Uma solução de baixo custo para digitalização baseada em Clara OCR e DJVU

Ricardo Ueda Karpischek - EdIC
Imre Simon - IME/USP

(*parcialmente suportados pelos processos CNPq 380374/03-0 e CNPq 465901/00-0*)

Através do programa Clara OCR, e do uso do formato DJVU e do pacote djvulibre, obtivemos uma solução prática para a digitalização de livros, capaz de obter uma primeira versão eletrônica na digitalização de livros, apta para ampliar o volume de materiais de referência atualmente disponíveis eletrônicamente, inclusive para o trato computacional da língua. Os procedimentos descritos neste artigo utilizaram dois livros diferentes [7], [8].

Introdução

A solução mais barata para a digitalização de livros consiste em limitar-se a escanear as suas páginas, gerando uma coleção de arquivos, cada um contendo a imagem digital de uma página. É assim por exemplo que a Biblioteca Nacional da França está disponibilizando algumas dezenas de dicionários publicados nos séculos XV-XIX (http://gallica.bnf.fr). Também é dessa forma que diversos pesquisadores têm tornado eletrônicamente acessíveis a sua produção, como por exemplo a iniciativa de Gratzer [4].

Uma segunda alternativa consiste em produzir um texto eletrônico através da digitação ou do reconhecimento óptico dos caracteres (OCR). Essa alternativa tem as vantagens de permitir buscas de palavras, e também o reaproveitamento do conteúdo por programas que lidam com linguagem. O Projeto Gutenberg [5] é provavelmente o exemplo mais conhecido de uma iniciativa baseada nessa opção. Ela tem como obstáculo o custo elevado da digitação e/ou da revisão. Isso motivou recentemente a criação de um sistema de revisão cooperativa através da Internet bastante engenhoso e eficaz [6]. O Projeto Runeberg [9] tem um sistema de revisão cooperativa mais simples.

Procuramos uma solução de meio termo capaz de ser refinada no tempo, e que oferece tanto imagens digitalizadas para consulta manual, quanto texto reconhecido não revisado para trato computacional, fundidos no formato djvu, cujas boas propriedades tornam essa solução prática e atraente. Nossos testes fizeram uso de dois livros: Aspectos Teóricos da Computação [7] e a quarta edição do Dicionário de Cândido de Figueiredo [8], que serão referidos respectivamente por ATC e CF.

O Formato DJVU

O djvu (Déjà Vu, do francês) é uma ampla plataforma para disponibilizar documentos num browser pela teia. Ela é cuidadosamente otimizada para aspectos de versatilidade dos documentos representáveis, aspectos da qualidade gráfica das suas imagens, aspectos da eficiência da compressão e recuperação dos documentos representados. É um formato aberto e existem componentes de software livre para praticamente todas as etapas da preparação e recuperação de documentos em djvu. No todo, o djvu é uma alternativa vantajosa à plataforma pdf da Adobe. A plataforma djvu incorpora também um projeto de investigação e documentação de grande interesse científico [6]. Apesar destas qualidades notáveis, o formato e a plataforma são relativamente pouco conhecidos ainda nos meios da Internet.

Uma característica importante da plataforma djvu é a capacidade de acoplar à imagem digitalizada de um documento uma camada UTF-8 de texto digitalizado. Isto permite a realização de buscas textuais enquanto o usuário examina a imagem do documento.

Existem vários locais na teia para demonstrar as qualidades desta tecnologia, mencionamos em particular os sítios [2]. Há também um dicionário de inglês disponível na Internet no formato djvu com a camada texto presente [11].

A plataforma djvu começou a ser desenvolvida nos Laboratórios da AT&T Bell em 1996 por uma pequena equipe de cientistas liderada por Yann LeCun [1]. Em 2000 a empresa LizardTech [3] adquiriu a tecnologia e posteriormente abriu partes substanciais da mesma sob a licença GPL. A partir do código liberado alguns dos desenvolvedores originais da plataforma lançaram o projeto de software livre DjVuLibre http://djvu.sf.net onde desenvolveram uma plataforma praticamente completa.

O programa Clara OCR

A dificuldade de aplicar OCRs do mercado a livros antigos e/ou com ortografia antiquada levou-nos a desenvolver uma ferramenta simples mas especializável, isto é, que pudesse ser adaptada às características peculiares de cada livro. Com isso em mente foi criado o Clara OCR, que implementa heurísticas de reconhecimento simples, mas com grande capacidade de ajustes.

Os atuais resultados obtidos pelo Clara OCR podem ser considerados bons em alguns casos. Para cada 1000 caracteres do CF, 995 estão sendo reconhecidos. Essa taxa de 99.5% é particularmente significativa pelo fato de ser puramente óptica, isto é, não depender de heurísticas baseadas em ortografia (o Clara OCR não inclui atualmente heurísticas baseadas em ortografia). Recentemente, o Clara OCR ganhou suporte ao formato djvu.

O Dicionário de Cândido de Figueiredo

As características gráficas da quarta edição do dicionário de Cândido de Figueiredo, relevantes para fins de OCR, podem ser sumarizadas assim: apenas um tamanho de fonte (exceto nas introdução), presença de negrito e itálico, apenas uma língua (Português), ortografia antiquada, e, na maior parte das páginas, a impressão é nítida, o papel apresenta pouca ferrugem, e é não transparente.

O livro Aspectos Teóricos da Computação

De um livro publicado há relativamente pouco tempo como o ATC, espera-se em geral bons resultados de reconhecimento óptico. Fizemos uma primeira tentativa utilizando o Clara OCR que obteve resultados medianos. Jim Rile (criador do portal djvuzone.org) realizou uma segunda tentativa que obteve resultados muito bons para o texto, conforme comentaremos na seção seguinte.

Procedimento adotado

O escaneamento dos dois livros utilizou o pacote SANE e um scanner popular (Genius HR5), resolução 600 dpi e 256 tons de cinza. A qualidade do reconhecimento depende do treinamento e também da maturação do Clara OCR. Por isso a estratégia adotada foi realizar um treinamento básico, e escrever um script que a partir das páginas escaneadas e dos dados de treinamento fizesse de forma automática o reconhecimento de todas as páginas, e montasse o arquivo djvu final.

Dessa forma, na medida em que o Clara OCR evoluir, poderemos de forma automática produzir novas versões do CF e do ATC, consumindo para isso apenas tempo de CPU. Esse fato tem um alcance prático significativo, pois o resultado poderá melhorar com o tempo, mesmo sem a aplicação direta de esforço humano.

A medida da qualidade do resultado está sendo feita de forma ingênua como segue: escolhemos ao acaso um trecho do livro, contamos quantas palavras ele tem (N), e contamos quantas delas foram corretamente segmentadas e reconhecidas (R). O quociente R/N é a nossa medida de qualidade.

Convém notar que as medidas típicas de qualidade de reconhecimento envolvem caracteres, e não palavras. A medida que adotamos é pessimista, porque o erro no reconhecimento de uma palavra reduz-se geralmente a um erro no reconhecimento de apenas uma das suas letras. Por outro lado, a medida adotada tenta refletir a probabilidade de sucesso na localização de uma ocorrência de uma palavra através de um visualizador djvu, o que parece ser justo do ponto de vista das necessidades típicas do usuário.

Algumas medições

Os números mais relevantes obtidos nos nossos testes estão sumarizados na tabela que segue:

	CF	**ATC**
imagem pgm (megabytes)	46369	5386
imagem djvu (megabytes)	58	4
texto djvu (megabytes)	13	1
páginas	2156	304
qualidade (Clara)	0.82	0.70
qualidade (FineReader)	-	0.99
treinamento (horas)	16	2

Observações:

1. Atualmente não há OCRs disponíveis (livres ou comerciais) com suporte a DJVU além do Clara OCR. Jim Rile possui uma ferramenta baseada no ABBYY FineReader OCR, capaz de produzir a camada texto djvu, mas não se trata de um produto do mercado. Ele tomou conhecimento dos nossos testes com o ATC, e aplicou a sua ferramenta, partindo das imagens comprimidas com djvu, obtendo os resultados que estão sumarizados na linha "erros (FineReader)".

2. A linha "imagem pgm" indica o tamanho total das imagens não comprimidas das páginas escaneadas.

3. A saída HTML, não editada, produzida pelo Clara OCR para as primeiras 100 páginas do CF pode ser examinada em http://www.claraocr.org/cgi-bin/dict.cgi/R/cf/page e um arquivo djvu parcial que contém apenas as 10 primeiras páginas pode ser obtido em http://www.claraocr.org/cf-10.djvu

4. O desempenho do Clara OCR depende bastante dos ajustes do programa a cada livro, o que inclui mas pode não se limitar ao treinamento.

Comentários adicionais

Os resultados do Clara OCR (e de qualquer OCR) variam significativamente com os documentos utilizados, principalmente quando se admite trabalhar com documentos publicados há várias dezenas ou centenas de anos, multilíngues (caso típico de dicionários) e incluindo ou o uso de línguas mortas, ou de ortografias antiquadas.

Atualmente os resultados do Clara OCR dependem apenas dos fontes utilizados na impressão e do estado do papel. A razão disso é a ausência de heurísticas baseadas em ortografia ou consulta a dicionários. Assim, os resultados obtidos para o Cândido de Figueiredo irão repetir-se para outros livros com as mesmas características gráficas desse dicionário. Outros livros poderão apresentar resultados piores ou melhores. De modo geral, pode-se dizer que dicionários do século XIX apresentarão resultados piores.

Os autores pretendem prosseguir os trabalhos com o CF e disponibilizar livremente uma primeira versão eletrônica com qualidade de reconhecimento possivelmente melhor do que a atual, lançar mão de um processo de revisão cooperativa a fim de obter um resultado definitivo. Um dos atuais obstáculos para essa revisão é a atual inexistência de uma ferramenta de revisão simples e capaz de lidar diretamente com o formato djvu.

É possível também que venhamos a realizar em breve testes de OCR com o Dicionário de Pedro José da Fonseca [10], para o qual temos já uma versão eletrônica completa (imagem apenas).

Referências

[1] DjVuLibre History and Credits, http://djvu.sourceforge.net/credits.html

[2] DjVuZone (http://www.djvuzone.org) Planet DjVu (http://www.planetdjvu.com) e Yann’s DjVu Page (http://yann.lecun.com/ex/djvu/)

[3] LizardTech, Inc. Imaging Solutions http://www.lizardtech.com

[4] G. Gratzer, Publishing legacy document on the Web, TUGboat. ftp://server.maths.umanitoba.ca/pub/gratzer/articles/A14.pdf

[5] Projeto Gutenberg, http://promo.net/pg/

[6] Distributed Proofreaders, http://texts01.archive.org/dp/

[7] Cláudio L. Lucchesi, Imre Simon, Istvan Simon, Janos Simon e Tomasz Kowaltowski, Aspectos Teóricos da Computação, IMPA-CNPq, 1979, http://www.ime.usp.br/~is/atc/

[8] Cândido de Figueiredo, Novo Diccionário da Língua Portuguesa, Lisboa, 1926.

[9] Projeto Runeberg, http://runeberg.org/

[10] Pedro José da Fonseca, Diccionario Portuguez-Latino, nona edição, Lisboa, 1879, http:/www.linguateca.pt/pjf

[11] The Century Dictionary, http://www.global-language.com/CENTURY/