UNIVERSITÀ DEGLI STUDI DI CAGLIARI

FACOLTÀ DI SCIENZE

Corso di Laurea Triennale in Informatica

Analisi ed estrazioni di informazioni da referti medici: Un caso di studio e implementazione nel progetto SIMIOR

Relatore
Dott. Gianmarco Cherchi

Studente
Lorenzo Ludovico Concas
Matr. N. 65315

ANNO ACCADEMICO 2021/2022

In the treatment path of hospital patients, one of the fundamental steps to define the state of health is the execution of tests aimed at better understanding the pathologies encountered. This information is expressed in medical reports, which together determine the progression of the patient in his hospital career. But keeping track of these trends can be complicated by human error in transcribing the information.

This thesis analyses the methodology implemented in the SIMIOR project to solve the problem and its future developments

Nel percorso di cura dei pazienti ospedalieri, uno dei passaggi fondamentali per definire lo stato di salute è l'esecuzione di test mirati a comprendere meglio le patologie riscontrate. Queste informazioni sono espresse in referti medici, il cui insieme determina la progressione del paziente nel suo percorso ospedaliero. Tenere traccia di questi andamenti può però essere complicato dall'errore umano nella trascrizione delle informazioni.

Questa tesi analizza la metodologia implementata nel progetto SIMIOR per risolvere il problema e i relativi futuri sviluppi

Indice

Capitolo 1

Introduzione

1.1 Cos'è il progetto SIMIOR

Il *Portable Document Format* è nato nel 1993 da Adobe Inc. con lo scopo di creare un formato file che permettesse di rappresentare documenti testuali o con immagini indipendentemente dall'hardware o dal software che usato per generare o visualizzare il documento stesso. Al momento la sua versione attuale è la 2.0 del 2017 ma lo standard più diffuso è la 1.4 di cui esiste la norma ISO 32000 stabilita nel 2008. Essendo un formato libero, possono essere creati programmi che leggono e scrivono documenti PDF senza la necessità di pagare diritti di utilizzo del brevetto, tuttavia l'accesso ai documenti contenenti le specifiche ISO sono a pagamento. Nonostante il successo di questa tecnologia, il PDF non è considerato come adeguato alla conservazione sul lungo termine perché la riproduzione può dipendere da sorgenti esterne quali Font o oggetti esterni (link, immagini allegate esternamente ecc).

1.2 La struttura del PDF

I PDF sono espressi tramite una sequenza di caratteri ASCII a 7 bit, pertanto le strutture di questo formato seguono tale codifica; fanno eccezione possibili oggetti binari che mantengono il loro flusso di bit originale senza alterazioni. Ogni file inizia con un intestazione leggibile indicante la versione del formato:

$$\%\text{PDF}-1.4$$

La parte costituente e fondamentale di un PDF sono dei blocchi costruttivi chiamati "COS" *"Carousel" Object Structure* (nome del progetto originale diventato in seguito Adobe Acrobat), questi blocchi nonostante il nome non sono dei veri e

propri "oggetti" come nei linguaggi *object-oriented*, ma rappresentano la struttura dei documenti. Alcuni oggetti sono delimitati da dei cosidetti `marker` che possono essere parentesi tonde, angolari, particolari caratteri o parole chiave. Esistono 9 tipi di oggetti *COS*, ognuno con un compito specifico:

- Oggetto *null*: è indicato come la semplice sequenza di caratteri `"null"`, è comunemente utilizzato per indicare la mancanza di un valore
- Oggetti booleani: sono la rappresentazione diretta dei due valori dell'algebra booleana, sono indicati come `"true"` e `"false"`, alcuni PDF scritti scorrettamente presentano variazioni nelle maiuscole (es `TRUE`, `True`, `FALSE` o `False`)
- Oggetti numerici: possono essere di due tipi, interi o reali, equivalenti ai rispettivi tipi matematici.
  I tipi interi consistono in uno o più cifre decimali, precedute opzionalmente dal segno e rappresentano un valore in base 10. Di seguito alcuni esempi
  - 1
  - -2134
  - 96

  I tipi reali presentano in più rispetto agli interi il punto come separazione fra parte intera e parte decimale, non è supportata la notazione esponenziale ne le radici non decimali. Alcuni esempi
  - 0.0099
  - .200
  - -3.1415
- Oggetti di tipo "Nome": Sono una sequenza di caratteri univoca in formato UTF-8 (escluso il carattere ASCII `null`) preceduta sempre da una barra obliqua (`/`). E' usato per definire un set di valori fissi.
- Oggetti stringa: delle semplici serie di byte scritti come caratteri, racchiusi fra parentesi tonde o come dati in formato esadecimale racchiusi fra parentesi angolari. Una stringa letterale consiste in un arbitrario numero di caratteri 8 bit racchiusi fra parentesi. Dato che in questo caso qualsiasi carattere può apparire nella stringa le parentesi non bilanciate e il backslash la procedura di escape di questi caratteri viene fatta tramite un ulteriore backslash. Oltre a questo può essere usata la notazione `\ddd`. Le stringhe letterali possono essere di vari tipi:

  - ASCII : Semplice sequenza di bytes contenente soltanto caratteri ASCII
  - PDFDocEncoded: Sequenza di byte codificata secondo lo standard ISO 32000-1:2008
  - Text: Sequenza codifica come PDFDocEncoding o come UTF-16BE (ossia con Byte Order Marker (BOM) in testa
  - Date: Una stringa ASCII che segue le direttiva del formato ISO 32000-1:2008 formattata come segue : `D:YYYYMMDDHHmmSSOHH'mm`

- Oggetti array: Sono collezioni eterogenee di altri oggetti COS racchiuse fra parentesi quadre e separati da un `white-space`.

- Oggetti dizionario: E' il tipo oggetto più comune nei PDF ed è la rappresentazione diretta delle collezioni `chiave-valore`, le chiavi sono sempre degli oggetti di tipo `nome`. I dizionari sono racchiusi fra doppie parentesi angolari («») e non vi è limite alla loro dimensione

- Oggetti stream (flusso): Sono sequenze arbitrarie di byte potenzialmente illimitati, compressi o codificati. Vengono utilizzati per immagazzinare grandi blocchi di dati in altri formati standardizzati (per esempio `font`, `immagini`, `json`, ecc...
  Gli stream sono sempre preceduti da un oggetto dizionario che ne descrive alcuni attributi fondamentali quali lunghezza del contenuto (obbligatoria) e tipo, questo dizionario è chiamato `stream dictionary`.
  Gli stream sono delimitati dai marker `stream` e `endstream`

Esistono vari caratteri di tipo `white-space` utilizzati per separare i costrutti sintattici (es nomi, numeri ecc) fra di loro. La seguente tabella mostra i vari tipi di caratteri

Decimale	Hex	Nome
0	00	NULL
9	09	Horizontal Tab
10	0A	Line Feed (LF)
12	0C	Form Feed (FF)
13	0D	Carriage Return (CR)
32	20	Space

I caratteri `Carriage Return` e `Line Feed` sono conosciuti come i caratteri `newline` e sono trattati come indicatori di fine linea. Il carattere % indica i "commenti" (utili per denotare informazioni quali la versione dello standard seguito).

1.2.1 Le Sezioni del PDF

I documenti PDF sono suddivisi in 4 sezioni : `Header`, `Body`, `Cross-Reference Table` e `Trailer`.

L'`Header` inizia alla posizione 0 del file e consiste in almeno 8 byte seguiti da un marker di fine linea. Come accennato prima servono a contenere l'intestazione che identifica il documento come PDF e la versione del formato. Nel caso il documento contenga anche dati binari, seguirà una seconda linea contenente un carattere indicante i commenti (%) e 4 caratteri ASCII dal valore maggiore di 127, i più diffusi sono: âãÏÓ. Quindi la presenza alla seconda linea di %âãÏÓ indica che nel documento è presente un file binario.

Il `Body` (corpo) è come suggerisce il nome il centro dove vengono inseriti tutti gli oggetti COS che compongono il documento che poi verrà renderizzato e reso visibile, inizia subito dopo l'`header` senza marker specifici.

La `Cross-Reference Table` è un attributo semplice ma fondamentale per il PDF, infatti questa tabella fornisce gli offset (scostamenti) binari rispetto all'inizio del file di ogni (e per ogni) oggetto indiretto, permettendo cosi all'analizzatore del documento di cercare e leggere più velocemente gli oggetti in qualsiasi momento garantendo così la possibilità di un accesso casuale piuttosto che una lettura sequenziale, velocizzando anche l'apertura e l'elaborazione. E' delineata dai marker `xref`

Infine il `Trailer` invece occupa gli ultimi byte del documento e consiste nel marker `trailer` seguito da un oggetto dizionario contenente la chiave `size` indicante la dimensione in byte del documento e la chiave `root` che indica il riferimento all'oggetto-catalogo, un particolare oggetto che contiene vari puntatori a vari tipi di oggetti speciali. Altri elementi che il trailer può contenere sono la chiave `Encrypt` usata per specificare il dizionario di crittografia del documento, o `ID` usata per dare un identificatore al file.

Capitolo 2

Tecniche di analisi del PDF

Esistono vari approcci per l'analisi e lo *scraping delle informazioni* da un PDF. In questa tesi ne analizzeremo 3.

2.1 L'analisi grezza

Si analizza il documento PDF senza ulteriori elaborazioni basandosi soltanto sulle conoscenza dell

2.2 Metodologie

2.3 La libreria Apache PDFBox

Capitolo 3

I referti medici

3.1 La struttura dei referti

3.2 Il problema dell'analisi

3.3 la soluzione

Capitolo 4

Il SIMIOR

4.1 Cenni sul soggetto di studio

4.2 L'implementazione pratica

4.3 Prestazioni?

Capitolo 5

Conclusioni

5.1 Non ne ho idea