Siegfried Reball
Forschungsinstitut Manfred von Ardenne, Dresden, DDR

Zur Bewertung von Informationsfonds durch die Theorie unscharfer Mengen (On the evaluation of information collections by the theory of fuzzy sets)

Reball, S.: **Zur Bewertung von Informationsfonds durch die Theorie unscharfer Mengen.** (On the evaluation of information resources by means of the theory of fuzzy sets.) (In German).
In: Intern. Classificat. 5 (1978) No. 3, p. 152–155

The information and documentation resources are described by means of the theory of fuzzy sets. Application of the allocation function f_R, well known to be document-oriented, to partial resources characterized as mean-value functions f_R^* permits resources to be evaluated in such a way as to make the effort and rate of recall calculable and experimentally veryfiable for an "optimum retrieval strategy" model derived from such resources. (Author)

Buchbestände, Sonderdrucksammlungen und andere Informationsfonds sind zweckmäßigerweise meist nach inhaltlichen Gesichtspunkten systematisiert. Je nach Sachbereich, Zielstellung, Anforderung und Umfang dieser Fonds sind dabei sowohl die Grenzen zwischen jeweils zwei oder mehr im System unterscheidbarer Teilmengen von Objekten unscharf, als auch die Zugehörigkeit von Objekten zu einer solchen Teilmenge unsicher.

Beide Tatsachen, die sich von der Komplexität der Objekte herleiten und in engem Zusammenhang miteinander stehen, lassen sich durch Verfeinerung des Systems, zumindest im Hinblick auf die Ökonomie, nur begrenzt in ihrer Wirkung vermindern, wie die Versuche zur Steigerung der Systemeffektivität durch „verbesserte", ausdrucksfähigere Dokumentationssprachen nachdrücklich gezeigt haben (1).

Einen möglichen Zugang zur Beschreibung solcher Systemunsicherheiten liefert die Theorie unscharfer Mengen (= fuzzy sets). Mit ihrer Hilfe wird zunächst ein Informationssystem S, durch das ein bestimmter Informationsfond verwaltet wird, als Quintupel dargestellt:

$$S = \langle D, Q, T, K, \psi \rangle,$$

wobei:
D die Menge der Objekte des Informationsfonds (= Dokumente),
T die Menge der Ordnungsmerkmale (= z.B. Deskriptoren), die zur Formulierung von Indexierungen und Anfragen verwendet werden,
Q die Menge der Anfragen,
K eine Menge von Relationen und
ψ eine binäre Zuordnungsfunktion der Art ist:

$$\psi(q) = \begin{cases} D_n, \text{ falls Indexierung } D_n = q \\ 0 \ , \text{ falls Indexierung } D_n \neq q: \end{cases}$$

Die Zuordnungsfunktion leistet die Bildung einer Teilmenge $D_n \subset D$ von Dokumenten, deren Indexierungen mit der Anfrage $q \in Q$ übereinstimmen.

Auf eine Anfrage $q \in Q$ ist die Menge D_n eine Antwort des Systems (man vergl. (2)).

Von den möglichen Relationen über D, T und Q beschreiben die Relationen K den Zusammenhang zwischen D und Q in der Form:

$$K = \{ \langle q, d, f_R(q, d) \rangle / q \in Q, d \in D \} .$$

Die Funktion f_R (q, d) weist dabei jedem geordneten Paar aus einer Anfrage und einem Dokument ⟨q, d⟩ eine reelle Zahl aus dem diskret gedachten Intervall [0, 1] zu. Die Relationen K sind damit Unschärferelationen im Sinne der Theorie unscharfer Mengen. Sie ordnen die Dokumente der Menge D_n nach dem Grad des Zutreffens der Dokumenteninhalte auf die Anfrage bzw. im gleichen Sinn nach dem Grad des Zutreffens der Indexierung, wenn es gelingt, vorab jedes Dokument mit einem entsprechenden „Zugehörigkeitsmaß" der Indexierung (3) als reelle Zahl größer Null und kleiner Eins zu kennzeichnen unter der Annahme, daß bei der Indexierung wie bei der Anfrageformulierung ein gleicher Grad des Zutreffens des Dokumenteninhaltes zur Indexierung wie zur Anfrage vorhanden ist. Dann kann $\psi(q)$ schärfer formuliert werden mit:

$$\psi(q) = \begin{cases} D_n, & \text{falls Indexierung } D_n = q \text{ und} \\ & f_R(q,d)_{D_n} \geqslant f_R(q,d)_{min} \\ 0, & \text{falls Indexierung } D_n \neq q \text{ oder} \\ & f_R(q,d)_{D_n} < f_R(q,d)_{min} \end{cases}$$

$f_R(q, d)_{min}$ muß dabei von Fall zu Fall vom Nutzer festgesetzt werden.

Da man sich einen Informationsfond immer zerlegbar in Teilfonds vorstellen kann, wobei die Bildung solcher Teilfonds apriori beispielsweise durch ein Klassifikationssystem erfolgt oder aposteriori durch Anfrageformulierungen, die je nach der Kombination von Ordnungsmerkmalen entsprechende Dokumentmengen zusammenstellen, wird eine weitere Funktion f_R^* eingeführt, die für beliebige Teilfonds festsetzt:

$$f_R^*(q, d_{T^*}) = \frac{1}{N_{T^*}} \Sigma f_R(q, d_{T^*}),$$

d.h. eine Mittelwertbildung aller zugeordneten Zahlen aus dem Intervall [0, 1] zu den Dokumenten 1 bis N des betrachteten Teilfonds T^*. Im Grenzfall $N_{T^*} = 1$ ist $f_R^* = f_R$.

Damit ist, wie im folgenden gezeigt werden soll, ein experimenteller Zugang zur Bewertung von Informationsfonds geschaffen, als Ersatz für die geforderte praktikable Lösung der Bestimmung von f_R (q, d) (man vgl. (3) (4)).

Im Extremfall totaler Undifferenziertheit eines Informationsfonds ist bei einer beliebigen Anfrage der Gesamtfond zu durchmustern, was durch die Angabe

$f_R^*(q,d) = \text{const.}$, mit $0 < \text{const.} \leqslant 1$,

beschrieben werden kann. Das bedeutet, daß bei jeder Unterteilung des Gesamtfonds in Teilfonds keine Anreicherung an relevanten Dokumenten erzielt werden kann. Mit f_R^*-Werten nahe 1 ist ein großer Anteil relevanter Dokumente des Gesamtfonds angezeigt, mit f_R^*-Werten nahe 0 ein großer Anteil nicht relevanter Dokumente. Die genaue Bestimmung der f_R-Werte für jedes Dokument steht dabei aus; man kann vereinfacht z.B. drei Werte für f_R festsetzen: $f_R = 1, f_R = 0{,}5, f_R = 0$.

Entsprechend ((5), S. 200) werden folgende Maßgrößen eingeführt:

Dokument / Bei der Recherche	relevant	nicht relevant
aufgefunden	a Treffer	b Ballast
nicht aufgefunden	c verfehlte Treffer	d verfehlter Ballast

und mit: $a + b = R$; $a + c = P$

wird festgesetzt:

Pertinenz (Sachdienlichkeit): $\frac{a}{K}$ Relevanzrate $\frac{100\,a}{R}$

Recallfaktor (Wiederaufruf): $\frac{a}{P}$ Recallrate $\frac{100a}{P}$

Wird der Recallfaktor im Hinblick auf die o.g. Teilfondbildung für jeden Teilfond getrennt ermittelt, so wird, abgesehen von dem Grenzfall $f_R^*(q,d) = \text{const.}$, der Recallfaktor für verschiedene Teilfonds verschieden groß sein. Das Vorhandensein von Teilfonds mit sehr großem Recallfaktor zeichnet Systeme aus, in denen bei richtiger Anfrageformulierung hohe Wiederaufrufraten erzielt werden. Ordnet man die Teilfonds nach fallenden Recallfaktoren (= optimale Suchstrategie), so wird festgesetzt, daß sein soll:

$\frac{a}{P} = \sqrt[\epsilon]{n}$, mit $\epsilon \geqslant 1$ und $0 < n \leqslant 1$ und

$n = \frac{N_{\text{bisher durchmusterter Teilfonds}}}{N_{\text{Gesamtfond}}}$

und der Zusammenhang des Recallfaktors mit der Zuordnungsfunkuon $f_R^*(q,d)$ soll sein:

$$f_R^*(q,d)_i = \left(\frac{a}{P}\right)_i - \left(\frac{a}{P}\right)_{i-1}$$

für den i-ten Teilfond.

Ein Informationsfond mit der Möglichkeit zur Teilfondbildung ist dann beschreibbar in der Form:

$S = \langle D, Q, T, K, \psi \rangle$

$K = \{\langle q, d, f_R^*(q,d)_i \rangle / q \in Q, d \in D\}$ Teilfond i

$$f_R^*(q,d)_i = \left(\frac{a}{P}\right)_i - \left(\frac{a}{P}\right)_{i-1} = \frac{\sqrt[\epsilon]{i N_T^*} - \sqrt[\epsilon]{(i-1) N_T^*}}{\sqrt[\epsilon]{N_{Ges}}}$$

und

$$\left(\frac{a}{P}\right)_i = \sqrt[\epsilon]{\frac{i \cdot N_T^*}{N_{Ges}}}.$$

Abb. 1 zeigt den Zusammenhang zwischen Recallrate und „Suchraum", wenn dieser in optimaler Weise durchmustert wird, d.h. Teilfond für Teilfond nach fallenden Recallfaktoren.

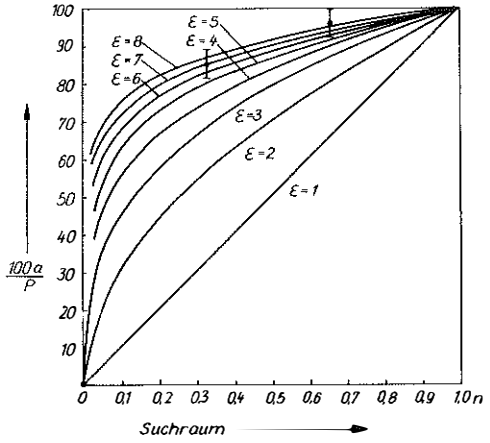

Abb. 1: Modell der Recallrate $\frac{100a}{P}$ bei optimaler Strategie und der Abhängigkeit $\frac{a}{P} = \sqrt[\epsilon]{n}$ eines Dokumentenfonds

Ein praktisches Beispiel dieses funktionellen Zusammenhanges zeigt Abb. 2. Die Makulaturfaktoren von Zeitschriften bestimmen die Recallrate, so daß die Ordnung von Zeitschriften nach zunehmendem Makulaturfaktor der optimalen Strategie entspricht. Die Kurve gehorcht einigermaßen der postulierten $\sqrt[\epsilon]{\ }$-Abhängigkeit mit $\epsilon \approx 10$.

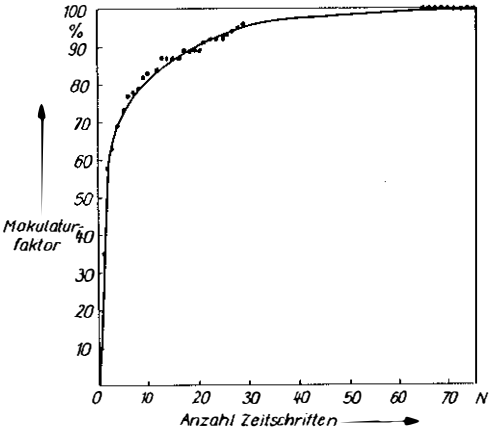

Abb. 2: Makulaturfaktoren institutseigener Zeitschriften nach der Zitierhäufigkeit geordnet (nach (6))

Um die $\sqrt[\epsilon]{n}$-Abhängigkeit als Maß für die Unschärfe von Dokumentenfonds zu prüfen, wurde ein Gesamtfond von 1129 Dokumentennachweisen über den Sachbereich „Dünne Schichten" untersucht. Der Gesamtfond besteht entsprechend der Indexierung aus drei Teilfonds:

A: Dünne Schichten, Eigenschaften (698)
B: Dünne Schichten, Messungen (142)
C: Dünne Schichten, Herstellung (289)[1].

Je nach thematischer Breite (ohne Bezug auf Substanzen oder Substanzgruppen, Bezug auf Substanzgruppe, Bezug auf einzelne Substanz) der Dokumente wurde anhand Dokumententitel und Referat geprüft, wieweit Dokumente auch in die Nachbarfonds gehören können, da oft nicht nur über die Herstellung einer Dünnen

Schicht, sondern *gleichzeitig* über die Eigenschaften oder die Messung (z.B. während der Herstellung) *in einem* Dokument berichtet wird. Je nach Gewicht der einzelnen Anteile gehört das Dokument mehr oder weniger genau in nur einen Teilfond, was durch die Funktion f_R und bei entsprechender Teilfondbildung durch f_R^* ausgedrückt werden kann. Diese Dokumente sollen bi-, triusw. thematisch genannt werden. Bei eventueller Duplizierung des Nachweises für zwei oder mehr Teilfonds wird der Ballast größer, ohne die Relevanz (eine optimale Suchstrategie vorausgesetzt) zu verbessern.

Das Ergebnis der Untersuchung, eingeschränkt auf bithematische Dokumente, zeigt Tabelle 1.

Tab. 1: Anzahlen mono- und bithematischer Dokumente eines Fonds

Aspekt / Substanz	Eigenschaften Dünner Schichten A	Messungen an Dünnen Schichten B	Herstellungsmethoden Dünner Schichten C
Allgemein	57 A 19 (A+C) 3 (A+B)	14 B 24 (A+B) 4 (B+C)	86 C 10 (A+C) 3 (B+C)
Metall	32 A 1 (A+B)	2 B 1 (A+B) 1 (B+C)	25 C 4 (A+C)
Legierung	3 A 1 (A+C)	1 (A+B)	5 C 1 (A+C)
Plast	8 A	1 (A+B)	9 C 1 (B+C)
Aluminium	24 A 5 (A+B) 3 (A+C)	2 (A+B)	8 C 1 (A+C) 1 (B+C)
Nickel	22 A 5 (A+B) 2 (A+C)	4 (A+B)	2 C 1 (A+C)
CdS	11 A 2 (A+B) 2 (A+C)	-	-
Summa	200	54	157
Gesamtteilfond	698	142	209

Aufsummiert ergibt sich daraus Tabelle 2, die unmittelbar Treffer und Ballast bei Recherchen ausweist unter der Annahme, daß jeweils 50 % der bithematischen Dokumente relevant bzw. nicht relevant für die Fragestellung sind[2]. So können im Mittel beim ersten Suchschritt im jeweils richtigen Fond 85 % relevante Dokumente gefunden werden. Beim zweiten Suchschritt im jeweils nächstliegenden Teilfond können weitere 11 % relevante und beim dritten Suchschritt weitere 4 % gefunden werden. Diese Werte wurden in die Abb. 1 eingetragen und weisen ein ϵ von 7 bis 9 nach[3].

Tab. 2: Summenwerte mono- und bithematischer Dokumente eines Fonds

Aspekt / Alle Substanzen	Eigenschaften A	Messungen B	Herstellung C
Summenwerte	127 A 46 (A+B) 27 (A+C) } 240	16 B 33 (A+B) 5 (B+C) } 54	135 C 17 (A+C) 5 (B+C) } 157
Prozentanteile relevanter Dokumente 1)	77 für A 18 für B 4 für A 8 für C 7 für A	35 für B 8 für A 36 für B 2 für C 5 für B	83 für C 4 für A 5 für C 5 für B 2 für C
Auf der Suche nach relevanten Dokumenten werden im 1. Suchschritt gefunden	80 %	76 %	90 %

1 errechnet nach der im Text genannten 50 : 50 Prozedur.

Unter der Annahme unterschiedlicher Strategien ergeben sich im Mittel folgende Relevanz- und Recallraten:

	Relevanzrate	Recallrate
A) nur ein Suchvorgang im entsprechenden Teilfond	81	85
B) zusätzlich zweiter Suchvorgang im nächsten Nachbarfond	45	96
C) zusätzlich dritter Suchvorgang im übernächsten Nachbarfond (entspricht der einmaligen Suche im Gesamtfond)	32	100
D) nur ein Suchvorgang im nächsten Nachbarfond	13	11
E) nur ein Suchvorgang im übernächsten Nachbarfond	4	4

Diese Werte zeigt die Abb. 3. Das Wertepaar für Suchvorgang A ist das wegen der gefundenen Unschärfe bestmögliche Ergebnis bezüglich Recall und Relevanz. Jede weitere Steigerung an Treffern verschlechtert die Relevanz und bei nichtoptimaler Suche (E oder D) verschlechtern sich Relevanz und Recall sogar erheblich.

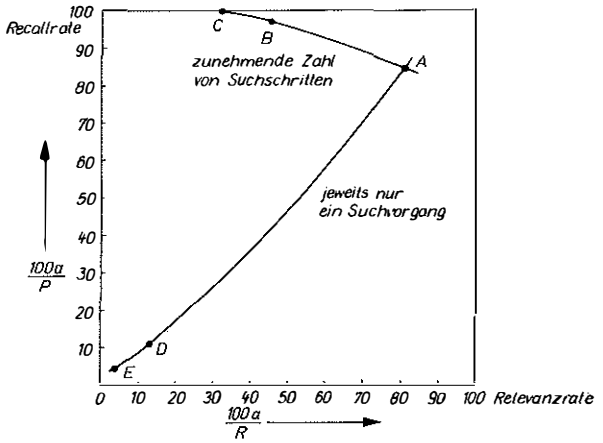

Abb. 3: Die optimale Recallrate als Funktion der Relevanzrate bei unterschiedlichen Suchstrategien in 3 Teilfonds des Fonds „Dünne Schichten"

Da die Zahl der Suchschritte direkt ein Maß für den Suchaufwand ist, zeigt Abb. 4 die Recallrate der Strategie A + B + C als Funktion der Suchschritt-Anzahl.

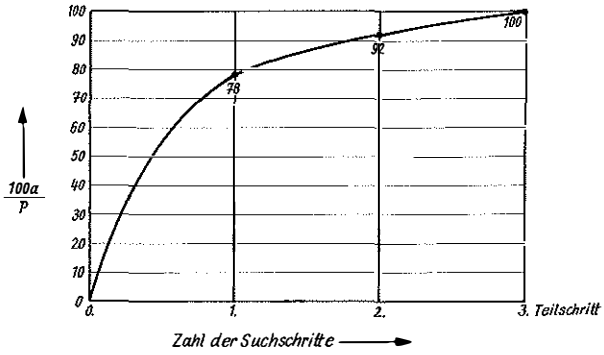

Abb. 4: Die Recallrate $\frac{100a}{P}$ bei optimaler Strategie und drei Teilfonds des Gesamtfonds „Eigenschaften Dünner Schichten"

In Systemen mit großen ϵ-Werten kann man mit wenigen Suchschritten in einem stark unterteilten Fond große Recallraten erzielen. Sind die ϵ-Werte klein, nahe

1, so sind die Recallraten klein und der Aufwand groß. Diesen Zusammenhang zeigt Abb. 5. Die Kurven stellen für konstante Recallraten den Aufwand als Funktion von ϵ dar. Der Aufwand ist hier von der Anzahl gleichgroßer Teilfonds abgeleitet worden. Gelingt es, viele Teilfonds zu bilden, so kann, wiederum bei optimaler Strategie, nach dem Durchmustern entsprechend geringer Dokumentmengen eine hohe Recallrate erzielt werden. Bei einer vorgegebenen Recallrate kann damit bei bekanntem ϵ der Aufwand direkt bestimmt werden.

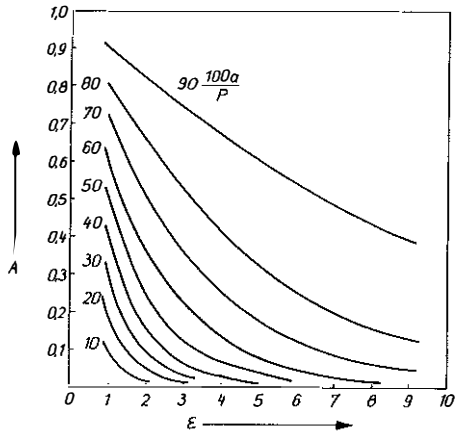

Abb. 5: Der Aufwand A für den Wiederaufruf in Informationsfonds als Funktion von ϵ

Die entsprechende Empfehlung für optimale Systemgestaltung ist: Bei bekanntem ϵ und vorgegebener Recallrate ist die Teilfondbildung der Aufwandzahl anzupassen mit:

Anzahl der Teilfonds $T^* = \frac{1}{A}$. Bei optimaler Strategie ist dann das bestmögliche Suchergebnis erreichbar.

Unübersehbar ist die starke Vereinfachung des beschriebenen Modells. U.a. zeigt die detaillierte Berechnung der Relevanz- und Recallraten für die Teilfonds beträchtliche Abweichungen vom Mittelwert, wie Tabelle 2 ausweist. Setzt man jeweils die monothematischen Dokumente zu den bithematischen in Bezug, so ergibt sich die abschließend vorgelegte Tabelle 3. Sie enthält in ihren Zeilen die Anteile von bithematischen Dokumenten der beiden Nachbarfonds im jeweils betrachteten Fond. So wird in Dokumenten über die Messung an Dünnen Schichten sehr häufig (zu 61 %) auch über die Eigenschaften Dünner Schichten berichtet, andererseits sind die Herstellung und das Messen nur gering (zu 3 %) miteinander verknüpft.

Tab. 3: Anteile mono- und bithematischer Dokumente in drei Teilfonds des Gesamtfonds „Dünne Schichten"

betrachteter Teilfond \ bezogen auf den Teilfond	Eigenschaft	Messung	Herstellung
Eigenschaft	1	0,08	0,13
Messung	0,61	1	0,09
Herstellung	0,11	0,03	1

Anmerkungen:

1 Dies bedeutet verallgemeinert eine Beziehung zwischen den Ordnungsmerkmalen T und den Anfragen Q als ein einfacher aber häufig anzutreffender Fall:

$Q = R^{i,u}_{n,m}({}_nT_i, {}^m_nT_u)$.

Nach (7) sind dabei n ein Index für Situationsmuster, m ein laufender Index und i und u Verknüpfungsanzeiger. Die Situation hier ist „Dünne Schicht" mit drei möglichen Relationen (Prädikation, „Observation", „Produktion") oder auch der Sachverhalt aufzufassen als unterscheidbar durch drei Situationen. Je nachdem ist der Index n fest oder läuft von 1 bis 3.

2 Damit erfolgt eine, für lediglich mono- und bithematische Dokumente mögliche, oben bereits vorgeschlagene Dreiteilung in $f_R = 0$, $f_R = 1$ und $f_R = 0{,}5$, wovon sofort die Dokumente zu $f_R = 0{,}5$ aufgeteilt werden. Wie die später gezeigte Tabelle 3 vermuten läßt, ist die 50 : 50-Aufteilung nicht korrekt, denn die bithematischen Dokumente sind unsymmetrisch auf die Nachbarfonds verteilt.

3 ϵ könnte als Selektivitätsmaß bezeichnet werden. Sollte ein ϵ von 7 bis 9 typisch für Informationsfonds sein, so liefert beispielsweise die UDC bei einer Teilfondbildung in jeweils rund 10 Teilmengen beim ersten Suchschritt einer optimalen Strategie eine Recallrate von rund 75. Sicher ist dies ein genügend gutes Ergebnis, wenn man mit den Erfahrungswerten bei anderen Informationssystemen vergleicht.

Literatur

(1) Sparck Jones, K., Kay, M.: Linguistik und Informationswissenschaft. München: Verlag Dokumentation 1976, S.31, 82 u. 163.

(2) Radecki, T.: Mathematical Model of Information Retrieval System based on the Concept of fuzzy Thesaurus. In: Inform. Process. & Management 12 (1976) Nr. 5, S. 313–318.

(3) Sachs, W. M.: An Approach to Associative Retrieval through the Theory of Fuzzy Sets. In: J. Amer. Soc. Inform. Sci. 27 (1976) Nr. 2, S. 85–87.

(4) Laus-Maczynska, K.: Die Effektivität von Informationsrecherchesystemen (IRS) unter dem Aspekt unscharfer Mengen. Vortrag X. Koll. Inf. u. Dok. der TH Ilmenau, 8.–11. Nov. 1977 Oberhof.

(5) Vickery, B. C.: Zur Theorie von Dokumentationssystemen. München-Pullach: Verlag Dokumentation 1970.

(6) Reball, S.: Rationalisierung der Informationstätigkeit in kleinen Informationsstellen. Vortrag 4. Fachtagung für wiss.-techn. Information 25.–26.11.1976, Karl-Marx-Stadt.

(7) Reball, S.: Semantischer Bereich und Sachverhaltsbeschreibung von Dokumenteninhalten. In: Internat. Classificat. 3 (1976) Nr. 1, S. 18–22.