PG-Seminar

MIDOS/ KEPLER

Ein Algorithmus zur multi-relationalen Entdeckung von Subgruppen

Referent: Daniel Wiese

PG-Seminar
Am Fachbereich Informatik
Der Universität Dortmund

Montag, den 14. November 2001

Betreuer:

Prof. Dr. Katharina Morik
Stefan Haustein

Inhaltsverzeichnis

1 Überblick

Diese Ausarbeitung im Rahmen der PG 402 der Universität Dortmund beschäftigt sich mit dem MIDOS Algorithmus von Stefan Wrobel, einem Algorithmus zur Subgruppenentdekkung. Die vorliegende Ausarbeitung ist folgender maßen gegliedert:

Zunächst wird das MIDOS Verfahren in den Kontext der Referate der anderen PG Telnehmer eingeordnet.

Um einen Überblick über den praktischen Einsatz der Subgruppenentdeckung zu bekommen, wird das Data-Mining-Werkzeug Kepler vorgestellt, welches das MIDOS-Verfahren beherrscht und auf Datenbanken anwenden kann.

Als nächstes wird das Problem der Subgruppenentdeckung genauer definiert. Hierzu gehört auch die genaue Betrachtung einer von Stefan Wrobel definierten multirelationalen Hypothesensprache.

Um die Qualität der Subgruppen beurteilen zu können, benötigen wir eine Evaluationsfunktion. Diese wird in dieser Ausarbeitung in einem eigenen Abschnitt behandelt.

Bevor zum Schluss der MIDOS Algorithmus selbst vorgestellt wird, betrachtet diese Ausarbeitung die möglichen Suchstrategien im Hypothesenraum. Vorgestellt werden hier verschiedene Möglichkeiten, den Hypothesenraum zu ordnen und zu beschneiden, so dass der Suchaufwand selbst verringert werden kann.

2 Einleitung

Data Mining oder auch Wissens Entdeckung in Datenbanken wird dazu verwendet, neues und interessantes Wissen aus großen Datenbanken zu extrahieren. Ein Verfahren welches im Rahmen von KDD eingesetzt werden kann, ist die Subgruppenentdeckung. Die Subgruppenentdeckung ist ein deskriptives Lernverfahren.

Hier ist man daran interessiert, durch Hypothesen beschriebene Teilbereiche des Instanzenraums zu identifizieren, über die lokal interessante Aussagen gemacht werden können.[1]

Bei dem hier vorgestellten MIDOS Verfahren von WROBEL wird die Subgruppenerkennung multirelational, d.h. über mehrere Relationen hinweg, durchgeführt.

Das MIDOS Verfahren basiert auf dem System EXPLORA , welches 1996 von W. Klösgen vorgestellt wurde. Das EXPLORA System, welches nur ein Ein-Relationales Problem behandeln konnte, wurde hier um eine multi-relationale Funktionalität erweitert.

Nach KLOESGEN gehört das statistische Auffinden von Subgruppen zu den am meisten populären und einfachsten Formen des Wissens.[2]

---

[1] WROBEL/MORIK/JOACHIMS
[2] KLOESGEN

Beispiele von interessanten Subgruppen sind folgende:

- *Die Arbeitslosenrate ist überproportional hoch bei jungen Männern mit niedrigen Ausbildungsgrad*

- *Die Todesrate bei Lungenkrebs ist bei Frauen signifikant in den letzten 10 Jahren gestiegen*

- *Junge arme Frauen sind stärker mit AIDS infiziert als ihre männliche Vergleichsgruppe*

3 Einordnung des Verfahrens

Die PG402 der Universität Dortmund beschäftigt sich mit den Bereich des Wissensmanagements. Demzufolge lässt sich die zu bearbeitende Domäne grob in folgende Teilbereiche gliedern:

- Wie lässt sich Wissen repräsentieren und verwalten?
- Welche Verfahren zur Wissensentdeckung können im Rahmen dieser PG eingesetzt werden?
- Welche Verfahren können zur Wissensextraktion verwendet werden?[3]

3.1 Wie lässt sich Wissen repräsentieren und verwalten?

Beispielsweise haben sich STEWART, DOUBLEDAY, BORGHOFF etc. mit der Frage „Wie Wissen Repräsentiert werden kann“ beschäftigt. Eine Möglichkeit das Wissen zu strukturieren sind Ontologien.

Ontologien[4] sorgen für eine konsistente Repräsentation relevanter Bereiche des Wissens. Mit diesen wird uns die Spezifikation einer Konzeption ermöglicht. Zur Beschreibung von Ontologien gehören Verfahren wie RDF(S) und DAML+OIL[5]

3.2 Welche Verfahren zur Wissensentdeckung können im Rahmen dieser PG eingesetzt werden?

Im Rahmen der Seminarphase wurden außer dem MIDOS-Verfahren folgende Lernverfahren vorgestellt:

- Data Cubes
- Der Apriori-Algorithmus
- Zeitaspekte
- RDT/DB

[3] Wird im Rahmen dieser Ausarbeitung nicht behandelt, da der Bezug zu MIDOS nicht gegeben ist
[4] GENESERETH/NILSON/1987
[5] CHAMPIN u. W3C u. HARMELEN

Data Cubes[6] verfolgt das Ziel eine Datenmenge als einen n-dimensionalen Raum darzustellen. Die Dimensionsreduktion wird hier durch Aggregation entlang der weggelassenen Dimension durchgeführt.

Der Apriori-Algorithmus[7] entdeckt Assoziationsregeln der Form X =>Y. Wie beim MIDOS-Verfahren wird hier ein minimaler Support (minimale Menge der Transaktionen) gefordert. Für den Konfidenzwert (Konfidenz misst die Gültigkeit der Assoziationsregel selbst) existiert bei MIDOS keine Äquivalenz. Allerdings verwendet MIDOS eine Evaluationsfunktion, die eine ähnliche Aufgabe hat wie der Konfidenzwert bei Apriori.
So wie das MIDOS Verfahren nutzt Apriori eine Ordnung über dem Hypothesenraum (Raum aller Assoziationsregeln) aus.

Unter dem Begriff Zeitaspekte sind verschiedene Algorithmen zur univariaten und multivarianten Zeitreihenanalyse zusammengefasst. Das MIDOS Verfahren kann hier im Rahmen eines Preprozessing durchgeführt werden, um die Datenmenge über den die Zeitreihenanalyse durchgeführt wird einzuschränken.

RDT/DB[8] ist ein System welches Regeln- bzw. Begriffslernen direkt auf Relationalen Datenbanken durchführen kann. Wie bei MIDOS ist hier der mögliche Hypothesenraum eingeschränkt und nach Allgemeinheit partiell geordnet.
Im Gegensatz zum RDT/DB, wird bei der Subgruppenerkennung nicht versucht eine Hypothese zu finden, welche die Daten global beschreibt und uns eine Vorhersage ungesehener Instanzen ermöglicht, sondern man ist nur einzig und allein an einer interessanten Teilmenge der gegebenen Daten interessiert.
Das MIDOS Verfahren selbst findet eine Teilmenge der gegebenen Daten. Diese Fokussierung auf eine Teilmenge von Daten unterscheidet sich von der Aufgabe eine Funktion zu Approximieren.

4 Visualisierungstools und Kepler

Kepler ist ein erweiterbares Softwaresystem für die Datenanalyse. Die Erweiterbarkeit von Kepler ermöglicht den Einsatz von neuen Verfahren, da diese Verfahren als ein Plug-in integriert werden können.

Das MIDOS Verfahren wird als ein solches Plug-in im Rahmen der Analysesoftware Kepler verwendet.

Kepler selbst, stellt ein Interface, eine graphische Benutzeroberfläche zu vielen Datenquellen sowie eine Sammlung von vielen gängigen Algorithmen bereit.

Das MIDOS-Verfahren selbst wird über folgendes Fenster konfiguriert (s. Abbildung 1).

[6] GRAY/HARINARAYAN
[7] AGRAVAL/IMIELINSKI/SWAMI
[8] MORIK/BROCKHAUSEN

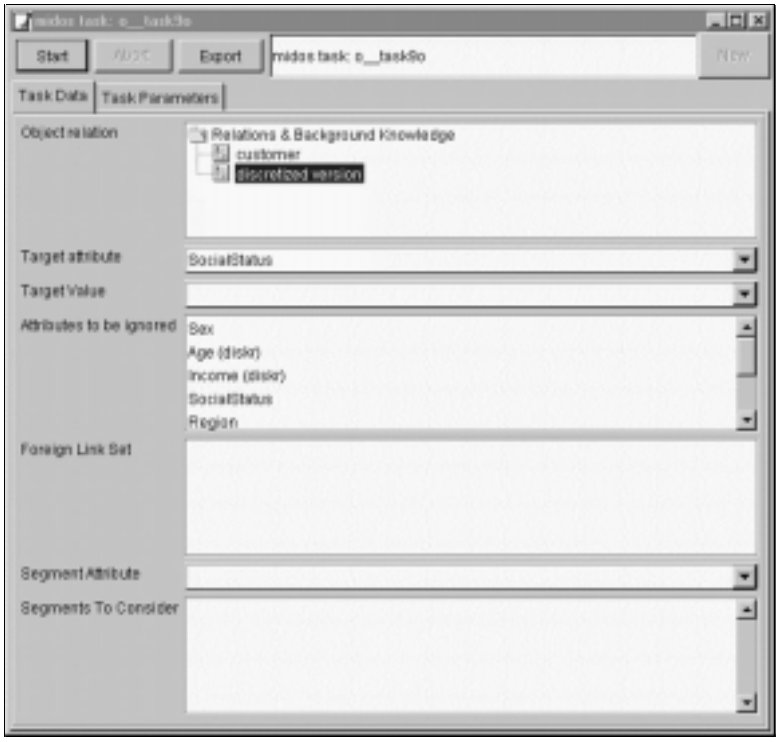

Abbildung 1

Wie in der Abbildung 1 zu sehen ist, wird im ersten Feld die Objektrelation für die zu analysierende Relation bestimmt.

Im Feld *Target attribute* wird das Zielattribut festgelegt. Die Werteverteilung dieses Zielattributes in der Subgruppe wird mit der Verteilung in der gesamten Population verglichen. Eine Subgruppe wird als interessant eingeschätzt wenn die Verteilung von der Gesamtverteilung abweicht.

Im Feld *Foreign Link Set* wird die Fremdschlüsselmenge eingetragen. Hier wird das Hintergrundwissen angegeben. Durch die Fremdschlüssel werden zusammenhängende Relationen bestimmt.

Das Ergebnis der Subgruppenerkennung wird vom Kepler in der folgenden Form präsentiert (s. Abbildung 2).

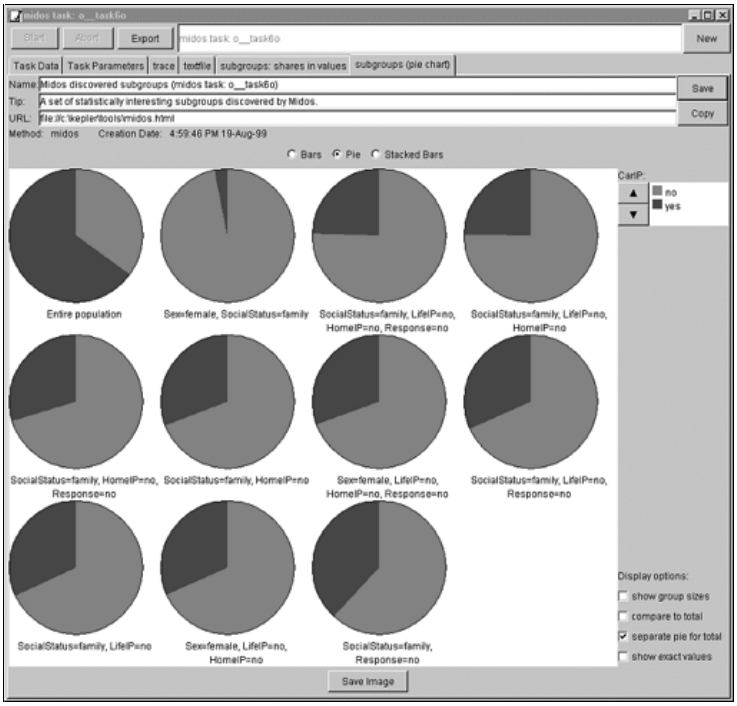

Abbildung 2

Das erste Kuchendiagramm stellt die Verteilung der Gesamtpopulation in der Objektrelation dar. Die anderen 10 Kuchendiagramme sind die gefundenen Subgruppen. Die Subgruppen wurden als interessant ausgewählt, da hier die Verteilung bezüglich des binären Attributes CarIP von der Gesamtpopulation abweicht. Ein Beispiel für eine gefundene Subgruppe ist *(Kuchendiagramm 1) Frauen mit einer Familie haben selten ein Auto angemeldet*.

5 Definition des Problems

5.1 Welche Hypothesensprache kann verwendet werden?

Üblicherweise können Subgruppen in einer einzelnen Relation einer Datenbank durch eine Konjunktion von Einschränkungen auf den Attributen beschrieben werden.[9]

WROBEL hat die im EXPLORA System verwendete Hypothesensprache für den multirelationalen Fall folgender maßen erweitert. Mit der hier im folgenden vorgestellten Hypothesensprache ist es möglich, weitere Relationen durch einen Join zu verbinden.

Der Grund, warum eine multirelationale Hypothesensprache benötigt wird, lässt sich gut am folgenden Beispiel illustrieren:

Die Daten eines Krankenhauses seien in einer relationalen Datenbank mit 7 Relationen gespeichert.

R1: Krankenhäuser und Abteilungen [300 Tupel]
R2: Patienten [6.000Tupel]
R3: Diagnosen [25.000 Tupel]

[9] Kloesgen/96a

R4: Therapien [43.000Tupel] usw..

Etwas formaler lassen sich die Relationen folgender maßen ausdrücken:

patient(PatientID,Name,Age,Sex, ...)
diagnose(PatientID,DiagnosisID,Date,HospitalID)
therapie(PatientID,TherapyID,Dosage,Date,HospitalID)
krankenhaus(HospitalID,Name,Location,Size,Owner,Class)

Eine mögliche entdeckte multirelationale Subgruppe kann dann lauten: *Patienten älter als 65 Jahre, die in einem kleinen Krankenhaus behandelt werden, haben eine überdurchschnittlich hohe Mortalitätsrate.* Diese Subgruppe kann mit dem folgenden Ausdruck beschrieben werden:

patient(ID,N,A,S) & A > 65 & diagnose(ID,D_ID,Dt,H) & krankenhaus(H,_,_,small,_,_).

Dieser Ausdruck erstreckt sich über die Relationen *patient, diagnose* und *krankenhaus*. Eine ein-relationale Hypothesensprache hätte nicht die Mächtigkeit diesen Ausdruck darzustellen.

5.2 Forderungen an die Hypothesensprache

Die von WROBEL definierte Hypothesensprache verwendet eine vorbestimmte Zielrelation. D. h. es werden nur Subgruppen von Objekten gebildet, welche die Zielrelation beinhalten *z.B. Subgruppen von Personen.* Weiterhin existieren genau spezifizierte Links zwischen Relationen. Um einer kombinatorischen Explosion vorzubeugen, hat WROBEL vorgeschlagen, einen *forein key* Link als zusätzliche Information zu verwenden. Als Konsequenz dürfen Links nur entlang vorher spezifizierten Pfaden gebildet werden (s. Abbildung 3).

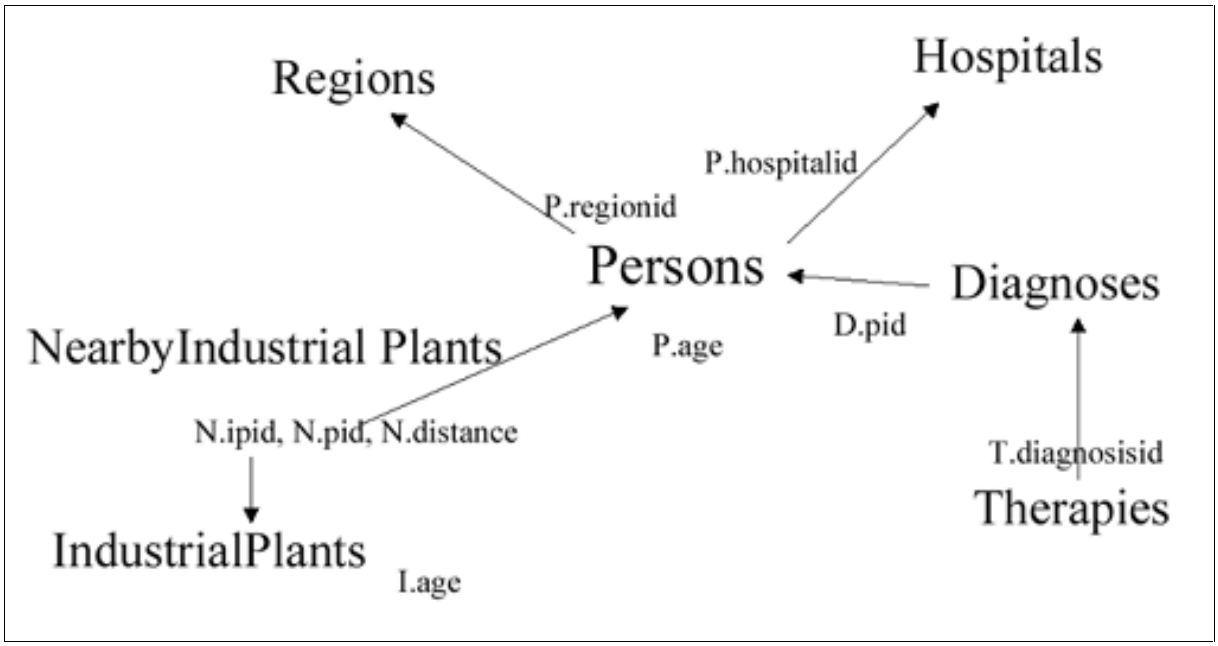

Abbildung 3

5.3 Definition der Hypothesensprache

Gegeben ist eine Datenbank D mit Relationen R und Fremdschlüssel-Links F. Die Hypothesensprache besteht aus einer Menge von verknüpften Prädikaten

$$C = A1 \,\&\, ... \,\&\, An$$

der folgenden Form:

$A_i = r_i\,(V_{i1}, ..., V_{ia})$ jede Relation wird durch ein korrespondierendes Prädikat beschrieben

$A_i = V_j\,[<|=|>]$ Prädikate mit booleschen Ausdrücken

Es gilt:
Die Hypothese startet immer von einer vorher bestimmten Objektrelation r_0. Zwei Prädikate teilen sich eine Variable nur dann, wenn ein korrespondierender Link (Fremdschlüssel) existiert.

Beispiel einer erlaubten Hypothese:

$$D=\{r_0, r_1, r_2, r_3, ...\},\ F = \{r_0[2]->r_1[1],\ r_0[3]->r_2[1],\ r_1[2]->r_3[1],\ r_2[2]->r_3[2]\}$$

$$r_0(X,Y,Z)\ \&\ r_1(Y,U)\ \&\ r_2(Z,R)\ \&\ r_3(U,R)\ \&\ (X = x_0)\ \&\ (R \geq medium)$$

Die Literale Y,R kommen in zwei Prädikaten vor, da zwischen diesen beiden Prädikaten auch tatsächlich ein Fremdschlüssellink existiert. Im Gegensatz hierzu sind folgende Hypothesen verboten:

- $r_0(X,Y,Z)\ \&\ r_3(U,R)$ *[nicht verlinkt]*
- $r_0(X,Y,Z)\ \&\ r_3(X,R)$ *[link nicht erlaubt/nicht definiert]*
- $r_1(Y,U)\ \&\ r_2(Z,R)\ \&\ r_3(U,R)$ *[r_1 ist nicht die Objektrelation]*

5.4 Definition des Subgruppenproblems

Damit kann eine Multi-relationale Subgruppe beschrieben werden als:

Gegeben:
eine relationale Datenbank D mit Relationen $R = \{r_1; :::::: ; r_m\}$
eine Sprache für Hypothesen L_H (um die Subgruppen zu beschreiben)
eine Qualitätsfunktion, um die Qualität der gefundenen Subgruppe zu beurteilen $d: h \in L_H$, $D \rightarrow [0..1]$
eine Integerzahl $k>0$

Finde:
eine Teilmenge $H \subseteq L_H$ von Hypothesen; mindestens k-viele ;
so, dass für jede gefundene Hypothese gilt: $h \in H$, $d(h, D) > 0$ (die Qualität ist größer null) und für jedes $h' \in L_H \setminus H$ gilt: $\min_{h \in H} d(h,D) > d(h',D)$ (die anderen Hypothesen h' sind schlechter)

6 Evaluationsfunktion

Eine gefundene Subgruppe muss durch ein geeignetes Verfahren als interessant eingeschätzt werden. Das MIDOS Verfahren schätzt eine Subgruppe als interessant ein, wenn diese ein abweichendes statistisches Verhalten aufweist.

Eine auffällige statistische Abweichung lässt sich besonders einfach bei einem binären Attribut illustrieren:

Angenommen das binäre Ziel-Attribut unserer spezifizierten Objektrelation sei der Behandlungserfolg (binär, ja | nein).

Die Wahrscheinlichkeitsverteilung in der gesamten Objektrelation (Referenzpopulation) bezüglich dieses Attributes sei [61%=ERFOLG und 31%=MISSERFOLG]

Eine Statistisch auffällige Subgruppe währe dann beispielsweise:
„Patienten älter als 65 Jahre, die ihre Erstdiagnose in einem kleinen Krankenhaus erhielten“
Wahrscheinlichkeitsverteilung [43%= ERFOLG 57%=MISSERFOLG]

Für das Aufstellen der Evaluationsfunktion welche von KLOESGEN aufgestellt und von WROBEL übernommen wurde wird folgendes benötigt:

- $(h) := \pi[K](\{\sigma \mid h\sigma \in D\})$, die Abdeckung von h
- $T := \{t \in r_O \mid t[A_g] = 1\}$, das Zielattribut (nur die wahren „Tupel“)
- $g(h) := |c(h)| / |r_0|$, Generalität (wie viele „Tupel“ werden im Verhältnis zur Gesamtmenge durch die Hypothese abgedeckt?)
- $p_0 := |T| / |r_O|$, die Referenzwahrscheinlichkeit der Zielattributes (binär) in D
- $p(h) := |c(h) \cap T| / |c(h)|$, Wahrscheinlichkeit des Zielattributes (binär) in $c(h)$

Somit lässt sich die Evaluationsfunktion aufstellen als:

$$d(h) := g(h) \ |(p(h) - p_0)|$$

7 Suchstrategien

Theoretisch müssten wir die Menge aller Hypothesen vollständig durchsuchen. Die erste naheliegende Möglichkeit zur Reduktion des Hypothesenraumes ist folgende. Man beginnt mit der allgemeinsten Hypothese und verfeinert diese, zu immer kleineren Subgruppen.

WROBEL[10] hat weiterhin folgende Möglichkeiten zur Reduktion des Hypothesenraumes vorgeschlagen:

[10] Wrobel/97a

7.1 Systematisches und geordnetes Durchsuchen des Hypothesenraumes

Das Ziel ist hier die Spezifikation eines Verfeinerungsoperators ρ. Dadurch ist es möglich, eine Partialordnung über dem Hypothesenraum zu definieren. Selbst bei einer Parallelisierung des Verfahrens kann durch die Partialordnung sichergestellt werden, dass jede Hypothese nur einmal generiert werden muss.

Der Verfeinerungsoperator wird folgendermaßen definiert:
Konstruiere eine Ordnung *o* im Verfeinerungsoperator ρ mit folgender Eigenschaft:

$$\rho: L_H \rightarrow 2^{\wedge} L_H$$

für alle $h' \in \rho(h)$ soll gelten $o(h') > o(h)$

„Alle Verfeinerten Hypothesen haben eine höhere Ordnung als die aktuelle Hypothese“

Der Spezialisierungsoperator liefert zu einer Hypothese *h* **alle** unmittelbar speziellen Nachfolger $\rho(h)$. Das generieren von Duplikaten kann jetzt dadurch vermieden werden, indem man sich einfach nur die aktuelle Hypothese und die Ordnung der Verfeinerung merkt.

7.2 Beschneiden des Hypothesenraumes

Wie bereits zu Beginn erwähnt, beginnt man mit einer generellen Hypothese und spezialisiert diese. Zum Beispiel ist „*Gebiet mit großen Arbeitslosigkeit und einer großen Anzahl von medizinischen Einrichtungen*“ spezieller als „*Gebiet mit großen Arbeitslosigkeit*“. Wie man sieht, kann die Anzahl der abgedeckten Tupel durch eine Spezialisierung nur kleiner werden, aber niemals größer. Wir beschneiden den Hypothesenraum genau dann, wenn die Anzahl der Tupel in unserer Subgruppe eine bestimmte Anzahl von Elementen unterschreitet.

Weiterhin lässt sich in dem Fall, dass wir die k-besten Hypothesen suchen, wie durch KLOESGEN gezeigt, eine optimistische Schätzfunktion d_{max} bestimmen.

Die k-besten Hypothesen seien in der Menge *H*. Wenn die Vorhersage für unsere aktuelle Hypothese h_0 schlechter ist, als die schlechteste bisher gefundene, können wir den Hypothesenraum ohne Gefahr beschneiden

$$d_{max}(h_c) < min_{h \text{ in } H}\, d(h)$$

Eine in KLOESGEN vorgestellte einfache optimistische Schätzfunktion lautet dann:

$$d(h) := g(h)\ (p(h) - p_0) \Rightarrow\ d_{max} := g(h)\ (1 - p_0)$$

7.3 Vorgehen bei der Verfeinerung

WROBEL hatte nicht nur eine Partialordnung auf dem Verfeinerungsoperator ρ definiert, sondern durch eine genaue Reihenfolge der Verfeinerungsoperationen sogar totale Ordnung erreicht.

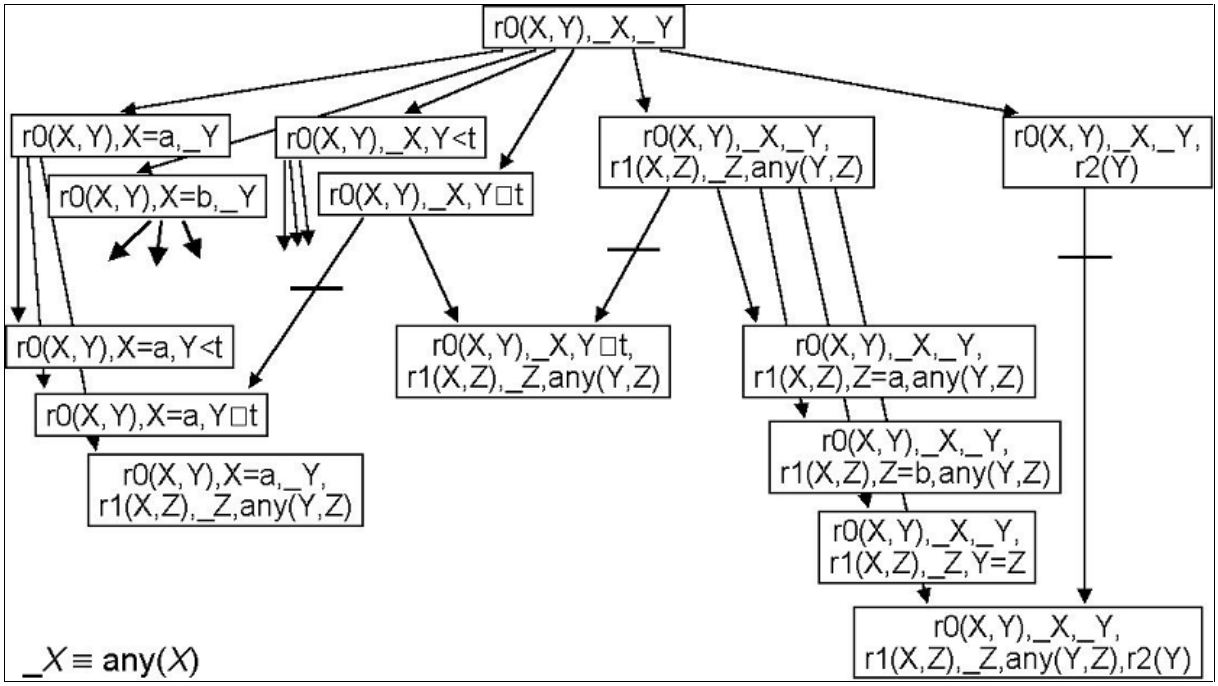

Abbildung 4

Die Reihenfolge, in der eine Hypothese verfeinert wird, ist in Abbildung 4 dargestellt. Auszugsweise wird in folgender Reihenfolge verfeinert:

- Verfeinere ein Literal *i* durch Spezialisierung
  - ersetze any(X) durch X=a oder X=b ... (abhängig von der Wertehierarchie)
  - ersetze any(X,Y) durch X=Y
  - usw..

- Verfeinere ein Literal *i* durch Verfolgen des Links. Füge dabei ein Literal nur mit Variablen (korrespondierend zur Zielrelation) hinzu. Z.B. wird in Abbildung 4, das Literal r1(X,Y) hinzufügt, da es einen Link von r0→r1 gibt.

Jetzt kann jede Hypothese der Ordnung *o* durch einen lexikographischen Vergleich total geordnet werden.

8 Der MIDOS Algorithmus

Im Groben arbeitet der MIDOS Algorithmus folgender maßen:

Zunächst nutzt der Algorithmus die Tatsache aus, dass die Hypothesen geordnet sind.

Die Suche von der generellen zur spezifischen Hypothese geschieht iterativ in 2. Phasen:

1. Generierung der verfeinerten Hypothesen
2. Evaluation der neuen Hypothesen

Im Anschluss hieran findet das Beschneiden des Hypothesenraumes statt. Das Beschneiden des Hypothesenraumes kann entweder durch das unterschreiten einer minimalen Größe der Subgruppe erfolgen (Eine speziellere Hypothese könnte niemals mehr Tupel abdecken als die vorhergehende) oder durch die Verwendung der Vorhersagefunktion. Unterschreitet die Vorhersage die schlechteste der bisher gefundenen Hypothesen, so wird der Hypothesenraum an dieser Stelle abgeschnitten.

Der Algorithmus selbst ist folgender:

Q:={V1, ..., Va(r0)}, ← start mit der gesamten Objektrelation (allgemeinste Hypothese);
H:=0 ← suche die k-besten Hypothesen

while nicht fertig
- wähle eine Teilmenge *C* aus *Q* gem. Suchstrategie
-$\rho(C):=\{\rho(h)|h \in C\}$ ← Menge aller unmittelbar spezielleren Hypothesen
-Teste jede Hypothese auf ihre Qualität (berechne *d(h)* mit $h \in \rho(C)$)
-**if** *d(h)=0* dann den Hypothesenraum abschneiden
-**else if** $d_{max}(h)$ schlechter ist als die bisher schlechteste Hypothese in *H*←Hypothesenraum abschneiden
-else
-**if** *d(h)* besser als die bisher schlechteste Hypothese→ schlechteste Hypothese entfernen
-füge *h* zu *H* hinzu
-füge *h* zu *Q* hinzu

9 Fazit

Wie in der Ausarbeitung gezeigt wurde, lässt dich das MIDOS-Verfahren auf multirelationale Datenbanken anwenden. Dieses hat den Vorteil, dass die zu untersuchende Teilmenge der Datenbank nicht im Rahmen eines Preprozessing zu einer einzigen Relation transformiert werden muss.

Der MIDOS-Algorithmus selbst ist ein typischer Top-Down Algorithmus, bei dem folgende Eigenschaften zu Effizienzsteigerung integriert wurden:

1. *Totale Ordnung auf dem Hypothesenraum*. Durch die totale Ordnung auf dem Hypothesenraum ist das MIDOS-Verfahren für eine Parallelisierung bestens geeignet.

2. *Beschneiden des Hypothesenraumes*. Bei einem unterschreiten einer minimalen Größe der Subgruppe und durch das Verwenden einer Vorhersagefunktion wird der Hypothesenraum abgeschnitten.

Meine persönliche Meinung ist, dass sich der Einsatz des MIDOS Verfahren im Rahmen der PG eignet. Teilprobleme der Aufgabenstellung wie z.B.: *finde besonders interessante Kundengruppen bezüglich einer Versicherung* lassen sich mit dem MIDOS Verfahren gut lösen.

Weiterer Vorteil im Hinblick auf die Zielgruppe der Wissensmanagementsoftware (Produktmanager eines Versicherungsunternehmens) ist, dass MIDOS intuitiv leicht zu interpretierbare Ergebnisse liefert.

10 Literaturverzeichnis

- [AGRAVAL/IMIELINSKI/SWAMI] Agraval, R., Imielinski, T., und Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, D.C.

- [CHAMPIN] Pierre-Antoine Champin. Einführung in die grundlegenden Aspekte von RDF. RDTF Tutorial. http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer2001-/SemanticWeb/papers/rdf-tutorial.pdf

- [GRAY] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichert, M. Venkatrao, F. Pellow, H. Pirahesh. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. Data Mining and Knowledge Discovery, 1(1): 29-53, 1997.

- [HARINARAYAN] V. Harinarayan, A. Rajaraman, J. Ullman. Implementing Data Cubes Efficiently. In Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, Seiten 205-216, Montreal, Kanada, Juni 1996.

- [HARMELEN] Frank van Harmelen, Peter F. Patel-Schneider, and Ian Horrocks. Reference Description of the DAML+OIL(March 2001) Ontology Markup Language. DAML+OIL Document, URL http://www.daml.org/2000/12/reference.html, Mar. 2001.

- [KLOESGEN] Willi Klösgen. Explora: A multipattern and multistrategy discovery assi-stant. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, Kapitel 3, s. 249–272. AAAI Press/The MIT Press.

- [MORIK/BROCKHAUSEN] Katharina Morik and Peter Brockhausen. A multistrategy approach to relational knowledge discovery in databases. In Ryszard S. Michalski and Janusz Wnek, editors, Proceedings of the Third International Workshop on Multistrategy Learning (MSL-96), Palo Alto, May 1996. AAAI Press. http://citeseer.nj.nec.com/morik96multistrategy.html

- [W3C] W3C. Resource Description Framework. http://www.w3.org/TR/rdf-schema

- [WROBEL/MORIK/JOACHIMS] Wrobel, S., Morik, K., Joachims, T. Maschinelles Lernen und Data Mining. In. Görz, Günther, Handbuch der künstlichen Intelligenz, s. 517-597, Oldenburg, 2000.
- [WROBEL] Stefan Wrobel. An algorithm for multi–relational discovery of subgroups. In J. Komorowski and J. Zytkow, editors, Principles of Data Minig and Knowledge Discovery: First European Symposium (PKDD 97), s. 78–87, Berlin, New York, 1997. Springer.

- [GENESERETH/NILSON] Genesereth, M. R., & Nilsson, N. J. (1987). Logical Foundations of Artificial Intelligence. San Mateo, CA: Morgan Kaufmann Publishers. Menlo Park, California, 1996.