in-house Literaturdatenbanken

Bequemer und schneller Zugriff über ein Programmpaket zur Literaturverarbeitung und -Verwaltung

In-House Databases for Literature
A Program System for Fast and Conveniend Literature-Search and -Handling

G. GAUGLITZ, Tübingen")

Zusammenfassung

Es werden die modernen Möglichkeiten der Verwaltung und Verarbeitung von Literatur in in-house Datenbanken vorgestellt. Diese Datenbanken dienen zur Ergänzung von kommerziellen, und können eigene Literaturexzerpte, Vortragsabstracts, Manuskripte, Sonderdruckbibliotheken und Diasammlungen enthalten. Dadurch ist eine Kombination der Information gegeben, die einerseits weltweit durch die modernen Zugriffsmöglichkeiten auf kommerzielle Datenbanksysteme erhalten werden können und die andererseits im Hause vorhanden sind oder gesammelt werden sollen. Auf die Leistungsfähigkeit des Programmpaketes wird im Detail eingegangen. Dabei wird insbesondere auf die doppelte Strategie hingewiesen, nicht nur über Indexdateien einen schnellen Zugriff nach Autoren oder Schlagworten zu gestatten, wie dies üblicherweise der Fall ist, sondern auch durch semantische bzw. phonetische Suche tippfehlerbehafte Zitate auch über die Kombination von Suchbegriffen wiederzufinden.

Summary

Modern possibilities of Organization and processing of literature in in-house data bases are discussed. These data bases complete the commercially available ones. They can be used to collect and organize own literature abstracts, reference lists, copies, and slides. Their use allows the combination of Information obtained by worldwide on-line research in commercial data bases and literature at home. The efficiency of the program System LITERA is discussed in detail. Special attention is directed to the double aspect used. Index files allow a very fast approach to author names and special key words as well as a semantic and phonetic search respectively. The later possibility gives the Chance to find even words, names or facts in case they are misspelled.

") Prof. Dr. G. Gauglitz, Institut für physikalische und theoretische Chemie, Auf der Morgenstelle 8, D-7400 Tübingen

Einleitung

In den letzten Jahren haben sich völlig neue Möglichkeiten des Zugriffs auf große, weltweite Datenbanken ergeben. Computerrecherchen solcher maschinenlesbarer Dateien, in denen Literaturstellen in Abstractform, im Volltext oder auch sogar als Strukturen gespeichert sind, können heutzutage aus der modernen Forschung, in der Industrie, an Forschungsinstituten und Hochschulen sowie aus der Lehre nicht mehr weggedacht werden. Dabei bietet einerseits z. B. das FIZ Chemie in Berlin die Möglichkeit der Unterstützung der Formulierung [1] von Suchbegriffen bzw. deren Kombination und der Auftragssuche in Datenbanken an. Andererseits hat der erfahrene Benutzer die Möglichkeit des direkten Zugriffs über STN Karlsruhe auf verschiedene Chemie-, Physik- und Biologie-Datenbanken sowie auf Faktendatenbanken, wie sie z. B. von der Dechema aufgelegt werden. Darüber hinaus stehen auch dem Analytiker - insbesondere in ausländischen Datenbanksystemen - weitgehende Informationsmöglichkeiten offen, die sonst nur mit zeitraubenden Literaturrecherchen in Bibliotheken zu erlangen wären.

Dabei stellt der Anschluß an solche Datenbanken (Host-Rechner) über Datex-P und zum Teil lokale Netzwerke kein größeres Problem dar. In vielen Fällen ist es sogar möglich, daß der Mitarbeiter über ein Terminal auf seinem Schreibtisch direkten Zugriff hat. Allerdings zeigt sich bei der Benutzung dieser Datenbanken, daß es häufig sinnvoll ist, die erhaltenen Datenmengen in-house weiterzuverarbeiten, zu sichten und zu klassifizieren, sowie diese Ergebnisse mit vorhandenen eigenen Literaturzitaten zu vergleichen.

Überlegungen zur Literaturverarbeitung und -Verwaltung

In unserer Arbeitsgruppe wurde seit nahezu zwei Jahrzehnten eine Literaturdatenbank aufgebaut, die zunächst auf Großrechnern installiert war [2]. Dabei stand die Überlegung im Vordergrund, daß bevor noch weltweite Datenbanken zur Verfügung standen jeder Mitarbeiter für Diplom-, Doktorarbeiten oder Publikationen Literaturzitate sammelte, sich entweder Xeroxkopien oder Sonderdrucke beschaffte und häufig auch Seminarmitschriebe oder Vortragsmanuskripte geordnet ablegen wollte. Da die Kenntnis dieser Sammlungen auf den einzelnen Mitarbeiter beschränkt war und häufig schon aus Lesbarkeitsgründen beim Ausscheiden des Mitarbeiters verloren ging, begannen wir frühzeitig eine gemeinsame Datenbank für äffe Mitarbeiter aufzubauen. Dazu war selbstverständlich eine gewisse Formalisierung notwendig. Dies bedeutete zunächst entsprechende zusätzliche Arbeit für den einzelnen Mitarbeiter. Dem stand aber der große Vorteil gegenüber, daß ab sofort alle Daten nicht mehr nur einem zur Verfügung standen, sondern auch andere Zugriff hatten. Außerdem mußten Sonderdrucke nicht mehrfach beschafft werden. Sie wurden auffindbar; ein Vorteil, den jeder zu schätzen wußte, der Wochen auf eine „Fernleihe" im Bibliotheksverkehr gewartet hatte.

Außerdem bot diese Literaturdatenbank einfache Möglichkeiten, Literaturverzeichnisse in größeren Arbeiten zu erstellen. Es gab daher wenig Probleme, die Mitarbeiter vom Sinn dieser Datenbank zu überzeugen und für ihre Mitarbeit zu gewinnen - natürlich mit unterschiedlichem Erfolg. Erfreulich waren die kurzen notwendigen Zeiten, um eine Literaturstelle wiederaufzufinden, obwohl immerhin 5000 bis 10000 Zitate in der Datenbank gespeichert waren. Andererseits zeigte sich aber ein Problem, daß nämlich die Wiederfindungsrate zunächst sehr gering war. Die Ursache bestand natürlich darin, daß beim Eintippen des Datenmaterials die Fehlerzahl endlich blieb und nicht völlig beseitigt werden konnte.

Daher wurden von uns zunächst mit Unterstützung der Dokumentationsgruppe des Zentrums für Datenverarbeitung in Tübingen Maschinenunterprogramme entwickelt, die einen Ähnlichkeitsvergleich zwischen Suchwort und dem Stichwort im Text der Literaturstelle gestatteten. Durch Eingabe einer Fehlerschranke konnten auch tippfehlerbehaftete Worte wiedergefunden werden. Da bei der geringen Größe der Datenbank die Suchzeit nur eine untergeordnete Rolle spielte, war hier die Kanalzeit nach Eingabe zunächst über Lochkarte,

später über das Terminal der zeitlich limitierende Faktor.

Im Laufe der Zeit wuchs einerseits die Anzahl der Literaturzitate und der aufgenommenen sonstigen Einträge. Andererseits ergaben sich neue Möglichkeiten durch die Einführung von Minicomputern. Daher wurden von uns die zunächst für die Großrechner CDC3300 und TR 440 geschriebenen Programme [3] auf einen Minicomputer Tektronix4051 übertragen. Während bei diesem Rechner die Daten noch sequentiell auf einem Magnetband abgespeichert waren, wurden in der nächsten Generation von Kleinrechnern mit den Prozessoren TI990 und Motorola 68000 Möglichkeiten der Speicherung auf Diskette bzw. Winchester eröffnet. Dabei ist die inzwischen auf 25000 Stellen angewachsene Datenbank in Volltext geschrieben. Sie enthält aber zusätzlich noch eine Anzahl von Indexdaten, die den Speicherbedarf von ca. 7 MByte verdoppeln.

Da nach einem Vortrag [4] das Interesse an dieser in-house Literaturverarbeitung sehr groß war, haben wir in neuerer Zeit das gesamte Programmpaket auf IBM bzw. IBM kompatible Rechner umgeschrieben. Dabei wurden die Möglichkeiten der Ähnlichkeitssuche beibehalten, allerdings noch zusätzlich eine Schnellsuche nach Autoren oder speziellen Schlagworten über Indexdateien integriert.

Programmpaket LITERA

Bei LITERA handelt es sich um ein Programmpaket zur Verwaltung und Verarbeitung von einzelnen Literaturstellen, die aus Autoren, Titeln, Schlagworten (Thesaurus, Bibliographie), Kommentaren und/oder Kurzbezeichnungen bestehen. Im Gegensatz zu üblichen Literaturdatenbanken beschränkt sich das Programmpaket nicht nur auf Schlagwortverzeichnisse bzw. auf Thesauruseinträge, sondern es bietet darüber hinaus die Möglichkeit, nach jedem Wort der Literaturstelle abzufragen. Dies kann semantisch geschehen, d. h. bei der Abfrage müssen Suchwort und Text in der Literaturstelle nur innerhalb wählbarer Grenzen ähnlich sein. Diese Abfrage geschieht in diesem Fall als Ähnlichkeitssuche über Korrelationsmatrizen. Dadurch können auch solche Worte gefunden werden, die entweder mit Fehlern eingegeben wurden oder die mit dem abzufragenden Stichwort nur orthographisch bzw. semantisch ähnlich sind. Das Programmpaket stellt eine Kombination aus zwei völlig entgegengesetzten Philosophien bei einer Literatursuche dar:

- Einerseits ist es möglich, in einer ***Schnellsuche*** nach Autoren oder speziell gewählten und voreingegebenen Schlagworten über erstellte Indexdateien abzufragen. Dazu ist es aber Voraussetzung, daß die Literaturstellen mit einer begrenzten Zahl von fehlerfreien Schlagworten indiziert worden sind, und die Autoren ebenfalls als Begriffe in Indexdateien eingeordnet wurden. Diese Form der Suche stellt den üblichen Thesaurusteil von Datenbanken dar.
- Andererseits kann in einer *Ähnlichkeitssuche* jedes Wort der Literaturstelle nach Suchworten bzw. Suchwortkombinationen abgefragt werden. Da das Ergebnis nicht über ein Indexregister geschieht, sondern in Echtzeit durchgeführt werden muß, ist natürlich das Verfahren sehr zeitaufwendig. Trotzdem kann das Ergebnis auch bei großen Datenbanken (25000 bis 40000 Literaturstellen) über Nacht auf eine Datei geschrieben und durch schnelles Blättern am nächsten Morgen nach den interessierenden Stellen ausgewählt werden. Dies wird als *semantische bzw. orthographische Ähnlichkeitssuche* bezeichnet.

In Abb. 1 ist ein Beispiel für eine Literaturstelle, die in Volltextdateien steht, wiedergegeben. Sie besteht aus mehreren Zeilen, die verschiedene Bedeutung haben. Sie werden durch ein Symbol in der ersten Spalte einer jeden Zeile charakterisiert.

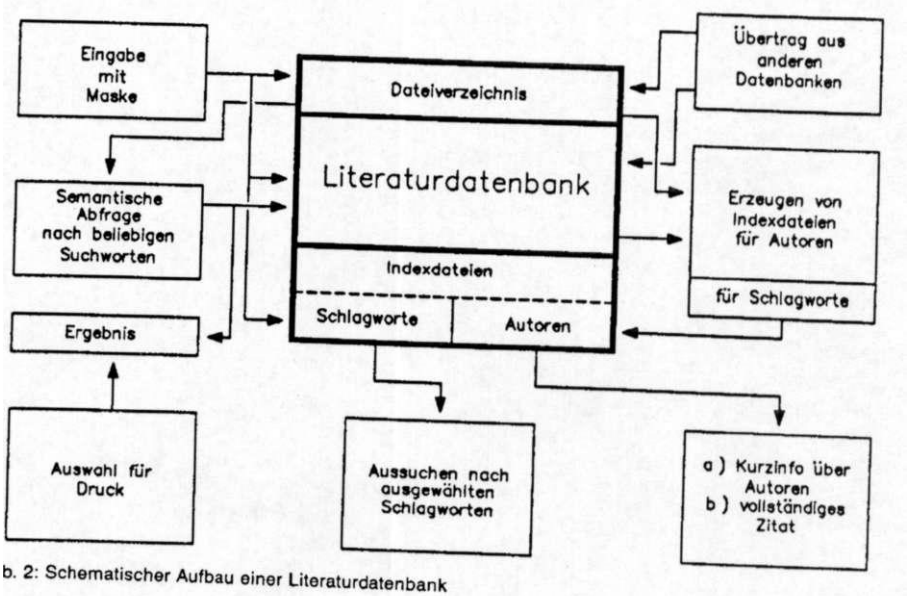
```
170315
$ Schmidt H N  Müller R G W  Hinterhuber H
  Der Einsatz eines Rechne« für die optimale Literatur«
  Verarbeitung, eingesetzt in Großrechnern, mit Erfolg
  umgeschrieben auf ein P-DOS-System und jetzt
  angepaßt an MS-DOS-Systerae
• Xerox 1273
# Dia COM-LI-1
+ Datenverabeitung Mikro Prozesor Thesaurus
  Phonetisch Abfrage belibige Stichwörte
% Literatur, EDV
= J. Comp. Sc. 55, 3423 (1982).
! Neueste Version in Vorbereitung
&

Die vorhandenen Tippfehler sind beabsichtigt.
```

Abb. 1: Muster einer Literaturstelle

In Abb. 2 ist der Aufbau der Literaturdatenbank schematisch dargestellt. Sie enthält neben den Volltextdateien außerdem Dateien für Schlagworte und Autoren als Indexregister. Aus diesen kann jeweils sofort die gewünschte Information erhalten werden. Die dazu notwendigen Indexdateien werden von Zeit zu Zeit korrigiert und in einem Nachtlauf auf den neuesten Stand gebracht. Allerdings wird auch bei jeder manuellen Eingabe von Literaturstellen sowohl die Autoren- als auch die Schlagwortindexdatei korrigiert. Darüber hinaus können die Volltextdateien nach beliebigen Suchworten bzw. Suchwortkombinationen abgefragt werden. Das Ergebnis wird dann auf dem Bildschirm „vorgeblättert", und es können die interessierenden Stellen zum Druck ausgewählt werden, da wegen der sinnvollerweise großzügigen Fehlertoleranz neben den gewünschten Stellen noch je nach Geschick der Frageformulierung eine größere Anzahl von weiteren Stellen gefunden wird.

Dabei muß zwischen den Begriffen Such-, Schlag- und Stichworte unterschieden werden. Suchworte sind Terme der Abfrageformulierung, fehlerfreie Schlagworte können eingegeben werden, um das Literaturzitat zu charakterisieren. Unter Stichworten sollen Begriffe verstanden werden, die zusätzlich zur Überschrift und gegebenenfalls zur Zusammenfassung aufgenommen wurden, um den Informationsgehalt zu vergrößern. Außer den Schlagworten können alle „Worte" der Literaturstelle tippfehlerbehaftet sein. „Fehlerhafte" Autoren können allerdings nicht über die Indexdateien, sondern nur über die schematische Abfrage gefunden werden.

Die Eingabe einer Literaturstelle kann von Hand über eine Maske oder aber auch durch Übertragung aus schon vorhandenen oder erzeugten Datenbanken geschehen. Dadurch ergibt sich die Möglichkeit, sich lokal (in-house) eine persönliche Literaturdatenbank zu halten, die den schnellen Zugriff auf eigene interessie-



b. 2: Schematischer Aufbau einer Literaturdatenbank

rende Stellen gestattet. Ein wesentlicher Vorteil des Konzepts ist sicherlich die Ähnlichkeitssuche, die hilft, Informationsverluste zu vermeiden, die auch in großen, mit viel Aufwand geprüften Datenbanken durch Tippfehler auftreten können.

In jeder Arbeitsgruppe liegt eine Vielzahl von fertigen Manuskripten, Sonderdrucken, Kopien, Abbildungen oder Dias vor. Alle diese vorhandenen Daten können mit in die Literaturdatenbank integriert werden. Numeriert man z. B. alle Sonderdrucke und Xeroxkopien mit einem Nummernstempel durch, so kann man sich in der Literaturdatenbank auf diese Nummer beziehen. Der Suchende weiß, daß er sich diese Stelle nicht erst in der Bibliothek oder über Fernleihe beschaffen muß. Er kann sie z. B. aus Aktenordnern entnehmen, in denen sie mit aufsteigender Nummer eingeordnet sind. Analoges gilt für die schnelle Suche nach einzelnen Dias zu einer speziellen Problemstellung. Darüber hinaus bietet die Einordnung in die Datenbank und die Zuordnung von einer größeren Zahl von Stichworten die Möglichkeit, Mitschriebe auf Karteikarten unter verschiedenen Gesichtspunkten schnell zu finden, wobei nicht für jedes Schlagwort eine Karteikarte angelegt werden muß.

Das Programmpaket kombiniert den schnellen Zugriff über Indexkarteien mit dem Ähnlichkeitsvergleich über semantische Abfrage mit geringer Zugriffsgeschwindigkeit und hoher Wiederfindungsrate. Es gestattet eine Ordnung von Sonderdrucken und Dias. Es kombiniert Index- mit Volltextdateien. Es ist somit eine geeignete Lösung für die Verwaltung auch größerer in-house Literaturdatenbanken, für die in unserer Arbeitsgruppe eine langfristige Erfahrung besteht.

Die Installation des Programmpaketes [5] geschieht von einer Diskette auf eine Festplatte Winchester in wenigen Minuten. Das Programmpaket ist menügesteuert. Obwohl ein ausführliches Handbuch vorhanden ist, kann bei jeder Eingabe eine Informationsdatei aufgerufen werden. Dadurch wird die Programmführung weitgehend selbsterklärend. Die Datenbank ist aus der Erfahrung des Naturwissenschaftlers und aus den Notwendigkeiten des täglichen Betriebes im Labor heraus entstanden. Sie ist sicherlich von gleichem Interesse für Hochschul- und Forschungsinstitute sowie für Labors in der Industrie.

Literatur

[1] KRIETSCH, W.: Computerrecherchen nach chemischen Stötten oder Stoffklassen, GIT Fachz. Lab. 1/88, S.20 (1988)

[2] NIEMANN, H. J.: CDC 3300 im Jahre 1971

(3] GAUGLITZ, G.: CDC 3300 (1973), TR 440 (1975)

[4] GAUGLITZ, G.: Vortrag auf der 3. DASp-Diskussionstagung „Praktische Molekülspektroskopie' in Dortmund, September 1986

[5] Attempto Verlag der Universität Tübingen (Wilhelmstraße 5, D-7400 Tübingen), Programmpaket „LITERA", Anfrage an Herrn Dr. FUNKE, Tel. *07071/29025030* oder 296473