Kapitel 18: Semistrukturierte Web-Daten

Im Gegensatz zu reinen Textdokumenten sind die Daten des World-Wide-Web (WWW) insofern *semistrukturiert*, als daß über die Textinhalte von HTML-Dokumenten (Hypertext Markup Language) hinaus auch explizite Referenzen (Links) und Markierungen (Tags, Labels) für Suchoperationen ausgenutzt werden können. Mit dem neuen XML-Standard (Extensible Markup Language) wird der strukturierte Anteil sogar weiter ausbaubar. Anfragesprachen für die Daten des WWW sollten daher Suchmöglichkeiten des Information Retrieval mit Elementen konventioneller Datenbanksprachen und der Suche in Graphen kombinieren.

18.1 Modellierung semistrukturierter Web-Daten

Die Daten des WWW können als eine (riesige) Menge von HTML-Dokumenten angesehen werden; künftig wird dies eine Menge von XML-Dokumenten sein. XML ist ebenso wie HTML aus der älteren Dokumentenbeschreibungssprache SGML (Standard Generalized Markup Language) abgeleitet. Im Gegensatz zu HTML wird in XML die strukturelle Information zu einem Dokument nicht mehr mit der Präsentationsinformation vermischt, und XML erlaubt durch explizite benannte *Elemente* bzw. *Attribute* eine reichere Beschreibung der Struktur eines Dokuments. Im Gegensatz zu konventionellen Datenbanken kann diese Beschreibung in jedes Dokument individuell eingebettet werden; es gibt also nicht notwendigerweise ein gemeinsames Schema mehrerer Dokumente. Allerdings ist es auch möglich, mit einer sogenannten DTD (Document Type Definition) ein Strukturschema für Dokumente vorzugeben. Darüber hinaus können mit Hilfe von Gateways zwischen DBS und WWW (z.B. Produkten wie dem Oracle Web Application Server) "virtuelle" XML-Dokumente dynamisch generiert werden, deren Inhalte und Struktur sich auf dahinterliegende Datenbanken beziehen.

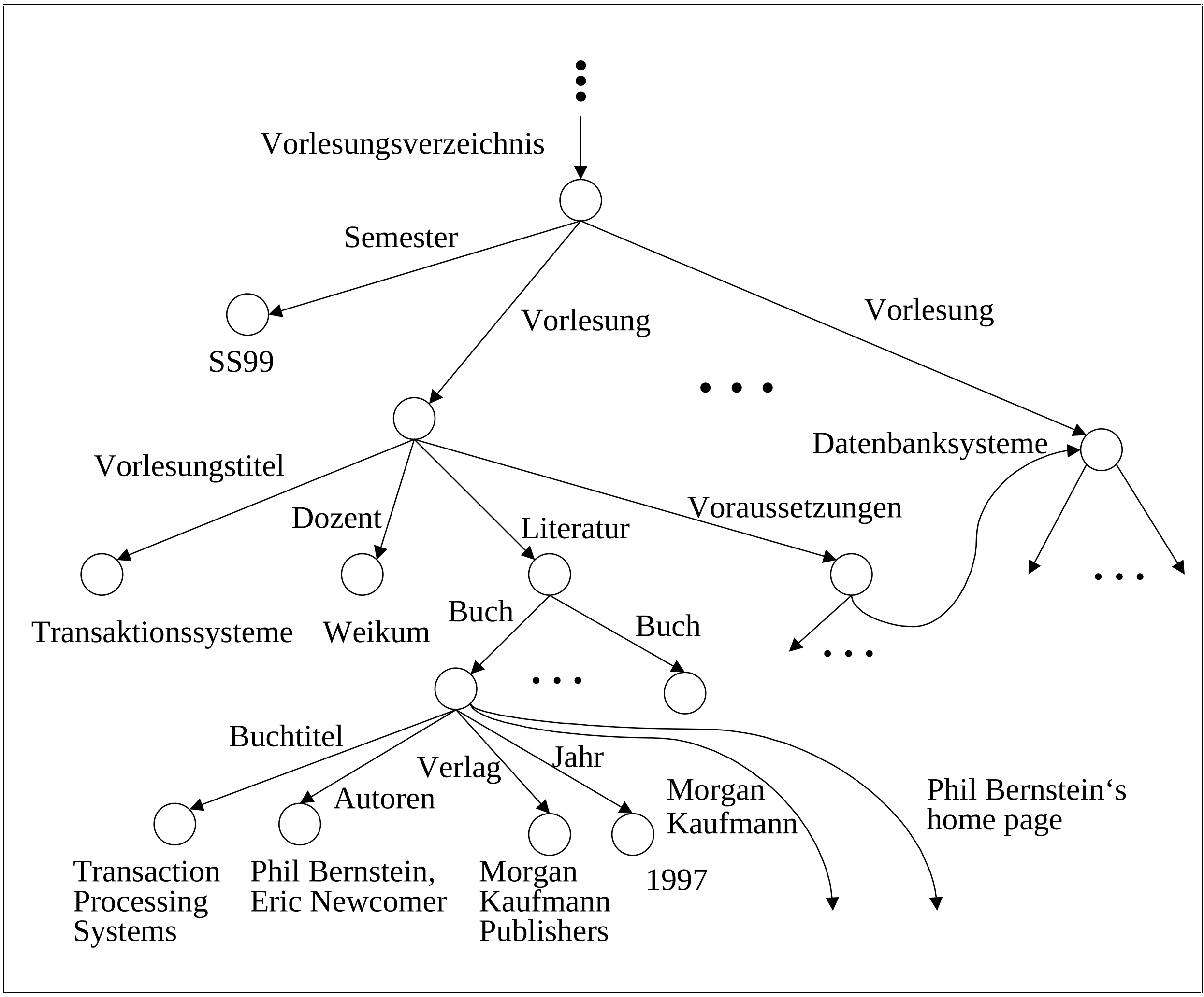
Beispiel eines XML-Dokuments:

```
<?XML ...>
<Vorlesungsverzeichnis>
<Semester> SS99 </Semester>
<Vorlesung>
<Vorlesungstitel> Transaktionssysteme </Vorlesungstitel>
<Dozent> Weikum </Dozent>
<Literatur>
<Buch
Buchtitel="Transaction Processing Systems"
Autoren="Phil Bernstein, Eric Newcomer"
Verlag="Morgan Kaufmann Publishers"
Jahr="1997">
<child href="http://www.mk.com" name="Morgan Kaufmann"/>
<child href="http://research.microsoft.com/philb" name="Phil Bernstein's home page"/>
</Buch>
<Buch
...
</Buch>
</Literatur>
<Voraussetzungen>
 <child href= "www.cs.uni-sb.de/Stammvorlesungen/DBS" name="Datenbanksysteme"/>
</Voraussetzungen>
</Vorlesung>
<Vorlesung>
<Vorlesungstitel> Multimedia </Vorlesungstitel>
...
</Vorlesung>
...
</Vorlesungsverzeichnis>
```

Formal ist ein XML-Dokument ein Baum von Elementen, die einen Namen haben und eine Menge von Attributen sowie eine (geordnete) Liste von Referenzen ("Kindern") beinhalten. Ein Attribut ist wiederum ein Paar bestehend aus einem Attributnamen und einem Datenwert.

Vereinfacht kann ein XML-Dokument somit als Baum modelliert werden, dessen Kanten mit den Namen der Elemente bzw. Attribute beschriftet sind und dessen Blattknoten die entsprechenden Datenwerte angeben. In Verbindung mit den benannten Links zwischen Dokumenten ergibt sich dann insgesamt ein Graph mit Werten als Knoten und Namen für Kanten. Ein so aufgebautes abstraktes Datenmodell für das WWW ist z.B. das Object Exchange Model (OEM).

Beispiel eines OEM-Graphen:



18.2 Anfragesprachen für semistrukturierte Daten

Anfragesprachen für OEM-artige Datenmodelle kombinieren prädikatenlogik-orientierte Filterbedingungen mit Suchbedingungen auf Graphen, z.B. regulären Ausdrücken für Kanten-Labels. Dies führt insbesondere auf Verallgemeinerungen der Pfadausdrücke, wie sie von OQL bekannt sind.

Beispiel:

Finde - ausgehend von einer Menge von Wurzel-URLs (die z.B. über eine vorhergehende IR-Suche oder durch Zugriff auf eine Web-Katalog bestimmt wurden) - alle Datenwerte, die den String "Transaction Processing" enthalten und über einen Pfad erreichbar sind, auf dem mindestens einmal die Labels "Vorlesung" und "Buch" auftreten.

```
Select Value
From Roots.(*.Vorlesung.*.Buch.* Or *.Buch.*.Vorlesung.*).Value
Where Value Contains "Transaction Processing"
```

Verschiedene Anfragesprachen für XML stehen zur Diskussion für die Standardisierung. Es wird erwartet, daß mit einer solchen Sprache traditionelle Web-Daten und Daten, die von Datenbanksystemen mit Web-Anbindung verwaltet werden, einheitlich abfragbar sind. Dazu müssen künftige Datenbanksysteme XML-Views auf ihre Daten unterstützen bzw. XML-Suchoperationen in lokale SQL-Operationen übersetzen. Komplementär dazu wird es auch möglich sein, beliebige XML-Dokumente in einer Datenbank abzulegen.

Ergänzende Literatur zu Kapitel 18

D. Florescu, A. Levy, A. Mendelzon, Database Techniques for the World-Wide Web: A Survey, ACM SIGMOD Record Vol. No. , 1998

IEEE Data Engineering Bulletin Vol.21 No.2, June 1998, Special Issue on Databases and the World Wide Web

S. Abiteboul, Querying Semi-Structured Data, International Conference on Database Theory (ICDT), Dephi, Greece, Lecture Notes in Computer Science, Springer-Verlag, 1997

offizielle Dokumente zu XML, http://www.w3.org/XML

diverse Dokumente zu XML, http://www.xml.com

XML Support in Oracle8i and Beyond, Whitepaper, http://www.oracle.com/xml/documents