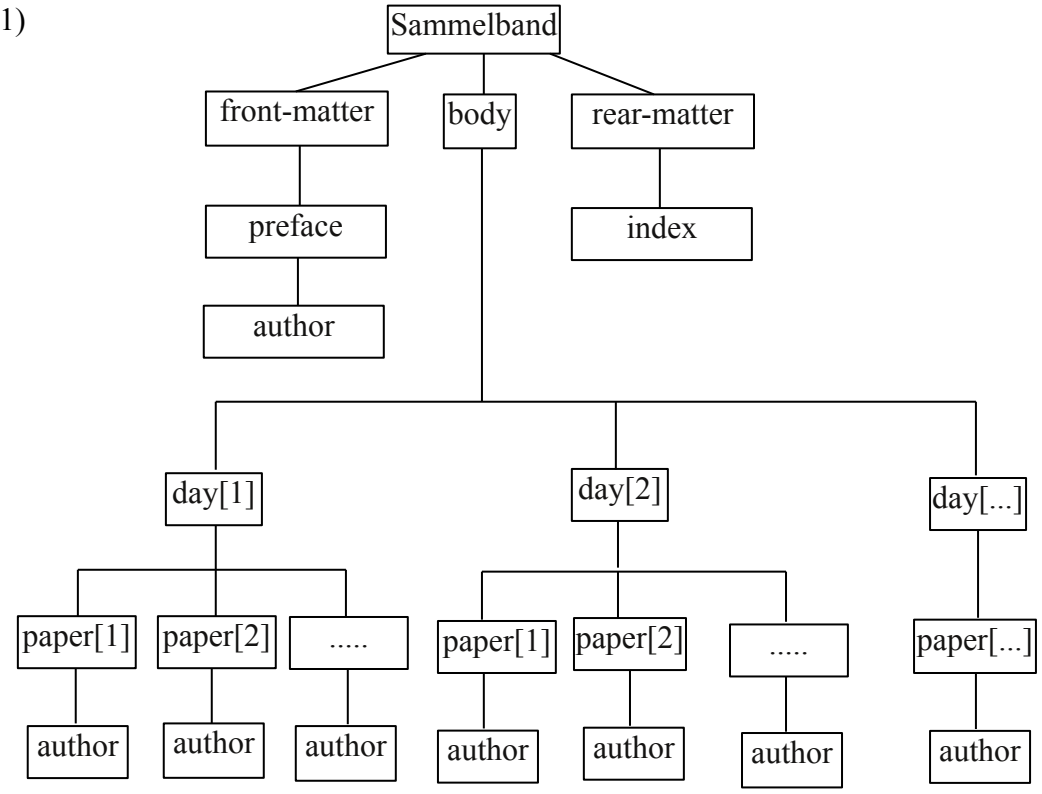
X-path

Ein Sammelband ist folgendermaßen aufgebaut: Die Wurzel enthält drei Elemente **front-matter, body** und **rear-matter. front-matter** enthält ein Vorwort **preface, body** enthält die Beiträge der Teilnehmenden als **paper** Element, wobei die Beiträge eines Tages jeweils in einem **day[X]** Element zusammengefasst sind (X ist die Nummer des Tages). **rear-matter** enthält ein Index **index.** Die Autoren aller Beiträge stehen in Vorwort und in den Beiträgen als **author** Element.

- wie sieht die Struktur vom Dokument aus?
- Mit welchem Pfad können folgende Ausdrucke erfasst werden?

a) der Autor des Vorworts
b) die Autoren der Beiträge
c) der Autor des zweiten Tages
d) alle Autor-Elemente

1)



2)

a) der Autor des Vorworts: *Sammelband/front-matter/preface/author*
b) die Autoren der Beiträge: *Sammelband/body/*/*/author*
Sammelband/body/ /author
(Der leere Eintrag „/ /“ symbolisiert beliebig tiefe Unterelemente, Der Eintrag /*/ beschreibt genau ein beliebiges Element)
c) der Autor des zweiten Tages: *Sammelband/body/day[2]/*/author*
d) alle Autor-Elemente: *Sammelband/ /author*

structural term

Wie viele strukturelle Terme gibt in Abbildung 10.1? (Manning Übungsaufgabe 10.3)

Ein strukturierter Term ist ein Xpath-Pfad, der mit einem einzelnen Vokabular beendet.
Angenommen ist die Größe für das Element „verse“ 6 Terme(„Will I with wine and wassail“). Die

Größe für den Text „Macbeth´s castle“ ist 2 Terme (Macbeth, Castle). Deshalb ist die Anzahl der strukturellen Terme:
3: Shakespeare, author/Shakespeare, play/ author/Shakespeare
3: Macbeth, title/Macbeth, play/ title/Macbeth
5: Macbeth, title/Macbeth, scene[vii]/title/Macbeth, act[I]/scene[vii]/title/Macbeth, play/act[I]/scene[vii]/title/Macbeth
5: castle, title/castle, scene[vii]/title/castle, act[I]/scene[vii]/title/castle, play/act[I]/scene[vii]/title/castle

5*6 für die Terme im „verse“
5: will, verse/will, scene[vii]/verse/will, act[I]/scene[vii]/verse/will, play/act[I]/scene[vii]/verse/will
u.s.w.
Die strukturierten Terme: 3+3+5+5+30= 46

Vektorraummodell

1) Beschreiben Sie kurz in eigenen Worten die wesentlichen Unterschiede zwischen einem Vektorraummodell das boolesches Retrieval implementiert und für XML Retrieval?
    - Ein Vektorraummodell für boolesches Retrieval betrachtet nur den textuellen Inhalt von Dokumenten. Es benutzt zum Ranking nur den Inhalt eines Dokuments.Die Darstellung des booleschen Retrieval im Vektorraummodell ist allerdings nur für einfache Anfragen möglich, bei denen alle Terme mit AND oder OR verknüpft sind. Anfragen, die komplex geschachtelt sind, lassen sich nicht einfach implementieren.
    - Im Gegensatz dazu nutzt das Modell in XML-Dokumenten enthaltet Strukturinformationen. Es ermöglicht, in Suchanfragen zusätzlich zu inhaltlichen auch strukturelle Bedingungen darzustellen. In diesem Modell wird die Häufigkeit des Auftretens von strukturierten Termen in strukturierten Dokumenten für das Ranking verwendet. Die Hoffnung ist dass durch die Ausnutzung der Strukturinformation die Qualität der Antworten verbessert wird und die Benutzer gezieltere Queries formulieren. Schwierigkeiten verursacht u. a. der unklare Dokumentenbegriff (was soll eigentlich zurückgegeben werden), was ist Grundlage der Berechnung der Gewichte, insbesondere für idf. Ausserdem ist die Formulierung der Queries aufwändig und benötigt ein spezielles Interface.

2) Geben Sie ein Beispiel für ein Query-Dokument Paar, dessen SIMNOMERGE (q, d) größer als 1,0 ist. (Manning Übungsaufgabe 10.11)

Beispielsweise besteht ein Query-Dokument Paar aus dem einzelnen Term „Gates“ sowohl in der Anfrage als auch im Dokument. Deshalb: CR(q,d) = 1.0, Gewicht des Terms in Anfrage weight(q, t, c) ist größer als 1, als wenn der IDF-Wert von Dokumente größer als 1 ist. Das normalisierte Gewicht des Terms im Dokument liegt bei 1. So SIMNOMERGE (q, d) ist größer als 1.