Informatik Forsch. Entw. (1989) 4: 89-107

Informatik
Forschung und Entwicklung

TWRM-TOPOGRAPHIC

Ein wissensbasiertes System zur situationsgerechten Aufbereitung und Präsentation von Textinformation in graphischen Retrievaldialogen

R. Kuhlen, R. Hammwöhner, G. Sonnenberger und U. Thiel

Universität Konstanz, Postfach 5560, D-7750 Konstanz 1

Zusammenfassung. TWRM-TOPOGRAPHIC ist Teil eines neuartigen Informationssystems, das sich auf inhaltsorientierte Repräsentation von Volltexten stützt. Die beiden wesentlichsten Leistungsmerkmale von TWRM-TOPOGRAPHIC sind die *graphische Retrievaldialogführung* über ein flexibles, objektorientiert spezifiziertes *User-Interface-Management-System (UIMS)* und die flexible, situationsgerechte Aufbereitung und Präsentation von Textwissen: Die Dialogführung erlaubt dem Benutzer die direkte Navigation in den auf dem Bildschirm graphisch dargestellten Wissensstrukturen, die Selektion dargestellter Objekte zur Formulierung einer Suchfrage sowie das Wechseln des Abstraktionsniveaus der dargestellten Textinformation. Textwissen wird in unterschiedlichen Abstraktionsstufen präsentiert: von einer sehr generischen Ebene über *Wissensgraphen*, automatisch generierten *Abstracts* mit unterschiedlichen Themenschwerpunkten und variabler Ausführlichkeit bis zur diskursiven Form der *Textpassage*.

Schlüsselwörter: Automatisches Textkondensieren, interaktives graphisches Retrieval, pragmatischer Systemdesign

Abstract. TWRM-TOPOGRAPHIC is part of a novel type of information-systems, which is based on a semantic representation of thematic descriptions of fulltexts. TWRM-TOPOGRAPHIC employs graphical interaction to provide access to the knowledge bases (text and world knowledge) and presents information in a flexible, situation-specific way. The user may navigate directly within the knowledge structures depicted on the screen, select objects in order to formulate a query and vary the abstraction-level of the represented textual information. Textual knowledge is presented on variable layers of abstraction including a taxonomical layer, conceptual graphs, automatically generated abstracts (varying in detail and thematic focal point) and the discursive form of textpassages.

Key words: Automatic text condensation, interactive graphical retrieval, pragmatic system design

CR Subject Classifications: H.3.1, I.2.7, H.3.3, I.3.6, I.2.8

1 Entwurfsziele eines neuen Systemtyps: Prototypische Realisierung durch TWRM-TOPOGRAPHIC

Durch die fortschreitende Verwendung moderner Drucklegungstechniken, den vermehrten Einsatz dezentraler Arbeitsplatzrechner, durch *Mail- und Message-Systeme* und Formen des elektronischen Publizierens liegen in allen Bereichen der Fachkommunikation zunehmend mehr Texte maschinenlesbar (und damit maschinenverarbeitbar) vor. Diese fast schon flächendeckende Verbreitung elektronischer Volltexte korrespondiert aber keinesfalls mit entsprechend leistungsstarken Nachweis-, Verarbeitungs-/Aufbereitungs- oder Präsentationstechniken. Die Gefahr besteht, daß potentiell relevante Information in maschinenlesbaren Speichern verschwindet, ohne bekannt zu werden.

Angesichts des großen Aufwandes, Volltexte intellektuell über Indexierung oder durch Referieren inhaltlich zu erschließen, sollen nach Einschätzung vieler Informationspraktiker Volltext-Retrieval-Systeme die Aufgabe des Relevanznachweises über-

nehmen, und entsprechend haben Volltextdatenbanken von den zur Zeit ca. 3500 weltweit verfügbaren elektronischen Online-Informationsbanken die größten Zuwachsraten; innerhalb eines halben Jahres ist ihre Zahl von ca. 800 auf heute (4/88) 1100 angestiegen.

Den Fortschritten bei Speicherungstechniken und Zugriffsmöglichkeiten entsprechen aber nicht im gleichen Ausmaß Fortschritte bei der automatisierten Inhaltserschließung oder beim Retrieval. Bislang gibt es kaum eigens für die Verarbeitung von Volltexten entwickelte Retrieval-Systeme. Diese Defizite gelten für Inhaltserschließung und Suchtechniken (das eigentliche Retrieval) gleichermaßen.

1.1 Inhaltserschließung

Zwar werden in manchen Fällen zur Wahrung von Qualitätsstandards intellektuelle Erschließungstechniken unter Verwendung von Klassifikationssystemen oder Thesauri eingesetzt, bei den meisten Volltextsystemen wird aber eine Automatisierung der Inhaltserschließung angestrebt, und zwar dadurch, daß die Volltexte invertiert werden. Dieses Verfahren, bei dem also die Textwörter unverändert als Suchargumente beim Retrieval verwendet werden, kann man allerdings kaum als Inhaltserschließung bezeichnen. Zu wenig kann bei diesem Verfahren der Bandbreite morphologischer, syntaktischer und semantischer Varianten Rechnung getragen werden. Weitergehende Verfahren der *automatischen Indexierung*, soweit sie überhaupt zum Einsatz kommen, sind fast ausschließlich wort- oder partiell satzorientiert und beruhen entsprechend auf morphologischen oder syntaktischen Konzepten und einer schwachen und/oder statistisch basierten Semantik [28]. Textadäquate Analysetechniken kommen so gut wie gar nicht zum Einsatz (vgl. Literaturbericht [12]). Soweit die Inhaltserschließung nicht nur durch Indexate auf die Originaltexte verweist, sondern Textzusammenfassungen liefern soll, ist man auf intellektuell erstellte Referate angewiesen. Die zahlreichen, seit den sechziger Jahren laufenden Experimente des *automatischen Abstracting* auf statistischer und/oder oberflächenlinguistischer Basis (vgl. [21]) haben in keinem Fall zu einem realen Einsatz in Informationssystemen geführt. Die verschiedenen Ansätze aus der Künstlichen Intelligenz, die versuchen, aus Wissensstrukturen Textzusammenfassungen zu erzeugen (vgl. Abschn. 3.1), sind weniger auf eine praktische Anwendung in der Fachkommunikation ausgerichtet, sondern mehr an der Simulation einer kognitiv hochstehenden humanen Leistung interessiert.

1.2 Retrieval

Angesichts der Semantikdefizite bei der Inhaltserschließung realer Informationssysteme können natürlich keine leistungsfähigen Retrievalsysteme entstehen. Der methodische Stand der sechziger Jahre wird insgesamt kaum überschritten (vgl. [28]), so daß Veränderungen auf der Retrieval-Seite entweder formalsyntaktische Verbesserungen (Menütechnik, Ikonengraphik) sind oder aber die Idee der Kontextoperatoren, mit denen Kontext- bzw. Abstandsbedingungen in der Kombination von Fragewörtern definiert werden können, weiter verfeinern (z. B. Berücksichtigung des Satz- oder Absatzkontexts, Lokalisierung der Suchargumente an bevorzugten Stellen, z. B. am Anfang von Absätzen). Entsprechend ist sich die Forschung nach verschiedenen Evaluierungsstudien (z. B. [25]) einig, daß Volltextsysteme in ihrer jetzigen Ausprägung, obgleich sie immer stärkere Verwendung finden, nur ein unzureichendes Mittel der Volltextverarbeitung sind.

Nach unserer Einschätzung sind bisherige *(Volltext-)Informations-Retrieval-Systeme* eher „Konfektionsware" und als solche natürlich auch weiterhin unersetzlich; sollen jedoch „maßgeschneiderte" Systeme entstehen, müssen Inhaltserschließung, Retrieval und die Präsentation der Suchergebnisse mehr auf intelligente Verfahren abgestützt werden. *Informationssysteme* – und dies folgt aus unserem informationswissenschaftlichen Verständnis von Information, nach dem Information lediglich als die Teilmenge von Wissen verstanden werden sollte, die aufgrund konkreter Benutzerbedürfnisse und Bedarfssituationen aktuell gebraucht wird – verdienen erst dann ihren Namen, wenn sie über pragmatische Komponenten verfügen, also in der Lage sind, flexibel, d. h. benutzer- und situationsgerecht zu reagieren. Neben leistungsstarken Semantik-Komponenten und, im Falle von textorientierten Systemen, robusten und auf die Möglichkeiten der Wissensrepräsentation zugeschnittenen Textanalyseverfahren sind also auf der Retrieval-Seite flexible Ableitungs- und Darstellungsformen von Wissens- und Textstrukturen notwendig. Auf unterschiedliche Anforderungen müssen Retrieval-Systeme differenziert reagieren können.

Diese Überlegung, den Zugang zu immer größer werdenden Textmengen durch leistungsstärkere Verfahren offen zu halten, war Ausgangspunkt der seit 1982 durchgeführten Forschungsprojekte TOPIC[1],

[1] TOPIC (*T*ext *O*riented *P*rocedures for *I*nformation Management and *C*ondensation of Expository Texts, Projektträger: GID, Förderungskennzeichen: 10200160) wurde in C entwickelt. Die Software ist portierbar und läuft z. Zt. auf einem Cadmus 9200 unter UNIX™

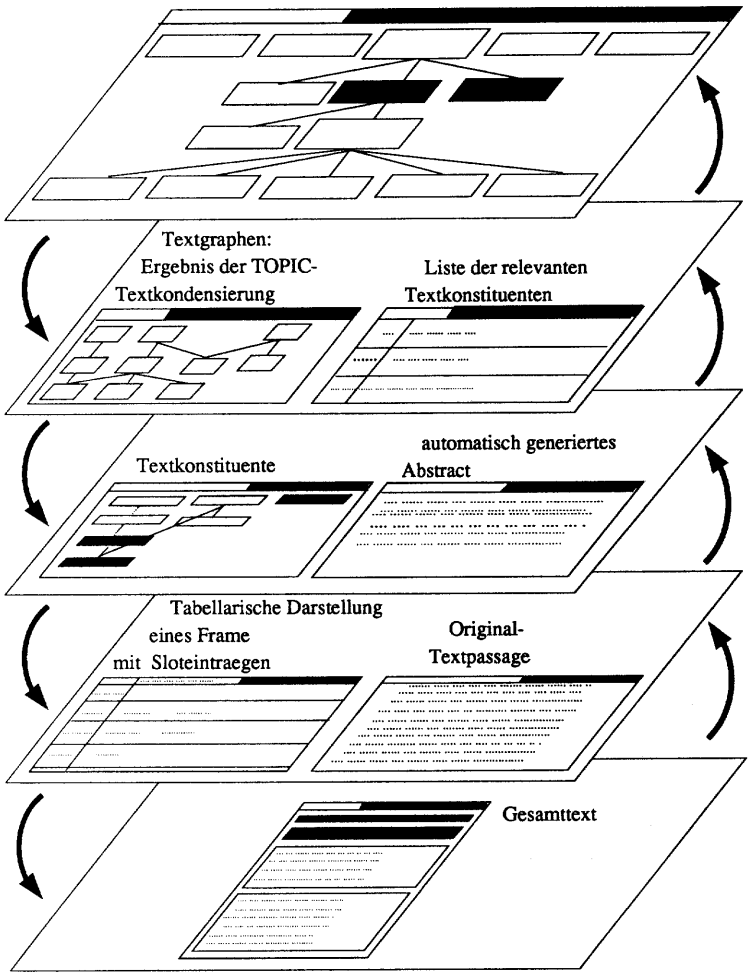

Abb. 1. Stufen der kaskadierten Kondensierung in TWRM-TOPOGRAPHIC

TOPOGRAPHIC[2] und TWRM-TOPOGRAPHIC[3]. Dabei haben wir das zum allgemeinen Gestaltungsprinzip gemacht, was uns die besondere Stärke rechnergestützter Systeme zu sein scheint: die Flexibilität der Verarbeitung und Darstellung. Anstatt primär mit der menschlichen Leistung zu konkurrieren, die in unserem Fall darin besteht, aus einem zehnseitigen Text eine zehnzeilige Zusammenfassung zu erstellen, sollte das Ziel verfolgt werden, aus den bei der Textanalyse erstellten Textwissensstrukturen variable „Kondensate" in unterschiedlichen, sich im Benutzerdialog entwickelnden Kaskaden zu präsentieren. Entsprechend ersetzen wir das Konzept des *automatischen Abstracting* durch das des *kaskadierten Kondensierens*.

Speziell bei dem Entwurf von TWRM-TOPOGRAPHIC sind wir davon ausgegangen, daß die erwünschte Flexibilität wesentlich über graphische Darstellungsformen erreicht werden soll (vgl. [11], [21]), wobei allerdings auch textuell ausgestaltete Kaskadierungsstufen als graphische Objekte aufgefaßt und präsentiert werden. TWRM-TOPOGRAPHIC steht also in der Tradition des graphikorientierten Retrieval. Insgesamt beruht die graphische Ausrichtung der Dialogführung und der Präsentation auf kognitiv-ergonomischen Prinzipien. Das System berücksichtigt die begrenzte Aufnahmekapazität von Benutzern und stellt die Bedeutung der zeitlichen Anordnung von Informationseinheiten für Wahrnehmung und Gedächtnisspeicherung in Rechnung.

Die Idee des kaskadierten Kondensierens wird in TWRM-TOPOGRAPHIC nach dem gegenwärtigen Stand über die folgenden graphisch realisierten Stufen verwirklicht (vgl. Abb. 1):

[2] TOPOGRAPHIC (*TOP*ic *O*perating with *GRAPH*ical *I*nteractive *C*omponents, Projektträger: GID, Förderungskennzeichen: 10200160) ist in C und IF-Prolog implementiert und z. Zt. auf einem Cadmus 9200 unter UNIX™ installiert

[3] TWRM-TOPOGRAPHIC (*T*ext*w*issens*r*ezeptions*m*echanismus TOPOGRAPHIC, Projektträger: GMD, Förderungskennzeichen: 10200181) ist als eine Erweiterung von TOPOGRAPHIC ebenfalls in C und IF-Prolog programmiert und läuft auf einem Cadmus 9200 unter UNIX™

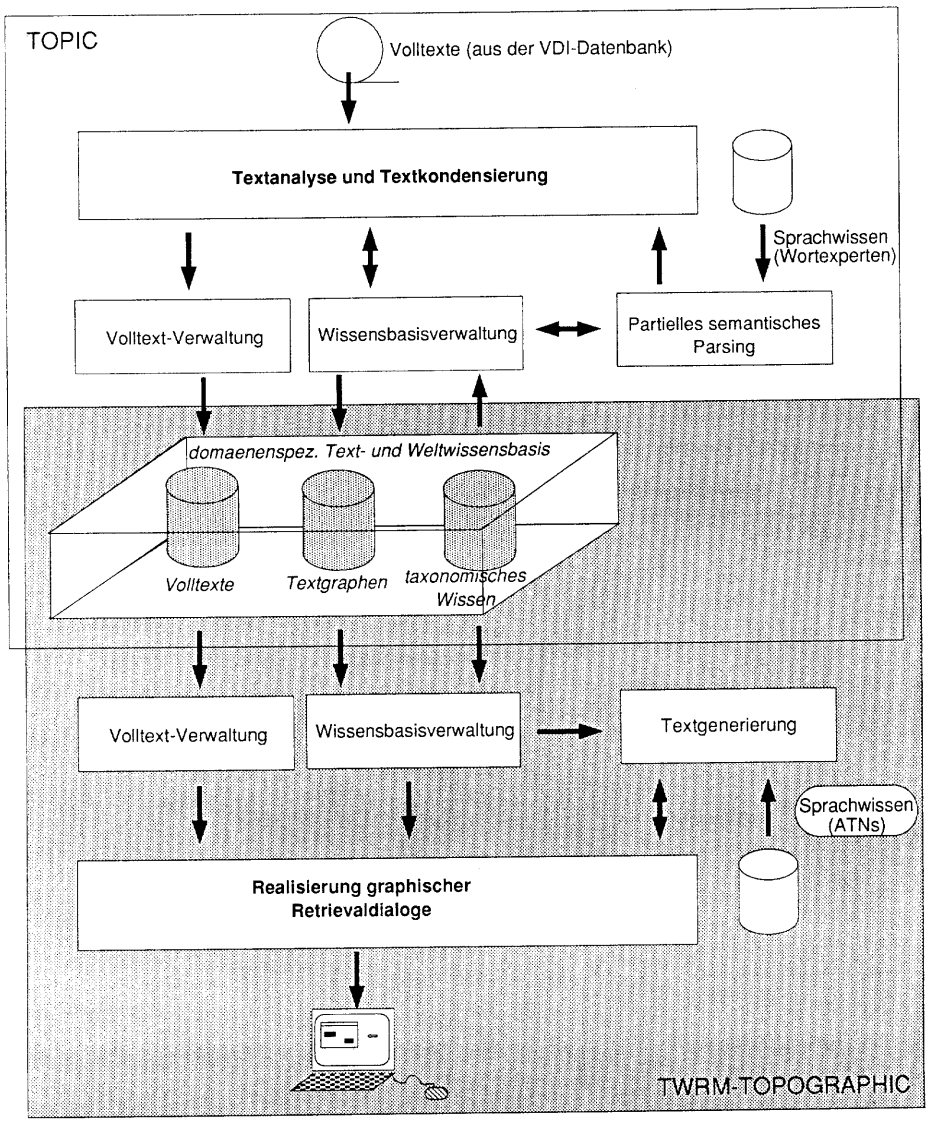

Abb. 2. Architektur des Informationssystems TOPIC/TWRM-TOPOGRAPHIC

• Taxonomische Informationen über den Diskursbereich eines Textes (bzw. einer Textmenge) werden über die durch Relationen verbundenen Konzepte der Weltwissensbasis bereitgestellt. Eine graphische Präsentation der Begriffshierarchie erlaubt eine Auswahl der relevanten Konzepte.
• Die semantische Struktur der jeweiligen Texte wird durch einen Textgraphen dargestellt, der als Ergebnis der TOPIC-Analyse bereitgestellt wird.
• Nach der prinzipiellen Relevanzentscheidung für einen Text wird die graphische Darstellung der thematischen Struktur der Textpassagen zugänglich. Sie erlaubt dem Benutzer,
- die Detailinformation zu den als einschlägig erkannten Konzepten zu betrachten, die in tabellarischer Form angeboten wird, oder
- die entsprechende Passage als Ganzes im Original zu lesen.
• Die in einem Textgraphen vorgegebene thematische und faktische Information bietet zusätzlich die Möglichkeit, nach situationsspezifischer Selektion relevanter Netzteile ein benutzerangepaßtes Abstract, also eine textuelle Kurzfassung, anzubieten.
• Nach Bedarf ist das ‚Blättern' im Gesamttext und die Anzeige von Grafiken im Text möglich.

Mit diesem Beitrag geben wir einen Überblick über die Systemkomponenten und den Leistungsstand des Gesamtsystems. Die Ergebnisse der Textanalyse, basierend auf einem *Frame-Modell* [26] und einem lexikalisch verteilten *Text-Parsing* [13], werden

kurz in Abschnitt 2 zusammengefaßt, allerdings nur insoweit, als es für das Verständnis der Ausgabe- bzw. Interaktionsleistungen erforderlich ist. Das Ergebnis wird - wie erwähnt - in einem konzeptuellen Textgraphen repräsentiert, welcher die Basis für die weitere, hier näher darzustellende Verarbeitung in TWRM-TOPOGRAPHIC ist. Abschnitt 3 stellt die Erweiterung der ursprünglich rein graphisch konzipierten Ausgabe durch einen Abstract-Generator dar, der aus den in den Textkonstituenten enthaltenen Wissensstrukturen natürlichsprachige flexible Zusammenfassungen erzeugt. Abschnitt 4 enthält eine Beschreibung des graphischen Kernstücks von TWRM-TOPOGRAPHIC und zeigt, wie syntaktisch definierte graphische Objekte im Dialog als informationelle Objekte interpretiert werden. Abschnitt 5 diskutiert die kognitiven und ergonomischen Prämissen, die dem Prinzip des kaskadierten Kondensierens und der flexiblen Dialogführung zugrunde liegen und demonstriert die Systemleistung an einem Beispieldialog. Abbildung 2 zeigt den Zusammenhang der wesentlichen Teile der beiden Systeme TOPIC und TWRM-TOPOGRAPHIC [15] sowie die Struktur der weiteren Darstellung.

Seit Beginn der Arbeit an den Projekten ist zumindest in der Forschung von Information-Retrieval-Systemen einiges in Bewegung gekommen. Wir haben den Eindruck, daß die seit ca. 15 Jahren festzustellende konzeptionelle Stagnation beim Entwurf von Volltextsystemen durch intensive Forschungsaktivitäten auf dem Gebiet intelligenter Information-Retrieval-Systeme überwunden werden kann (vgl. [4]). Auch das System TWRM-TOPOGRAPHIC, über das hier in erster Linie berichtet wird, steht im Zusammenhang des Entwurfs intelligenter Information-Retrieval-Systeme.

2 Textgraphen: Ergebnis der TOPIC-Textanalyse/Textkondensierung und Ausgangsbasis für den TWRM-TOPOGRAPHIC Retrievaldialog

TOPIC [14] analysiert deutschsprachige Texte, vollständige Zeitschriftenartikel aus dem Gebiet der Informations- und Kommunikationstechnologie (aus der VDI-Volltextdatenbank) und überführt thematisch zusammenhängende Textabschnitte in eine Themenbeschreibung, im weiteren auch Textkonstituente genannt. Eine solche Themenbeschreibung ist als hierarchisches Netz aufgebaut, dessen Knoten Frames und, je nach Spezifität, auch *Slots* und *Slot-Entries* zugeordnet sind. Diese repräsentieren die thematisch relevanten Konzepte eines Textabschnitts, deren semantische Beziehungen durch sie verbindende Kanten aufgezeigt werden. Die auftretenden *Frames* können durch die Ober-/Unterbegriff- und Prototyp/Instanz-Beziehung relationiert sein. Ausgehend von derartigen Textkonstituenten wird durch Ableitung weiterer konzeptueller Graphen, die in verallgemeinerter Form die Gemeinsamkeiten der beteiligten Textkonstituenten beschreiben, ein sogenannter Textgraph (das Textkondensat) gebildet, dessen Knoten die Textkonstituenten zugeordnet sind. Die am höchsten liegenden Knoten des Textgraphen enthalten die generalisiertesten und die Blattknoten die spezifischsten Themenbeschreibungen. Die Kanten des Textgraphen zeigen die Abstraktionsbeziehungen an, die zwischen den Textkonstituenten existieren. Folgende vier Kantentypen sind zwischen den Knoten und den ihnen zugeordneten Konstituenten definiert:

Is-a/Instance-Relation:
Eine Kante dieses Typs zwischen einem Knoten n und einem tiefer liegenden Knoten n′ zeigt an, daß die Textkonstituente des Knotens n′ eine Beschreibung eines Unterbegriffs (Is-a-Relation) bzw. einer Instanz (Instance-Relation) des in diesem Fall einzigen Konzepts der Themenbeschreibung des Knotens n enthält.

Slot/Slot-Entry-Relation:
Eine derartige Kante zwischen einem Knoten n und einem tieferen Knoten n′ signalisiert, daß die Themenbeschreibung des Knotens n in derjenigen des Knotens n′ enthalten ist und dessen Beschreibung durch einen zusätzlichen *Slot* bzw. *Slot-Entry* näher spezifiziert ist.

Zur weiteren Erläuterung soll Abb. 3 dienen, die einen Textgraphen-Ausschnitt zeigt, bestehend aus zwei Textkonstituenten, die jeweils einen thematisch zusammenhängenden Textabschnitt beschreiben, und zwei abgeleiteten Textkonstituenten, die die Gemeinsamkeiten der ihnen zugrunde liegenden Textkonstituenten in verallgemeinerter Form wiedergeben.

Die TOPIC-Textgraphen sind zusammen mit dem Volltext in einer Textwissensbasis abgelegt und bilden die Ausgangsbasis sowohl für das graphisch-interaktive Retrieval - aufgrund der Ähnlichkeit zwischen der Suchfrage des Benutzers und der Textrepräsentation werden diejenigen Repräsentationen ausgewählt, deren Textinhalte geeignet scheinen, die Suchfrage zu beantworten - als auch für die situationsspezifische Präsentation von Textwissen auf der vom Benutzer gewählten Kaskadierungsstufe.

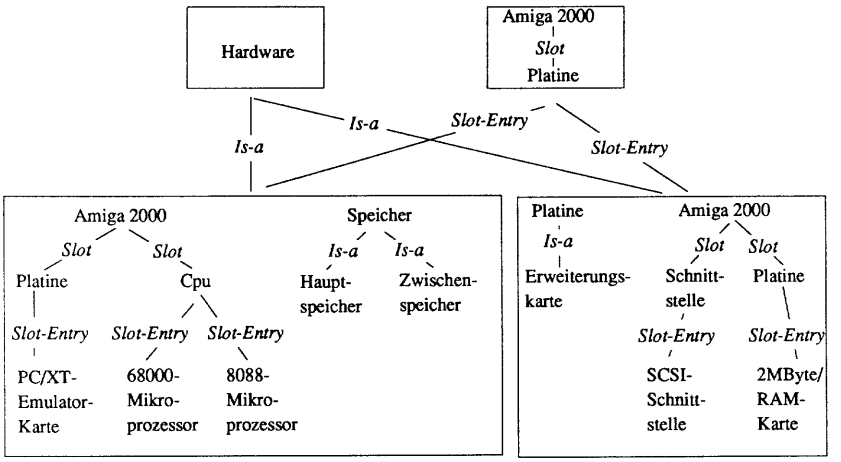

Abb. 3. Beispiel für eine Basis-Textkonstituente

3 Die informationslinguistische Komponente des Systems: Natürlichsprachige Präsentation von Textwissen

3.1 Situationsspezifische Generierung von natürlichsprachigen Abstracts aus den TOPIC-Textgraphen

Die gegenwärtige Beschäftigung mit der natürlichsprachigen Präsentation von Textwissen läßt einen entscheidenden Paradigmenwechsel insofern erkennen (vgl. [21]), als versucht wird, nicht mehr wie bei den früheren statistischen Verfahren aus den Texten direkt, sondern aus semantischen Repräsentationen der Texte Darstellungen der Textinhalte zu erstellen. Für TWRM-TOPOGRAPHIC stellen die vom Textanalyse-/Textkondensierungssystem TOPIC erstellten *semantischen Repräsentationsstrukturen* für Textinhalte (Textgraphen) die Basis für die Präsentation von Textwissen dar. Da die TOPIC-Textgraphen die thematischen Schwerpunkte eines Textes vorwiegend auf indikativem (lediglich anzeigendem) Niveau beschreiben, jedoch auch signifikante Fakteninformation beinhalten können, ist somit die Möglichkeit zur Erzeugung indikativer als auch indikativ-informativer Abstract (Zusammenfassungen)[4] gegeben. Die Hauptforderung, die an diese beiden Abstract-Typen gestellt wird, ist die, den wesentlichen Textinhalt anzuzeigen bzw., im Fall des indikativ-informativen Abstract, auch teilweise wiederzugeben.

Der Forderung, daß Länge, Komplexität und Abstraktionsniveau eines Abstracts sowie der angebotene Inhalt die jeweiligen Benutzerinteressen reflektieren sollen [8] [19] [20], konnte bisher nicht befriedigend entsprochen werden, da Abstracts in der Regel *einmalig* für *einen* bestimmten Zweck und *einen* angenommenen Benutzertyp angefertigt wurden. Mit der Generierungskomponente von TWRM-TOPOGRAPHIC können jedoch aus einem Text (bzw. dessen Repräsentation) *situationsspezifische* Abstracts mit unterschiedlichem Themenschwerpunkt und unterschiedlicher Ausführlichkeit produziert werden. Derart situationsspezifische Abstracts erfordern, im Gegensatz z. B. zum Abstracting-System SUSY [8], in dem der Benutzer Schemata angeben muß, die die Textanalyse und die Erstellung des Abstracts steuern, kein Eingreifen des Benutzers, sondern werden durch Auswertung der Vorgaben der Dialogführung von TWRM-TOPOGRAPHIC produziert.

3.2 Problematik der Generierung natürlichsprachiger Abstracts aus den Repräsentationsstrukturen der TOPIC-Textgraphen

Zwar liegt mit dem TOPIC-Textgraphen bereits eine kondensierte Textrepräsentation vor, doch sind im Textgraphen auch Themenbeschreibungen enthalten, die zwar für einen Textabschnitt, nicht aber für den gesamten Text von zentraler Bedeutung sind, da TOPIC die Relevanz eines Konzepts bezüglich eines Textabschnitts beurteilt. Ziel eines Abstracts dagegen ist, den wesentlichen Textinhalt (mit gegebenenfalls signifikanter Fakteninformation) bereitzustellen; aus diesem Grund müssen zunächst die zentralen Textthemen und die zugehörigen Konzepte identifiziert werden. Diese relevanten *Textthemen* sind je nach Interessenschwerpunkt und gewünschter Ausführlichkeit jedoch nicht für jeden Benutzer gleichermaßen relevant, so daß zur Bestimmung der

[4] Im Bereich der wissensbasierten Sprachverarbeitung wird *Zusammenfassung* meist für die Kurzwiedergabe der wesentlichen Handlung eines Narrativtexts gebraucht. Um Unklarheiten zu vermeiden, wird hier der Begriff *Abstract* bevorzugt

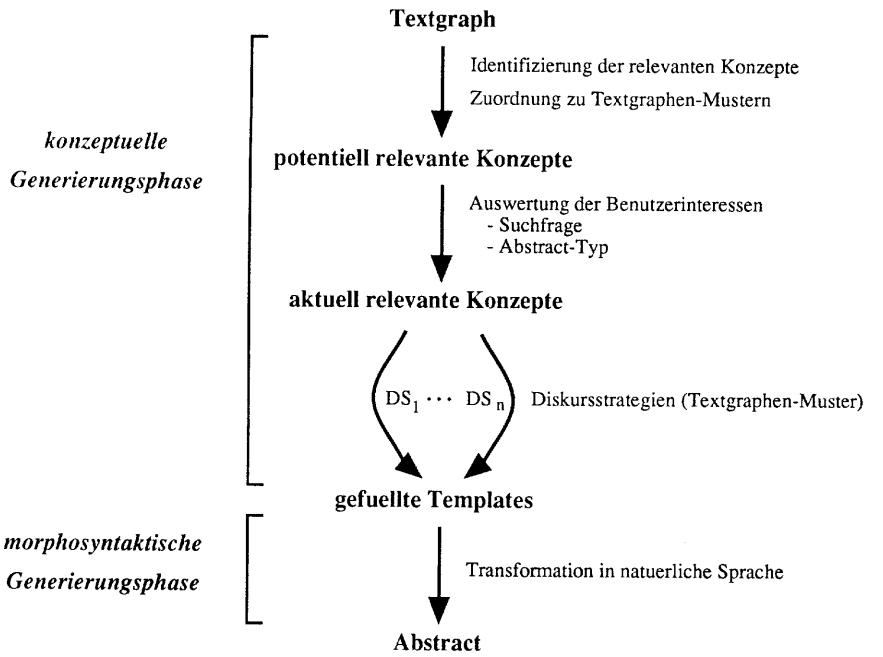

Abb. 4. Textgenerierungs-Modell

aktuell relevanten Konzepte eine weitere Bewertung durch Analyse der Vorgaben der Dialogführung (z. B. der Suchfrage des Benutzers) stattfinden muß.

Nachdem entschieden ist, welche Konzepte im Abstract ausgedrückt werden sollen, muß ein *Textplan* erstellt werden (vgl. zur skizzierten Problematik auch [3] [6] [22] [24]), der festlegt,

- welche Konzepte zusammengehören und in einem Satz ausgedrückt werden sollen,
- welche lexikalischen und syntaktischen Realisierungen am besten geeignet sind, die Beziehung zwischen den Konzepten auszudrücken,
- in welcher Reihenfolge die Sätze angeordnet werden sollen, wie sie zusammenhängen und wie dies verdeutlicht werden kann.

Es muß also die Generierung von zusammenhängendem Text gewährleistet werden, der sowohl die Expansion der einzelnen Textthemen und ihre Abgrenzung gegenüber anderen Themen *(Textkohäsion)* als auch die textuelle Relationierung der Themen *(Textkohärenz)* erkennen läßt (vgl. [17]) und darüber hinaus die speziellen Anforderungen, die an ein Abstract gestellt werden, erfüllt.

3.3 Textgenerierungs-Modell

Die angestrebte Funktionalität erfordert keine vollständig natürlichsprachige Generierung, so daß bei der Generierung vorgefertigte Satzmuster *(Templates)* verwendet werden können, in deren Lücken die ausgewählten Konzepte eingesetzt werden.

Die Textgenerierung wird in zwei Phasen, in eine konzeptuelle und in eine morphosyntaktische Phase, aufgeteilt (vgl. [24]). In der *konzeptuellen Phase* wird entschieden, welche Konzepte aus den Textkonstituenten des Textgraphen im Abstract ausgedrückt, wie sie gruppiert und angeordnet werden sollen, und welche lexikalischen und syntaktischen Realisierungen geeignet sind, diesen Inhalt auszudrücken. Aufgabe der *morphosyntaktischen Phase* ist die Transformation der erarbeiteten Struktur in natürliche Sprache.

Das hier kurz skizzierte Textgenerierungs-Modell legt den Schwerpunkt auf die Erstellung des Textplans, dementsprechend konzentriert sich die weitere Darstellung ganz auf die konzeptuelle Generierungsphase.

3.4 Konzeptuelle Generierungsphase

Im ersten Schritt der konzeptuellen Phase werden die zentralen Textthemen und die zugehörigen Konzepte identifiziert, also diejenigen Konzepte bestimmt, die aufgrund ihrer Bedeutung im Text *potentiell relevant* sind. Danach werden aus ihnen durch Auswertung der Vorgaben der Dialogführung, die die Benutzerinteressen reflektieren, diejenigen bestimmt, die für die gegebene Dialogsituation *aktuell relevant* sind.

Die ausgewählten relevanten Konzepte werden aufgrund ihrer thematischen Relationierung typischen Mustern zugeordnet. Entsprechende *Diskursstrategien* (vgl. z. B. [23] [24]) steuern sowohl die

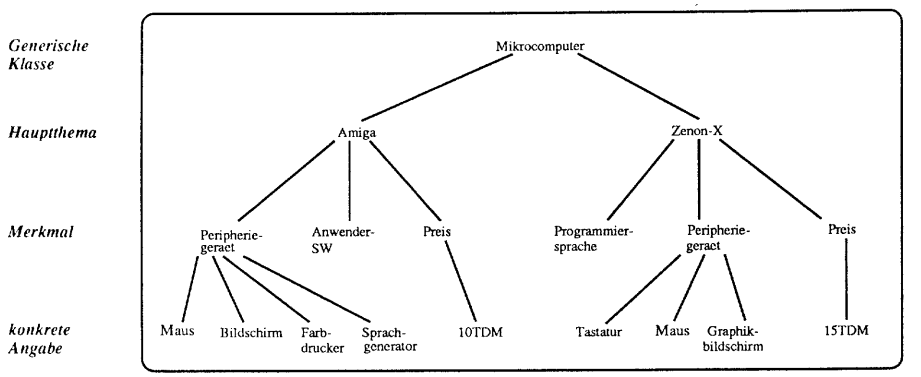

Abb. 5. Beispiel für die Relationierung der potentiell relevanten Konzepte im Textgraphen-Muster „Vergleichende Gegenüberstellung mehrerer verwandter Hauptthemen"

Auswahl und Anordnung der Konzepte im Text als auch das Einsetzen der Konzepte in adäquate Satzmuster.

Die Bestimmung der potentiell relevanten Konzepte beginnt mit der Identifizierung der zentralen Textthemen. Ein Konzept aus einer Themenbeschreibung des Textgraphen soll als ein zentrales Textthema gelten, wenn es im Text in mindestens zwei Abschnitten als thematisch relevant bewertet wurde (also nicht nur von lokaler Bedeutung ist), selbst durch andere thematisch relevante Konzepte näher spezifiziert wird und darüber hinaus genügend spezifisch ist, um Aussagekraft zu besitzen; eine formale Beschreibung der Bedingungen zur Bestimmung eines zentralen Themas enthält [30].

Anschließend werden die zu einem Hauptthema gehörigen Konzepte bestimmt und gemäß ihrer Relation zum Hauptthema unterteilt in:

- Generische Klasse (Prototyp oder Oberbegriff des Hauptthemas)
- Hauptthema
- Merkmal (ein *Slot* des Hauptthemas)
- konkrete Angabe zu einem Merkmal (*Slot-Entry* zu einem Merkmal)

Sind die potentiell relevanten Konzepte des Textgraphen bestimmt und thematischen Blöcken zugeordnet, werden aufgrund der Relationierung dieser Blöcke verschiedene, typische Textgraphen-Muster unterschieden, aus denen sich durch Kombination weitere, komplexere Muster ableiten lassen.

- *Einzelnes Hauptthema:*
Bei der Klassifizierung des Textgraphen wird ein einziges Hauptthema bestimmt, das durch Merkmale und konkrete Angaben zu den Merkmalen näher spezifiziert ist.

- *Mehrere, einfach linear relationierte Hauptthemen:*
Bei diesem Muster werden bei der Klassifizierung des Textgraphen mehrere Hauptthemen bestimmt. Dabei ist eines der näher spezifizierenden Konzepte des einen Hauptthemas ein anderes Hauptthema des Textgraphen, das wiederum näher spezifiziert wird. Gibt es mehr als zwei Hauptthemen, sind sie fortlaufend auf diese Weise relationiert.
- *Vergleichende Gegenüberstellung mehrerer verwandter Hauptthemen:*
Auch hier werden mehrere Hauptthemen bestimmt, die aber in diesem Fall alle derselben generischen Klasse angehören und darüber hinaus auch gemeinsame Merkmale besitzen. Aufgrund dieser thematischen Relationierung kann geschlossen werden, daß die Hauptthemen vergleichend gegenübergestellt werden (vgl. Abb. 5).

Um die für eine gegebene Dialogsituation aktuell relevanten Konzepte bestimmen zu können, werden die Suchfrage *(Query)* des Benutzers in Form semantisch relationierter Konzepte und die optionale Angabe über den gewünschten Abstract-Typ ausgewertet.

- *Abgleich mit der Suchfrage:*
Durch den Abgleich mit der Suchfrage werden aus den potentiell relevanten Konzepten diejenigen ausgewählt, die geeignet sind, die Suchfrage des Benutzers zu beantworten, die anderen Konzepte aber eliminiert. Damit es nicht zu einer Überbewertung der Relevanz des nachgewiesenen Textes kommt, muß jedoch auch der Teil des Textinhaltes angezeigt werden, der keinen direkten Bezug zur Suchfrage hat; d. h. es werden unter Beibehaltung der thematischen Relationierung die gesuchten Textthemen so detailliert wie möglich, die anderen nur so detailliert wie nötig beschrieben. Mit dieser Selektion wird der Grice'schen Quantitätsmaxime [10] entsprochen, so daß die Abstracts in den Benutzerdialog integriert