communications in mathematical
match
and in computer chemistry

no. 34, October 1996

ISSN 0340-6253

MATCDY (34) 79-90 (1996)

KLASSIFIKATION DURCH GRAPHENTHEORETISCHE KONSTRUKTION VON GERÜSTEN IN LINIENDIAGRAMMEN VON BEGRIFFSVERBÄNDEN

Hans-Georg Bartel

Institut für Chemie der Humboldt-Universität zu Berlin
Hessische Straße 1-2, D-10115 Berlin

1. Die Problemstellung

Es ist in vorhergehenden Arbeiten (s. z.B. [1,2,3]) bereits auf den Wert und die Möglichkeiten hingewiesen worden, welche mit der Formalen Begriffsanalyse (FBA) der Darmstädter Schule von R. WILLE [4,5] und ihrer Anwendung auch im Rahmen der Chemie verbunden sind. Durch Erfassung und Konstruktion der formalbegrifflichen Struktur eines vorgegebenen Kontexts ist sie neben vielem anderen in der Lage, auf der Basis einer begrifflichen Hierarchie eine Klassifikation der im Kontext vorhandenen Gegenstände zu gewährleisten. Dabei berücksichtigt sie die in diesem vorhandene Information ohne Einschränkung und zwingt nicht, wie etwa die hierarchische Cluster-Analyse, den gegebenen Daten eine Hierarchie mit Baumstruktur (und damit im allgemeinen eine Einschränkung) auf, wie sie die Darstellung ihrer Resultaten in Form von Dendrogrammen zum Ausdruck bringt (vgl.[6,7]). Außerdem liefert die FBA im Gegensatz zur Cluster-Analyse (auch der nichthierarchischen) neben der extensionalen zugleich auch die entsprechende intensionale Klassifikation.

Datensätze der Naturwissenschaften, also auch der Chemie, speziell der Analytischen Chemie, der Materialkunde, der Umweltchemie und Archäochemie, führen dabei sehr oft zu einer sehr großen Menge von Begriffen und daher zu einer komplexen begrifflichen Struktur. Es wäre daher in solchen Fällen eine vertretbare Beschränkung der gesamten begrifflichen Hierarchie auf eine solche mit Baumstruktur wünschenswert, die sich mit einem möglichst geringen Informationsverlust aus der vollständigen Begriffsstruktur des formalen Kontexts, der aus dem gegebenen Datensatz abgeleitet wird, gewinnen läßt. Obwohl die Erfahrung lehrt, daß die Bildung disjunkter Klassen in realen Fällen des Abstraktionsprozesses nicht die Regel ist, bleibt es doch im Sinne der rationalen Wissensaneignung günstig, eine Zerlegung zu erhalten.

Die mit diesen Zielen verknüpfte mathematische Aufgabe besteht darin, in dem mit der vollständigen Begriffsstruktur eines Kontexts zusammenhängenden gerichteten Graphen einen Baum bzw. ein Gerüst im graphentheoretischen Sinne (s.u.) zu konstruieren. Hierbei ist möglichst viel Information aus dem Kontext und dem Begriffsverband zu übernehmen.

Den ersten Überlegungen zu diesem Problem ist dieser Aufsatz gewidmet.

2. Grundbegriffe der Formalen Begriffsanalyse

Da die Grundbegriffe der FBA in dieser Zeitschrift [1,2] ausführlich dargestellt und demonstriert wurden, ist hier eine Beschränkung auf wenige Definitionen erlaubt:

Definition 1: Ein Tripel (G, M, I) heißt (formaler) Kontext, wobei G Gegenstandsmenge und M Merkmalsmenge genannt werden und $I \subseteq G \times M$ eine binäre Relation ist. Die Aussage $(g, m) \in I$ wird gIm geschrieben und „der Gegenstand $g \in G$ hat das Merkmal $m \in M$" gelesen.

Definition 2: Für $A \subseteq G$ bezeichnet $A' = \{m \in M \mid \forall\ g \in A: gIm\}$ die Menge aller Merkmale, die alle Gegenstände aus A gemeinsam haben. $B' = \{g \in G \mid \forall\ m \in B: gIm\}$ besteht aus den Gegenständen, auf die alle Merkmale aus $B \subseteq M$ zutreffen.

Definition 3: Das Paar (A,B) heißt ein (formaler) Begriff im Kontext (G,M,I), wenn die Beziehungen $A \subseteq G$, $B \subseteq M$, $A' = B$ und $B' = A$ gelten. A wird der Umfang und B der Inhalt des Begriffs (A,B) genannt.

Definition 4: (A,B) ist Unterbegriff von (C,D) (symbolisch $(A,B) \leq (C,D)$), falls $A \subseteq C$ ($\Leftrightarrow D \subseteq B$) zutrifft.

Bezeichnet $\mathcal{B}(C)$ die Menge der Begriffe zum Kontext $C = (G, M, I)$, so ist die halbgeordnete Menge $(\mathcal{B}(C), \leq)$ mit der Unterbegriffsrelation $\leq$ (*Definition 4*) ein Verband (zur Definition s.[8]), genannt der Begriffsverband von C.

Definition 5: (A,B) heißt unterer Nachbar von (C,D), falls (A,B) Unterbegriff von (C,D) ist und es keinen Begriff (X,Y) in $\mathcal{B}(C)$ gibt, für welchen $(A,B) < (X,Y) < (C,D)$ gilt. Man schreibt dann $(A,B) \prec (C,D)$.

Wie jede halbgeordnete Menge ist ein Begriffsverband graphisch in Form eines Liniendiagramms darstellbar. Zur Erklärung der Lesung eines solchen Diagramms können z.B. die Aufsätze [1,4] verwendet werden.

3. Gerichtete Graphen

Da sich die folgende Darlegung auf die Darstellung des Begriffsverbandes als Liniendia-gramm und somit als gerichteten Graphen bezieht, mögen einige Termini der Graphentheorie erwähnt werden (vgl. hierzu insbesondere [9,16,17]).

Es seien X und U Mengen und Γ: U $\rightarrow$ X×X eine Abbildung, die jedem Element von U ein geordnetes Paar (ξ,η) aus X×X zuordnet. Dann heißt das Tripel Γ = (X,U,Γ) ein *gerichteter Graph* mit der Knotenmenge X und der Bogenmenge U. Wird $\mathbf{u}\in\mathbf{U}$ das geordnete Paar $(\xi,\eta)\in\mathbf{X}\times\mathbf{X}$ vermittels Γ zugeordnet, so wird auch $\Gamma(\mathbf{u}) = (\xi,\eta)$ geschrieben. Ein Bogen u wird *Schlinge* (am Knoten ξ) genannt, wenn $\Gamma(\mathbf{u}) = (\xi,\xi)$ gilt. Bögen u und v, für welche $\Gamma(\mathbf{u}) = \Gamma(\mathbf{v}) = (\xi,\eta)$ zutrifft, heißen *parallel*. Ist $\Gamma(\mathbf{u}) = (\xi,\eta)$ und $\Gamma(\mathbf{v}) = (\eta,\xi)$, so spricht man davon, daß u und v *Doppelbögen* sind. Im Bogen (α,ω) heißen die Knoten α und ω Endpunkte, α Startpunkt und ω Zielpunkt.

Alle gerichteten Graphen in dieser Arbeit sind *schlicht im engeren Sinne*, d.h., sie enthalten keine parallelen Bögen und keine Schlingen. Solche Graphen werden auch *Digraph* genannt. Die hier betrachteten Digraphen werden zusätzlich zu den genannten Eigenschaften auch keine Doppelbögen enthalten.

Jedes Liniendiagramm des Verbandes $(\mathcal{B}(\mathbf{C}), \leq)$ stellt folgenden gerichteten Graphen dar. Seine Knoten sind die Begriffe aus $\mathcal{B}(\mathbf{C})$. $(\boldsymbol{b}_i, \boldsymbol{b}_j)$ $(\boldsymbol{b}_i, \boldsymbol{b}_j \in \mathcal{B}(\mathbf{C}))$ ist ein Bogen dieses Graphen, wenn $\boldsymbol{b}_i \prec \boldsymbol{b}_j$ gilt. Für die graphische Darstellung von Liniendiagranmmen wird vereinbart, daß der Startknoten $\boldsymbol{b}_i$ eines Bogens $(\boldsymbol{b}_i, \boldsymbol{b}_j)$ in der Papierebene stets unterhalb von $\boldsymbol{b}_j$ gezeichnet wird, wobei dann aber die Kennzeichnung der Richtung durch eine Pfeilspitze entfällt.

Sei (X,U,Γ) ein Digraph. Eine Knoten-Bogenfolge $\alpha_o, \mathbf{u}_1, \alpha_1, \mathbf{u}_2, \alpha_2, \ldots, \alpha_{n-1}, \mathbf{u}_n, \alpha_n$ mit $n\in\mathbb{N}_o$, $\alpha_o,\ldots,\alpha_n\in\mathbf{X}$ und $\mathbf{u}_1,\ldots,\mathbf{u}_n\in\mathbf{U}$ heißt eine *Kette* von α_o nach α_n, wenn für $i = 1,\ldots,n$ α_{i-1}, α_i die Endknoten von $\mathbf{u}_i$ sind und für $i,j = 1,\ldots,n$ aus $i \neq j$ auch $\mathbf{u}_i \neq \mathbf{u}_j$ folgt. Eine Kette von α_o nach α_n heißt eine *Bahn* (von α_o nach α_n), falls $\Gamma(\mathbf{u}_i) = (\alpha_{i-1}, \alpha_i)$ für $i = 1,\ldots,n$ gilt. Zur Beschreibung einer Bahn $\alpha_o,\mathbf{u}_1,\alpha_1,\mathbf{u}_2,\alpha_2,\ldots,\alpha_{n-1},\mathbf{u}_n,\alpha_n$ genügt die Folge $\mathbf{u}_1,\ldots,\mathbf{u}_n$. Eine Kette von α_o nach α_n wird ein *Zyklus* genannt, falls $n \geq 2$ und $\alpha_o = \alpha_n$ gilt.

(X,U,Γ) heißt *zusamnmenhängend*, wenn für alle $\alpha,\omega\in\mathbf{X}$ eine Kette von α nach ω existiert. Ist der Graph (X,U,Γ) zusammenhängend und zyklenfrei, so wird er ein *Baum* genannt. $(\mathbf{X}_1,\mathbf{U}_1,\Gamma_1)$ heißt ein *Gerüst* von $(\mathbf{X}_2,\mathbf{U}_2,\Gamma_2)$, falls $(\mathbf{X}_1,\mathbf{U}_1,\Gamma_1)$ ein Baum ist und $\mathbf{X}_1 = \mathbf{X}_2$, $\mathbf{U}_1 \subseteq \mathbf{U}_2$ sowie $\Gamma_1 \subseteq \Gamma_2$ gelten.

4. Kürzeste Bahnen in einem bogenbewerteten Graphen

Wie oben dargelegt wurde, verlangt die hier betrachtete Aufgabe insbesondere, in einem gerichteten Graphen (Liniendiagramm des Begriffsverbandes) ein Gerüst zu konstruieren. Das gelingt sofort, wenn ein bogenbewerteter Digraph Γ = (X,U,Γ,ℓ) gegeben ist, d.h., wenn in einem Digraphen eine Abbildung ℓ existiert, die jedem Bogen $\mathbf{u}\in\mathbf{U}$ eine reelle Zahl $\ell(\mathbf{u})\in\mathbb{R}$ zuordnet. $\ell(\mathbf{u})$ heißt die Länge des Bogens u.

Denkt man sich einen solchen Graphen $\Gamma = (\mathbf{X},\mathbf{U},\mathbf{F},\ell)$ gegeben und wird vorausgesetzt, daß für alle $\mathbf{u}\in\mathbf{U}$ $\ell(\mathbf{u}) \geq \mathbf{0}$ gilt, so ermöglicht der im folgenden beschriebene Algorithmus, bezüglich eines Knotens $\alpha\in\mathbf{X}$ einen Baum $\Delta(\alpha) = (\mathbf{Y},\mathbf{V},\mathbf{t})$ zu konstruieren, in welchem $\mathbf{Y}$ die Menge der Knoten ist, welche von α ausgehend auf Bahnen erreicht werden können. $\mathbf{V} \subseteq \mathbf{U}$ enthält Bögen, die Elemente solcher Bahnen sind, die α und die Knoten in $\mathbf{Y}$ mit jeweils minimaler Länge verbinden. Unter der Länge einer Bahn $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ wird die Summe $\sum_{i=1}^{n} \ell(\mathbf{u}_i)$ verstanden. Die Potential von η genannte Zahl $\mathbf{t}(\eta)\in\mathbb{R}$ gibt dabei die Länge einer kürzesten Bahn zwischen den Knoten α und η in $\Delta(\alpha)$ an. Der genannte Lösungsalgorithmus geht auf G.B. DANTZIG [10] zurück (bezüglich des Beweises s. auch [9]).

Algorithmus nach DANTZIG

gegeben: $\Gamma = (\mathbf{X},\mathbf{U},\mathbf{F},\ell)$ mit $\forall\mathbf{u}\in\mathbf{U} : \ell(\mathbf{u}) \geq \mathbf{0}$, $\alpha\in\mathbf{X}$
($\mathbf{X}$, $\mathbf{U}$ endlich)
gesucht: Baum $\Delta(\alpha) = (\mathbf{Y},\mathbf{V},\mathbf{t})$ mit den Eigenschaften s. Text

① $\mathbf{Y} = \{\alpha\}$, $\mathbf{t}(\alpha) = \mathbf{0}$, $\mathbf{V} = \emptyset$

② bilde $\mathbf{H} = (\mathbf{Y} \times (\mathbf{X} \setminus \mathbf{Y})) \cap \mathbf{U}$

③ $\mathbf{H} = \emptyset$?

nein: ④ ermittle ein $\mathbf{h} = (\eta^*,\mu)\in\mathbf{H}$
mit minimalem $\mathbf{t}(\eta^*)+\ell(\mathbf{h})$
⑤ $\mathbf{Y} = \mathbf{Y} \cup \{\mu\}$, $\mathbf{V} = \mathbf{V} \cup \mathbf{h}$, $\mathbf{t}(\mu) = \mathbf{t}(\eta^*) + \ell(\mathbf{h})$
⑥ → ②

ja: ⑦ $\Delta(\alpha) = (\mathbf{Y},\mathbf{V},\mathbf{t})$, Ende

5. Auffälligkeit von Merkmalen

Um den Algorithmus von DANTZIG auf den durch das Liniendiagramm eines Begriffsverbandes gegebenen gerichteten Graphen anwenden zu können, müssen dessen Bögen durch eine Abbildung ℓ bewertet werden. Zur Lösung dieses Problems wird der folgende Weg beschritten:

In der mathematischen Kunsttheorie wird der „Überraschungswert" oder die „Auffälligkeit" **A** eines Merkmals mit der relativen Häufigkeit p durch die Formel

$$\mathbf{A}(p) = -\,p\cdot\ln(p)$$

definiert [11]. Das Maximum dieser Funktion liegt bei $p_{max} = \frac{1}{e} \approx 0{,}37$, so daß die normierte prozentuale Auffälligkeit sich zu

$$\mathbf{A}(p) = -100\cdot e\cdot\, p\cdot\ln(p)\ [\%]$$

ergibt. Von den vielen Beispielen, die die Kunsttheorie (s. [11]) liefert, um diese Formel bezüglich ihrer Anwendbarkeit zu rechtfertigen, sei das Gemälde von ANSELM FEUERBACH (1829–1880) mit der Darstellung der IPHIGENIE im Hessischen Landesmuseum Darmstadt (Inv.-Nr. GK 488) [12] genannt. Die die Reinheit der mykenischen Fürstentochter symbolisierende Farbe Weiß (einschließlich des hellen Inkarnats) hat einen relativen Flächenanteil von $\frac{212}{595} \approx 0{,}36$ und liegt damit also tatsächlich in der Nähe des Auffälligkeitsmaximums.

Dieses Maß für die Auffälligkeit eines Merkmals soll noch mit zwei Begriffen aus der Petrographie kombiniert werden (s. [13]). Dort wird bezüglich der in Gesteinsproben auftretenden Mineralarten von deren Extensität (Verbreitungsweise) und Intensität gesprochen. Die erstere beschreibt den prozentualen Anteil der Proben, in denen die gegebene Mineralart auftritt, während die letztere ihren prozentualen Anteil in jeder einzelnen Gesteinsprobe angibt. In dem speziellen Fall der Gesteinskunde fungieren die Mineralarten als Merkmale und die Proben als Gegenstände. Bei einer Verallgemeinerung kann daher bei Bezugnahme auf einen Kontext (G,M,I) folgendermaßen definiert werden:

- Als relative Extensität des Merkmals $m\in M$ wird die Zahl $v(m) = |m'| : |G|$ bezeichnet, wobei $m' = \{m\}' = \{g\in G \mid gIm\}$ (Menge aller Gegenstände, bei denen das Merkmal m auftritt) ist.
- Als relative Intensität des Merkmals $m\in M$ bezüglich des Gegenstandes $g\in G$ wird das Verhältnis $i(m, g) = z(m, g) : |g'|$ verstanden, wobei $g' = \{g\}' = \{n\in M \mid gIn\}$ (Menge aller Merkmale, die der Gegenstand g besitzt), $z(m, g) = 1$, wenn $(g, m)\in I$, und $z(m, g) = 0$, wenn $(g, m)\notin I$, bedeuten. (Für diese Definition wird vorausgesetzt, daß im Kontext kein $g\in G$ mit $g' = \emptyset$ existiert.)
- Die Summe $j(m) = \frac{1}{|G|}\cdot\sum_{g\in G} i(m,g)$ wird als mittlere Intensität des Merkmals $m\in M$ bezeichnet.

Entsprechend sind die Auffälligkeiten zu definieren:

- extensive Auffälligkeit $\mathbf{A}_e(m) = -e\cdot v(m)\cdot\ln(v(m))$
- intensive Auffälligkeit $\mathbf{A}_i(m) = -e\cdot j(m)\cdot\ln(j(m))$
- Gesamtauffälligkeit $\mathbf{A}_g(m) = \frac{1}{2}(\mathbf{A}_e(m) + \mathbf{A}_i(m))$

In dieser Arbeit wird die Länge $\ell_{ij} = \ell(\boldsymbol{b}_i, \boldsymbol{b}_j)$ im Liniendiagramm zwischen $\boldsymbol{b}_i$ und seinem oberen Nachbarn $\boldsymbol{b}_j$ ($\boldsymbol{b}_i \prec \boldsymbol{b}_j$, $\boldsymbol{b}_i, \boldsymbol{b}_j \in \mathcal{B}(G,M,I)$) versuchsweise folgendermaßen festgelegt, wobei für spätere Überlegungen und Untersuchungen genügend Raum für Variationen bleibt: Es sei $\boldsymbol{b}_i = (O_i, Y_i)$ und $\boldsymbol{b}_j = (O_j, Y_j)$, so daß $O_i \subseteq O_j \subseteq G$ und $Y_j \subseteq Y_i \subseteq M$ gilt, und $Y_{ij} = Y_i \setminus Y_j$ die Menge der Merkmale bedeutet, die beim Übergang vom Unter- zum Oberbegriff wegfallen. Dann wird ℓ_{ij} definiert durch

$$\ell_{ij} = \frac{1}{|Y_{ij}|} \cdot \sum_{y \in Y_{ij}} \left(1 - \mathbf{A}_\iota(y)\right) \quad (\iota = e, i, g).$$

Hierbei ist $1-\mathbf{A}_\iota(y)$ die entsprechende Unauffälligkeit des Merkmals y. ℓ_{ij} ist somit eine mittlere Unauffälligkeit, die sich auf die Merkmale bezieht, welche im Inhalt des Unterbegriffs $\boldsymbol{b}_i$ zusätzlich zu denen im Inhalt des Oberbegriffs $\boldsymbol{b}_j$ vorhanden sind.

Da stets $\ell_{ij} \geq 0$ gilt, kann der Algorithmus von DANTZIG angewendet werden. Wird als Ausgangsknoten das Nullelement des Liniendiagramms (der „unterste" Begriff) gewählt, so liefert der Algorithmus ein Gerüst, da dieser Begriff Unterbegriff zu allen anderen Begriffen ist, so daß zu ihnen allen im Liniendiagramm vom Nullelement ausgehende Bahnen führen. Da ℓ_{ij} die (gemittelte) Unauffälligkeit derjenigen Merkmale darstellt, welche beim Fortschreiten im Liniendiagramm beim Übergang von einem Begriff zu dessen oberen Nachbarn (unmittelbaren Oberbegriffen) im Inhalt wegfallen bzw. in umgekehrter Richtung hinzukommen, sind die gesuchten bzw. konstruierten Bahnen somit solche minimaler Unauffälligkeit. Minimale Unauffälligkeit bedeutet aber maximale Auffälligkeit. Eine Bewegung auf einem Zweig des konstruierten Baumes ist also mit dem geringsten Verlust an Unauffälligkeit bzw. dem größten Gewinn an Auffälligkeit in Hinblick auf die durch sie bedingte Inhaltsänderung verbunden. Insofern entspricht die Zusammenfassung solcher Gegenstände zu einer Klasse, die zu Gegenstandsbegriffen eines bestimmten Zweiges des Baumes gehören, einer Klassifikation hinsichtlich kleinster inhaltlicher Unauffälligkeits- bzw. größter inhaltlicher Auffälligkeitsunterschiede.

Es muß noch erwähnt werden, daß das äußerlich viel einfachere Längenmaß $\ell_{ij} = |Y_{ij}|$ deshalb von Nachteil ist, weil es bei der Konstruktion des Gerüstes zu einer sehr großen Mehrdeutigkeit führt. Durch die Existenz sehr vieler gleichlanger Bögen würden sich sehr viele kürzeste Bahnen und damit verschiedene Gerüste ergeben.

6. Ein Demonstrationsbeispiel

Für die erste Darlegung des Problems, die mit diesem Aufsatz beabsichtigt ist, möge die Beschränkung auf die Verwendung der extensiven Unauffälligkeit der Merkmale $m \in M$ eines Kontexts (G,M,I) genügen, d.h. auf $1-\mathbf{A}_e(m)$. Als De-

monstrationsbeispiel wird ein kleiner Kontext ORGMOL gewählt, der schon in [1] (dort unter der Bezeichnung BEISPIEL2) verwendet wurde (Tabelle 1). Die Gegenstandsmenge enthält $|G| = 15$ organochemische Verbindungen, welche als Merkmale ein bzw. zwei Phenylringe (1Ph, 2Ph), ein bzw. zwei Hydroxylgruppen (1OH, 2OH) oder ein bzw. zwei Carboxylgruppen (1COOH, 2COOH) in ihrer Molekülstruktur enthalten. (Bezüglich der Gewinnung dieses einwertigen Kontexts durch ordinale begriffliche Skalierung aus dem entsprechenden mehrwertigen s. [1]).

Tabelle 1
Der Kontext ORGMOL

	Abk.	1Ph	2Ph	1OH	2OH	1COOH	2COOH
Benzen	I	X					
Diphenyl	II	X	X				
Essigsäure	III					X	
Benzoesäure	IV	X				X	
Salicylsäure	V	X		X		X	
α,γ-Diphenylacetessigsäure	VI	X	X			X	
Oxalsäure	VII					X	X
Äpfelsäure	VIII			X		X	X
Weinsäure	IX			X	X	X	X
Phthalsäure	X	X				X	X
Phenyläpfelsäure	XI	X		X		X	X
Benzilsäure	XII	X	X	X		X	
β-Resorcylsäure	XIII	X		X	X	X	
Diphensäure	XIV	X	X			X	X
Glykol	XV			X	X		

Die für die Merkmale berechneten relativen Extensitäten $\mathbf{v}(\mathbf{m})$ und prozentualen extensiven Unauffälligkeiten $(1-\mathbf{A}_e(\mathbf{m}))\cdot 100$ sind in Tabelle 2 gegeben. Die Werte der letzteren sind stark gerundet angegeben, da es auf die wirkliche Länge der kürzesten Bahnen nicht ankommt.

Das Liniendiagramm zum Kontext ORGMOL in Form des mit den prozentualen Unauffälligkeiten der Tabelle 2 bogenbewerteten Graphen zeigt Abbildung 1. In ihr ist auch das Ergebnis der Gerüstkonstruktion mit dem Algorithmus von DANTZIG dargestellt worden, wobei die Werte der Potentiale $\mathbf{t}(\boldsymbol{b}_i)$ $(\boldsymbol{b}_i \in \mathcal{B})$ in die Knoten geschrieben wurden und die Bögen des Gerüsts durch starke Linien hervorgehoben wurden.

Tabelle 2
Extensitäten und extensive Unauffälligkeiten der Merkmale des Kontexts ORGMOL ($|G| = 15$)

	$\nu(m)\cdot\|G\|$			$100\cdot(1-\mathbf{A}_e(m))$ [%]		
Anzahl	Ph	OH	COOH	Ph	OH	COOH
1	10	7	12	27	3	51
2	4	3	6	4	13	0,3

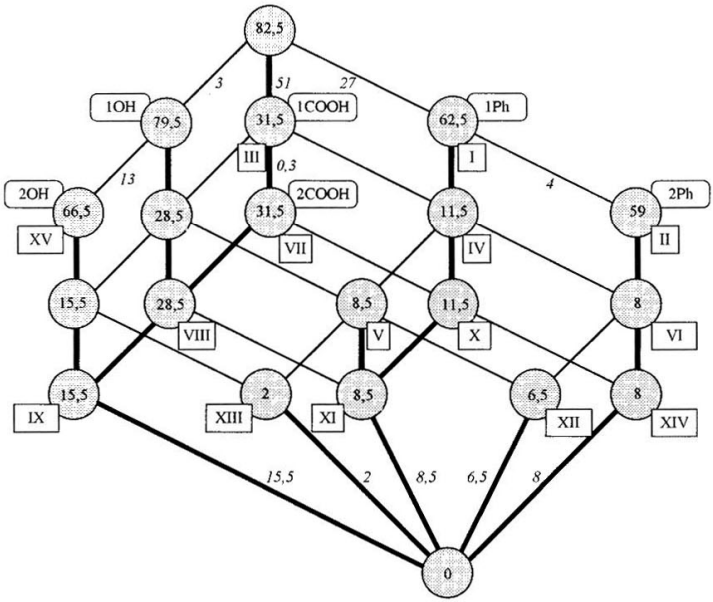

Abb. 1 Liniendiagramm des Begriffsverbandes zum Kontext ORGMOL und Ergebnis der Gerüstkonstruktion mit dem Algorithmus von DANTZIG (Bögen, an welche die Länge nicht angeschrieben wurde, sind mit der Zahl zu bewerten, welche an einem „parallelen" Bogen zu finden ist.)

Natürlich könnte der Algorithmus im Prinzip auch mit dem Einselement („oberster" Begriff) gestartet werden. Die hier verwendete Bogenbewertung führt

dann aber zu einer ziemlich großen Anzahl möglicher Gerüste und damit zu einer ungünstigen Mehrdeutigkeit.

In der Abbildung 2 ist das Gerüst aus Abbildung 1 isoliert gezeichnet worden. Damit ist die durch dessen Konstruktion verbundene Klasseneinteilung besser erkenntlich. Entsprechend dem großen Vorteil der FBA gegenüber den Cluster-Analyse-Methoden ist neben der Gegenstandspartition auch die intensionale Seite der so gewonnenen Zerlegung von **G** ersichtlich. Die einzelnen Klassen entsprechen den Zweigen des Baumes. Die chemischen Verbindungen wurden hier als Strukturformel angegeben, um das Resultat der erhaltenen Partition (und Hierarchie) besser erkennen zu können. Wichtige Merkmale zur Kennzeichnung eines jeden Zweiges sind am Inhalt der jeweiligen Endknoten ablesbar. Die Anzahl der hauptsächlichen Klassen entspricht natürlicherweise der Anzahl der Atome des Verbandes, der oberen Nachbarn des Nullelements. Eine feinere Zergliederung läßt sich an den weiteren Verzweigungen ermitteln.

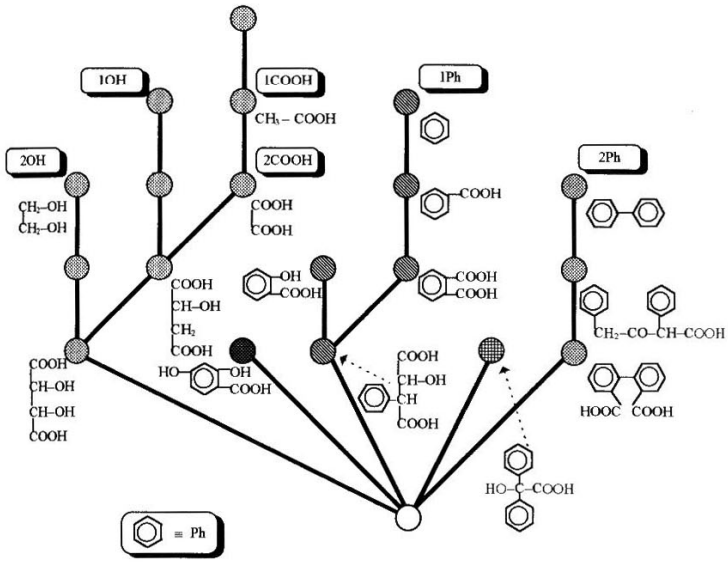

Abb. 2 Gerüst aus Abbildung 1

Im gewählten Beispiel findet man auf diese Weise die Klasse der aliphatischen Verbindungen, die der aromatischen Verbindungen mit einer Phenylgruppe (aber weniger als zwei OH-Gruppen) und die der aromatischen Verbindungen mit zwei Phenylgruppen (außer derartigen Hydroxycarbonsäuren) sowie die beiden Sonderklassen der β-Resorcylsäure (eine Phenyl-, zwei OH-, eine COOH-Gruppe) und der Benzilsäure (zwei Phenyl-, je eine OH- und COOH-Gruppe).

So befriedigend dieses Resultat auch ist, so kann es wegen der Kleinheit und des Beispielcharakters des Kontexts ORGMOL in Hinblick auf die Vielzahl organochemischer Verbindungen nicht für weiterreichende und verallgemeinernde Aussagen über das „Universum" der Molekülstrukturen in der organischen Chemie herangezogen werden. Dazu müssen selbstverständlich noch umfangreichere Studien betrieben werden, um dann umfassendere Kontexte zur Grundlage der Untersuchungen verwenden zu können.

Es muß darauf hingewiesen werden, daß der Algorithmus bei der Festlegung der Verzweigung auf höherer Ebene mehrere gleichberechtigte Lösungen zuläßt. Hierin spiegelt sich wider, daß die mit dem Begriffsverband gegebene Komplexität durch das Gerüst nur partiell erfaßt wird, so daß eine feinere Klasseneinteilung im umfänglichen wie im inhaltlichen Sinne nach verschiedenen Gesichtspunkten erfolgen kann. Es bleibt zu hoffen, daß die Benutzung der gesamten Unauffälligkeit als Bogenbewertung, welche für einen späteren Zeitpunkt vorgesehen ist, die beschriebene Mehrdeutigkeit einschränkt. Andererseits ist die Situation nicht besonders problematisch, da es auf jeden Fall möglich ist, alle gleichberechtigten Gerüste aufzufinden (s. [9]). Nur sollte im Sinne des Nutzers dieser datenexplorativen Methoden deren Anzahl nicht allzu groß sein.

7. Weitere Aufgabenstellungen

Es kann auf Grund der dargestellten Ergebnisse wohl festgestellt werden, daß die beschriebene prinzipielle Vorgehensweise die einleitend genannte Problematik bzw. Aufgabenstellung befriedigend lösen kann. Selbstverständlich sind noch ebenso viele Fragen offen, wie es Möglichkeiten zu ihrer Lösung gibt. Einige davon wurden schon erwähnt. Gedacht ist aber auch an die Anwendung eines anderen Verfahrens wie dem Kaskaden-Algorithmus von FARBEY, LAND und MURCHLAND [14] bzw. des gepflanzten Baumes [15]. Weiterhin sollten die aus der Gesteinskunde stammenden Begriffe der Merkmalsextensität und -intensität auf die Begriffsmenge angewendet werden, um so zu einer noch effektiveren Bogenbewertung zu gelangen.

Neben diesen mehr theoretischen Fragestellungen muß die Lösung der Aufgabe unbedingt automatisiert werden, um gerade in Verbänden mit sehr vielen Begriffen verwendbar zu sein.

Schließlich muß untersucht werden, welche Vorteile die Verwendung des Einselementes („oberster" Begriff) des Liniendiagramms als Startknoten (bei unproblematischer Umkehrung der Bogenrichtungen) für die Gerüstbildung bringt, die

allerdings im hier geschilderten Verfahren zu einer zu großen Mehrdeutigkeit führt. Auch die Benutzung der Gegenstandsbegriffe (Gegenstandsbegriff von $g \in G$: $\gamma g = (\{g\}'', \{g\}')$, hat den kleinsten Umfang, in dem g vorkommt) als Startknoten einer Baumerzeugung für die Herausbildung von (möglicherweise nicht disjunkten) Klassen ist zu diskutieren.

8. Literatur

[1] H.-G. BARTEL, match **30**, 9 (1994)

[2] H.-G. BARTEL, match **32**, 27 (1995)

[3] H.-G. BARTEL, *Formal Concept Analysis Studies: Roman Bricks and Wall Slabs*, in: Proceedings of the 29th International Symposium on Archaeometry, Ankara, May 1994, im Druck

[4] R. WILLE, *Bedeutungen von Begriffsverbänden*, in: B. GANTER, R. WILLE und K.E. WOLFF (Hrg.), *Beiträge zur Begriffsanalyse*, BI-Wissenschaftsverl., Mannheim, 1987, S. 161-211

[5] R. WILLE und B. GANTER, *Formale Begriffsanalyse, Dritter Teil, Mathematische Theorie der Formalen Begriffsanalyse*, Springer-Verlag, Heidelberg, im Druck

[6] L. LEBART, A. MORINEAU und J.-P. FÉNELON, *Statistische Datenanalyse*, Akademie-Verl., Berlin, 1984, S. 123-139

[7] H. SPÄTH, *Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion*, R. Oldenbourg Verl., München/Wien, 1975, S. 147-189

[8] G. BIRKHOFF, *Lattice Theory*, Amer. Math. Soc., New York, 1948

[9] H. SACHS, *Einführung in die Theorie der endlichen Graphen, Teil I*, B.G. Teubner Verlagsges., Leipzig, 1970, S. 122-128

[10] G.B. DANTZIG, Management Sci. **6**, 187 (1960)

[11] H. VÖLZ, *Computer und Kunst*, Urania-Verl., Leipzig/Jena/Berlin, 1990, S. 26-33

[12] *Anselm Feuerbach*, zusammengestellt von E. BARTKE und H. SCHIMPF, Landesbildstelle Rheinland-Pfalz, Speyer, 1980, S. 74-75; G. BOTT, *Malerei 1800 bis 1900*, Katalog des Hessischen Landesmuseums Nr. 7, Dr. Hans Peters Verl., Hanau, 1979, S. 35-36, Tafel 35

[13] R. JUBELT und P. SCHREIBER, *Gesteinsbestimmungsbuch*, Deutscher Verl. für Grundstoffindustrie, Leipzig, 1972, S. 21-22

[14] B.A. FARBEY, A.H. LAND und J.D. MURCHLAND, Management Sci. **14**, 19 (1967)

[15] U. PAPE und D. CONRADT, *Maximales Matching in Graphen*, in: H. SPÄTH (Hrg.), *Ausgewählte Operation Research Software in FORTRAN*, R. Oldenbourg Verl., München/Wien, 1979, S. 103-114

[16] C. POSTHOFF, D. BOCHMANN und K. HAUBOLD: *Diskrete Mathematik*. BSB B.G. Teubner Verlagsgesellschaft, Leipzig 1986, S. 160-190
[17] A.A. ZYKOV: *Theory of Finite Graphs*. Publishing House «Nauka», Siberian Branch, Novosibirsk 1969

Dank

Für eine sehr intensive, kritische und außerordentlich fruchtbare Diskussion über die vorliegende Arbeit bin ich Herrn Prof. Dr. KARL ERICH WOLFF (Fachhochschule Darmstadt) zu großem Dank verpflichtet.