Datenbanken zu einer einheitlich strukturierten Datenbank zusammen zu führen. Dabei sollen Redundanzen und Inkonsistenten verhindert werden. Diese können zum Beispiel dadurch entstehen, dass ein Objekt, das in verschiedenen Datenbanken mit von einander abweichenden Beschreibungen vorkommt, nicht als das gleiche erkannt wird und somit im Endergebnis der Integration als unterschiedliche Objekte mehrfach aufgeführt werden würde. Um einen Zusammenhang zwischen den Objekten verschiedener Datenbanken herstellen zu können, muss die Struktur der Daten, d.h. (semantische) Beziehungen zwischen den einzelnen Schemata bzw. Relationen einer Datenbank erkannt werden. Möglicherweise könnte ein solcher Zusammenhang auch eine noch unbekannte Beziehung zwischen den Objekten sein. Somit kann Integration zu neuen Erkenntnisse über den Aufbau bzw. der Struktur biowissenschaftlicher Objekte führen.

Eine semantische Beziehung zwischen Objekten wird in einer (relationalen) Datenbank durch eine Fremdschlüsselrelation wiedergegeben. Somit erlaubt das Auffinden von Fremdschlüsseln Rückschlüsse auf die Struktur einer Datenbank. Des weiteren stellt jede Fremdschlüsselbeziehung auch eine Inklusionsbeziehung dar. Ein Vorgehen zum Finden von FK-PK Beziehungen ist daher der folgende Weg: Zuerst sucht man alle Inklusionsbeziehungen. In einem zweiten Schritt muß dann „geraten“ werden, welche von diesen auch eine Fremdschlüsselbeziehung ist.

Dieser Ansatz wird im Projekt ALADIN[2] (ALmost Automatic Data INtegration)[3] verfolgt [2]. Ziel dieses Projektes ist es, den Integrationsprozess für biomedizinische Datenbanken so weit wie möglich zu automatisieren. Deshalb wurde zuerst ein Verfahren entwickelt, das alle Inklusionsabhängigkeiten einer Datenbank findet [3,4]. In der der Studienarbeit „Filtern von Fremdschlüsseln aus Inklusionsbeziehungen“ [1] wurde dann untersucht, wie man von Inklusionsabhängigkeiten auf Fremdschlüsselbeziehungen schließen kann. Als erste Annäherung zu diesem Thema wurden Heuristiken ermittelt, die geeignet waren, Fremdschlüssel zu erraten. Eine Heuristik repräsentiert jeweils ein Merkmal eines Fremdschlüssels, d.h. typische Eigenschaften der Werte eines Schlüssel – Fremdschlüssel – Paares, z.B. Abdeckung der Werte des einen Attributes im Wertebereich des anderen Attributes. Es wurden fünf Test-Datenbanken verwendet, bei denen die Struktur bekannt war und somit auch die vorhandenen Fremdschlüssel. So konnte überprüft werden, ob alle Fremdschlüssel richtig erraten wurden. Die Ergebnisse waren viel versprechend, wobei bemängelt werden kann, dass die Evaluierung auf Trainingsbeispielen erfolgte und somit die Gefahr von Overfitting besteht. Weiterhin wurde die Gewichtung der Fremdschlüssel-Merkmale noch rein manuell vorgenommen, was sehr inflexibel gegenüber möglichen neuen Heuristiken ist..

2. Zielstellung

Ziel dieser Studienarbeit ist es, diesen Ansatz zu systematisieren. Unter Berücksichtigung der ermittelten Heuristiken soll ein Klassifikationsverfahren [5] angewandt werden um damit die

[1] Objekte einer biowissenschaftlichen Datenbank sind zum Beispiel Gene, Proteine oder Sequenzen.
[2] http://zope.informatik.hu-berlin.de/forschung/gebiete/wbi/research/projects/aladin/
[3] http://www.informatik.hu-berlin.de/forschung/gebiete/wbi

Erkennung der Fremdschlüssel-Merkmale zu automatisieren. Um Overfitting zu vermeiden, soll bei der Evaluierung das Cross-Validierungsverfahren zum Einsatz kommen. Weiterhin soll versucht werden, weitere Indizien für Fremdschlüssel zu bestimmen und in die Klassifikation einzubeziehen.

3. Herangehensweise

Die Hauptaufgabe der Studienarbeit besteht darin, ein geeignetes Klassifikationsverfahren auszuwählen und zu evaluieren. Es existieren bereit fertige Machine Learning-Tools, die solche Aufgaben unterstützen. Eins davon ist WEKA (Waikato Environment for Knowledge Analysis[4]). Dabei handelt es sich um ein Open Source- Machine Learning-Tool der Waikato-Universität, dass in Java implementiert wurde.

Im Einzelnen sind folgende Aufgaben zu erfüllen:

1. ***Datenaufbereitung***
   Die Daten müssen zunächst für die in WEKA verfügbaren Klassifikationsalgorithmen aufbereitet werden. Dazu müssen die Daten in Form einer Relation in einer einzelnen Textdatei (flat file) vorliegen. In dieser sind die einzelnen Instanzen mit ihren jeweiligen Attributwerten aufgelistet. WEKA unterstützt auch den Zugriff auf SQL-Datenbanken. Außerdem muss die Trainingsmenge bestimmt werden.
2. ***Wahl eines geeigneten Klassifikationsverfahren***
   Aus den ermittelten Heuristiken müssen geeignete Maßzahlen abgeleitet und als Feature dargestellt werden. Mit Hilfe von WEKA sollen dann verschiedene Verfahren ausprobiert werden. In der vorangegangenen Studienarbeit[1][5] zeichnete es sich ab, dass Entscheidungsbäume gute Resultate liefern könnten. Das gilt es zu überprüfen.
3. ***Integration des Klassifikators***
   Nach der Entscheidung für ein Klassifikationsverfahren muss dieses im vorhandenen Projektrahmen integriert werden.
4. ***Evaluation***
   a. Es soll Cross-Validierung zur Evaluation verwendet werden. Dafür werden die vorhanden Beispiel-Datenbanken in Partitionen aufgeteilt. Die Einteilung in Test- und Trainingsmenge erfolgt so, dass jede Partition einmal die Testmenge ist, auf der evaluiert wird und alle übrigen die Trainingsmenge bilden.
   b. Die Evaluation wird einmal nach Fremdschlüsseln (Micro-Standard) und einmal nach Datenbank (Makro-Standard) erfolgen.
   c. Eventuell soll die Testmenge um weitere Datenbanken ergänzt werden.
5. ***Ermittlung weiterer Fremdschlüsseleigenschaften***
   Es soll versucht werden, noch weitere Eigenschaften von Fremdschlüsseln zu ermitteln.

[4] http://www.cs.waikato.ac.nz/~ml/weka/
[5] unter Punkt 3.3.3. Kombination der Heuristiken - 3.3.3.3. Alternative Methoden

4. Quellen

[1] O. Albrecht, „Filtern von Fremdschlüsseln aus Inclusionsbeziehungen“, 2007

[2] U.Leser, F.Naumann, „(Almost) Hands-off Information Integration for the Live Sciences“

[3] J. Bauckmann, U.Leser, F.Naumann, V.Tietz. „Efficiently Detecting InclusionDependencies“,International Conference on Data Engineering (ICDE 2007),Istanbul, Turkey

[4] J. Bauckmann. „Automatically Integrating Life Science Data Sources.“ VLDB2007 PhD Workshop, Vienna, Austria.

[5] Tom M. Mitchell [1997] “Machine Learning”, McGraw-Hill International Editions, Computer Sciences Series