4.3 Module Parser & Scanner

Đây là module làm việc trực tiếp với tệp truyện, và có thể nói là quan trọng nhất trong toàn ứng dụng. Mạch tư duy của toàn bộ phần này như sau:

1. **Yêu cầu**: đọc tệp nén ngẫu nhiên (sẽ giải thích ở Mục 4.3.1)
2. **Thuận lợi**: tệp nén ZIP có Central Directory, do đó có thể "nhảy cóc" (đã giải thích ở Mục 2.5.1)
3. **Khó khăn**: cần dùng SAF, do đó chỉ có thể đọc tuần tự (đã giải thích ở Mục 2.1.2)
4. **Giải pháp**: tìm offset ứng với mỗi tệp ảnh trong tệp ZIP và nhảy cóc đến đó khi cần (sẽ giải thích ở Mục 4.3.1)

4.3.1 `ComicParser`

`ComicParser` là một trong các thành phần trung tâm của yacv. Lớp này nhận vào URI trỏ đến một tệp truyện, và đọc nội dung tệp truyện đó ra. Cần gọi đủ tên là `ComicParser`, vì phải phân biệt với hai parser (bộ đọc/giải mã) khác nhưng tích hợp trong nó:

Bảng 4.2: Hai kiểu parser trong `ComicParser`

	Parser cho tệp nén	Parser cho metadata
Đầu vào	Tệp nén	Tệp metadata
Đầu ra	Các tệp con hoặc tương đương (`InputStream`,…)	Đối tượng `Comic`

Parser cho tệp nén

Parser cho tệp nén hiện gồm một giao diện và hai lớp:

- `ArchiveParser`: giao diện chung cho mọi parser tệp nén
- `CBZParser`: parser riêng cho tệp CBZ
- `ArchiveParserFactory`: giúp khởi tạo các parser

`ArchiveParser` là một giao diện (interface), định nghĩa một số phương thức chung mọi parser cho tệp nén đều phải có. Do hiện tại yacv mới hỗ trợ định dạng CBZ, chỉ có lớp `CBZParser` cài đặt giao diện này.

Trong `ArchiveParser`, có hai phương thức quan trọng:

- `getEntryOffsets()`: Phương thức này trả về một từ điển như sau:
  - Khóa: tên tệp lẻ
  - Giá trị: offset tệp lẻ, tức vị trí tệp lẻ trong tệp nén
- `readEntryAtOffset()`: Phương thức này nhận vào một offset, và trả về `InputStream` tương ứng với tệp lẻ ở offset đó bằng cách "nhảy cóc" đến đúng chỗ và đọc.

Ý tưởng cho thiết kế này xuất phát từ yêu cầu phải *đọc ngẫu nhiên* tệp nén. Như đã giải thích trong Mục 2.5.2, tệp ảnh trang truyện không bắt buộc lưu theo thứ tự nào trong tệp truyện nén. Ví dụ, tệp `2.jpg` được lưu trước tệp `1.jpg`, trong khi người dùng cần đọc trang 1 trước trang 2. Do đó, để người dùng được đọc *tuần tự*, ứng dụng phải đọc *ngẫu nhiên*.

Từ yêu cầu đọc ngẫu nhiên trên, ta thấy ngay thiết kế của `ArchiveParser` như trên là đơn giản và rõ ràng nhất.

`ArchiveParserFactory` là một lớp theo mẫu thiết kế factory, nhận vào URI của tệp truyện và trả về `ArchiveParser` để đọc loại tệp truyện đó (ví dụ, nếu URI có đuôi CBZ thì trả về một đối tượng `CBZParser`).

Parser cho metadata

yacv hiện hỗ trợ định dạng ComicRack, được giới thiệu chi tiết trong Phụ lục 2. Định dạng này là một tệp tin XML, do đó được đọc đơn giản bằng các thư viện XML sẵn có.

Để mở rộng định dạng tệp đọc, có thể dùng mẫu thiết kế factory như đã dùng với parser cho tệp. Theo cách này, các parser cần có hàm `parse()` trả về một đối tượng `Comic` và nhận hai tham số:

- Nội dung tệp metadata: ở dạng chuỗi thông thường
- Tên tệp metadata: tên tệp giúp phân biệt các định dạng tệp với nhau

4.3.2 `CBZParser`

`CBZParser` là một parser cho tệp nén CBZ/ZIP, cài đặt giao diện `ArchiveParser`. Như đã phân tích ở Chương 2, hệ thống đọc ghi tệp SAF của Android chỉ cho phép đọc ghi tuần tự. Việc tạo ra mảng offset không đơn giản, do phần mục lục của tệp ZIP nằm ở cuối, và có nhiều thao tác cần dò ngược từ cuối lên.

Để giải quyết vấn đề danh sách offset, có hai cách đơn giản nhất:

- Chép tệp truyện vào bộ nhớ riêng của ứng dụng, rồi xóa khi đọc xong

  - Ưu: Phần bộ nhớ này vẫn được dùng API File của Java, do đó có thể đọc ghi ngẫu nhiên, cho phép đọc mục lục rất nhanh.
  - Nhược: Các nhược điểm của việc dùng bộ nhớ riêng đã trình bày ở đầu chương.

- Đọc tệp ZIP ở chế độ đọc tuần tự
  - Ưu: Dùng ngay được với cơ chế đọc qua `InputStream` của SAF
  - Nhược: Do không có mục lục, dữ liệu phải được "dò" từ từ. Hậu quả là vừa tốn băng thông đọc, vừa tốn CPU để giải nén những tệp không cần thiết.

`CBZParser` giải quyết vấn đề này bằng cách làm giả một luồng nhập ngẫu nhiên, sẽ được miêu tả rõ hơn trong Mục 5.1.2. Cách làm đó có thể được tóm tắt như sau:

- Hai phần đầu tệp nén được lưu đệm trong RAM, do là hai phần có nhiều truy cập nhất trong khi đọc mục lục
- Các phần còn lại được đọc xuôi khi cần theo luồng nhập `InputStream`, nếu đọc ngược sẽ phải tạo mới luồng nhập

Kết quả là mục lục đọc được mà chỉ cần:

- Trung bình hai lần đọc tuần tự theo `InputStream`
- Không phải ghi ra đĩa
- Không phải giải nén những tệp không cần thiết

4.3.3 Tổng hợp lại `ComicParser`

Tương tác trong một ca sử dụng hiển thị của `ComicParser` được mô tả trong Hình 4.14.

Nhắc lại rằng do metadata không chỉ định, thứ tự trang truyện chỉ có thể suy ra từ thứ tự tên tệp ảnh. Không có quy chuẩn cho tên trang truyện, tuy nhiên đa số các tệp truyện đặt tên theo định dạng sau:

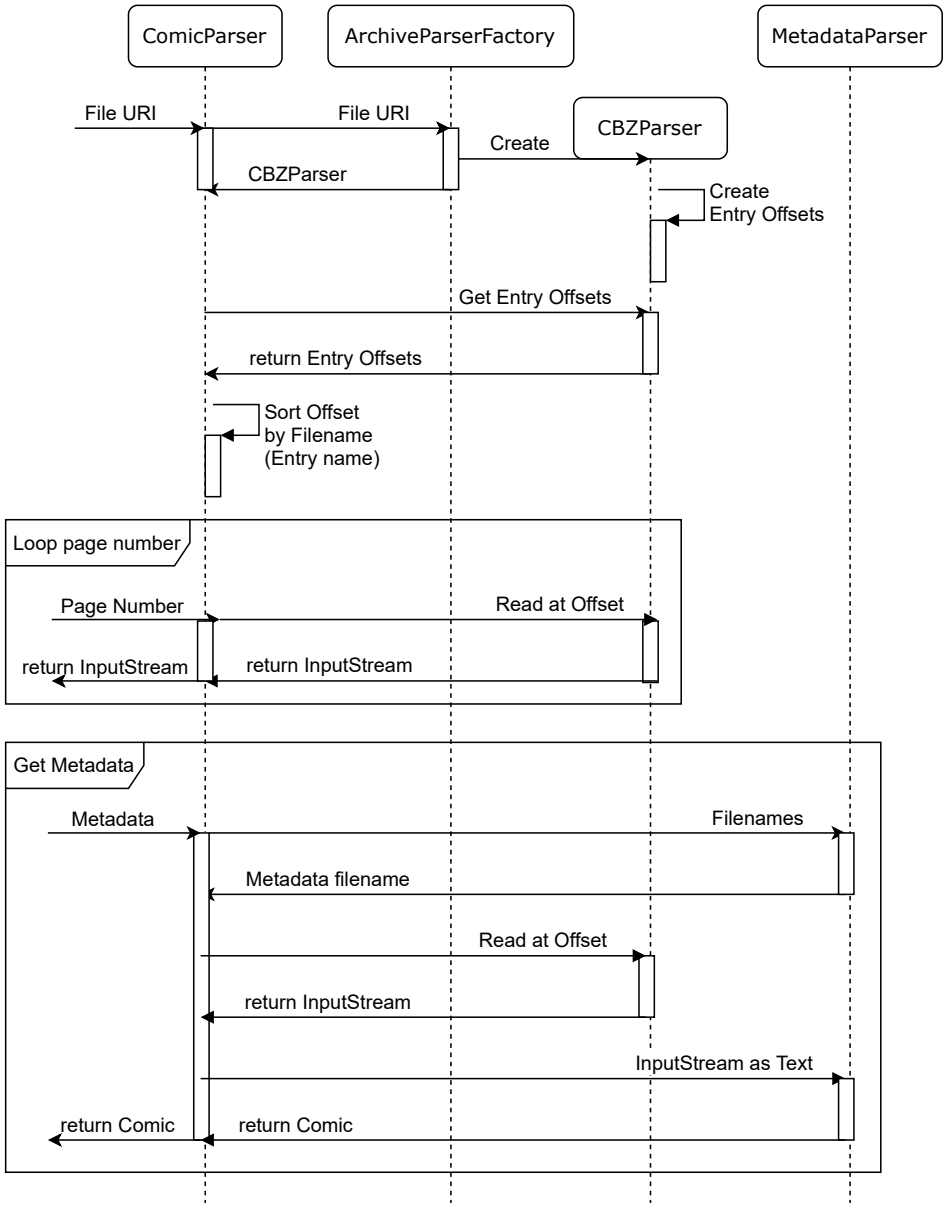
```
X-Men Vol 40 1.jpg
│     │      └─ Trang truyện số
│     └─ Số Volume, Number,...
└─ Tên tệp truyện
```

Vấn đề với định dạng này xuất hiện khi truyện có nhiều hơn 10 trang. Khi sắp xếp tệp ảnh theo ABC, các trang sẽ có thứ tự như sau:

```
X-Men Vol 40 1.jpg
X-Men Vol 40 10.jpg
X-Men Vol 40 11.jpg
...
X-Men Vol 40 19.jpg
X-Men Vol 40 2.jpg
...
```

Ta thấy ngay rằng thứ tự tệp ảnh trang truyện bị đảo lộn. Để giải quyết vấn đề này, cần viết hàm so sánh riêng cho tên tệp ảnh. Ý tưởng ở đây là gom những kí tự số liên tiếp với nhau thành một "kí tự" rồi mới so sánh. Đoạn mã giả sau đây trình bày thuật toán:

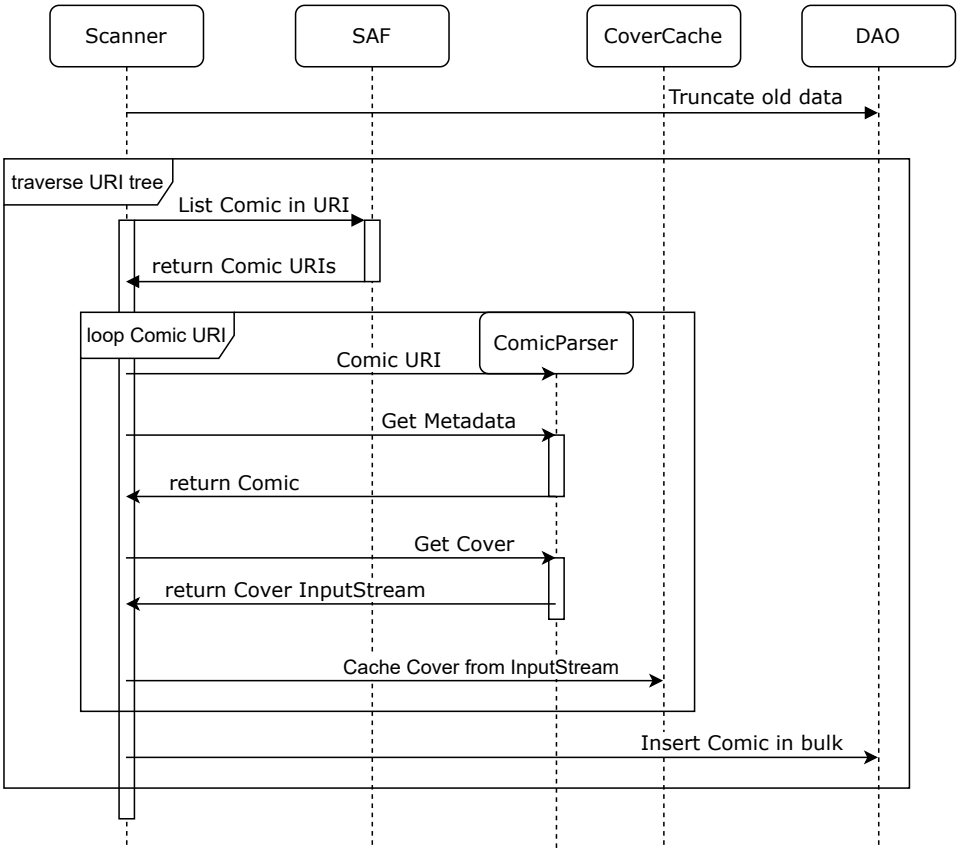
```
def compare(str1, str2):
    arrs = []

    for str in [str1, str2]:
        arrtmp = []
        acc = []

        for char in str:
            if is_number(char):
                acc.append(char)
            else:
                if len(acc) != 0:
                    acc = ''.join(acc)
                    acc = to_num(acc)
                    arrtmp.append(acc)
                    acc = 0
                arrtmp.append(to_codepoint(char))

    return compare_left_to_right(arrs[0], arrs[1])
```



Hình 4.14: `ComicParser` và các thành phần của nó

Scanner

Đây là lớp phục vụ cho tính năng quét truyện trong yacv. Biểu đồ luồng của ca sử dụng *quét mới* được thể hiện trong Hình 4.15.



Hình 4.15: Biểu đồ tuần tự của ca sử dụng quét mới tệp truyện

Một chi tiết mới trong biểu đồ này là lớp `ImageCache`, sẽ được giới thiệu ở mục về cache ảnh.

Ca sử dụng quét lại cũng có thiết kế gần tương tự, trong đó bỏ bước xóa dữ liệu, và thêm một bước quét cơ sở dữ liệu sau cùng để xóa truyện không còn trong bộ nhớ. Việc cập nhật và thêm truyện mới được thực hiện trong vòng lặp lớn bình thường.

Scanner nhận vào URI của thư mục gốc, rồi lặp qua từng tệp con, cháu,... Nếu đó là tệp truyện, nó gọi 'ComicParser' để lấy metadata, rồi lưu vào cơ sở dữ liệu qua DAO.

Quá trình quét tệp này giống như duyệt cây, do đó có hai cách cơ bản:

- Duyệt theo độ sâu (depth-first search, gọi tắt là DFS)
- Duyệt theo độ rộng (breadth-first search, gọi tắt là BFS)

Trong trường hợp cụ thể này, DFS được chọn. Lý do cho lựa chọn này là DFS có thể phát hiện *thư mục* nhanh hơn nhiều so với BFS. Mỗi khi gặp thư mục, DFS xử lí (thêm vào cơ sở dữ liệu) ngay, thay vì thêm vào hàng đợi. Do phát hiện được thư mục nhanh hơn BFS, Màn hình Thư viện, vốn hiển thị danh sách các *thư mục*, cũng hiển thị sớm hơn. Dù thời gian quét tổng thể không thay đổi, người dùng được thấy thư mục sớm hơn giúp tạo cảm giác ứng dụng khá nhanh.