

可搜尋中文字詞的 PDF 檔案

翁鴻翎*

2002.9.1

可以搜尋中文字詞的 PDF 檔案, 有好幾種途徑可以達到。例如, Adobe 公司的 ACROBAT distiller 5.0 就可以將 PS 檔案轉換成內嵌中文 TrueType 字形之 PDF 檔案, 而且檔案可以搜尋中文字詞。對於 cwTeX 使用者而言, 翁鴻翎先生把原先 cwTeX 的 PostScript 字形, 轉為標準的 UNICODE TrueType 字形。透過此一途徑, 我們可以產生可搜尋中文字詞之 PDF。本文說明如何以 cwTeX 製作可搜尋中文字之 PDF 檔案。

1 可搜尋中文字詞之 PDF 檔案

以 TeX 排版程式要產生可以中文搜尋的 PDF 檔案, 目前必須透過 DVIPDFM-CJK 工具程式。從技術層面而言, pdfTeX 應該在不久也可以達到同樣的功能。DVIPDFM-CJK 是以 Mark A. Wicks 所寫的 Dvipdfm 程式為基礎, 它可以將 TeX 編譯後所產生的 dvi 檔案轉換出 PDF 檔案。後來, Shunsaku Hirata 與 Jin-Hwan Cho 兩人, 在 Dvipdfm 中加入了 CJK cid 字形的支援, 這個計畫稱為 DVIPDFM-CJK。

如果你是使用 cwTeX 系統, 只要再加裝 cwTeX TrueType 字形 (以下簡稱 TTF 字形), 就可以直接使用 DVIPDFM-CJK 產生可以中文搜尋的 PDF 檔案。使用 cwTeX TTF 字形的一個好處為, 可以跟 PostScript 字形使用同一份的 tfm 與 fd 檔案, 因為 cwTeX 的 PostScript 與 TTF 字形的字形高度、寬度和深度是一樣, 避免了很多設定上的麻煩。底下將說明安裝及排版方法。

1.1 安裝方法

欲在 PDF 檔案中搜尋中文字詞, 必須使用 ACROBAT 5.0 版以上之版本。如果你目前的

* cwTeX 發展小組。

版本是舊版,請改裝新版。

請自 cwTeX 網站下載 `cwtex-ttf.zip`, 其中包含 5 套 cwTeX 中文 TTF 字型檔, 及製作可搜尋中文字詞 PDF 檔案之工具程式及檔案。若電腦中安裝的是 MiKTeX 2.1 版, 直接將 `cwtex-ttf.zip` 解壓至 `c:\texmf` 即可。以上之安裝動作會將 `c:\texmf\miktex\bin` 檔案夾內之 `dvipdfm.exe` 以新版取代。安裝之後, 請執行:

```
c:>initexmf -u
```

更新 TeX 之檔案系統。

如果你是使用 fpTeX 系統, 須自行將各檔案解壓至 TeX 系統內。其中, DVIPDFM-CJK 程式檔置於 `cwtex-ttf.zip` 之 `\texmf\cwtex\util` 檔案夾內。

1.2 排版測試

以測試檔 `examp1.ctx` 為例, 排版方法如下。先以 cwTeX 將檔案轉換成 `examp1.tex`, 再以 \TeX 排版成 `examp1.dvi`。之後, 進入 DOS 視窗, 執行:

```
c:\xtemp>dvipdfm examp1
```

以 ACROBAT 5.0 版開啓 `examp1.pdf`, 即可搜尋檔案內中文字詞。

2 使用限制與問題

由於要由 DVIPDFM-CJK 產生 PDF 檔案, 多少會對於文件的排版產生一定的影響。像功能強大的 `pstrick.sty` 和 `psfrag.sty` 就無法直接使用。不過由於 PDF 的檔案的目的就是定位在網路上可攜式文件, 你可以嵌入 `.jpg` 與 `.png` 等圖檔格式。其他可能的問題, 如同時引入 `hyperref.sty` 與 `url.sty` 巨集套件, 會出現錯誤。

根據實際的測試, `hyperref.sty` 並無法完整的支援 CJK 的語系。譬如, 當 PDF 檔案很大時, `bookmark` 還是會出現亂碼。測試的方法很簡單, 只要寫一個大概 30 個左右的 `\section{中文字}`, 就會出現亂碼了。主要的原因是 CJK 語系的字數通常都有幾千個, `hyperref.sty` 轉換的過程很有可能出現問題; 再加上 `hyperref.sty` 是讀取檔案再處理, 這一些 `moving arguments` 在 TeX 系統下是很容易流失的。

我們希望在不久的將來可以完整的支援。目前唯一的辦法只能在 ACROBAT 中修改。不過使用 DVIPDFM-CJK 也有好處的, 因為產生 PDF 的速度很快, 我們用來轉換 cwTeX 的使用手冊, 兩三分鐘就完成, 而且程式不會當掉, 如果用 DVIPS 情形就沒有這樣好了。

3 cwTeX 使用商業字形

當然, cwTeX 也可以直接使用其他廠商的商業字形, 使用的方法簡述如下。首先, 用工具程式 `ttf2tfm` 產生想要的 `tfm` 檔案, 例如, 假如你有文鼎的 UNICODE 細明體字形, 你想要用來替代 cwTeX 的明體, 假設文鼎細明體的檔名是 `wdm.ttf`, 你可以先建立一個臨時目錄, 把 `wdm.ttf` 拷貝到此臨時目錄下, 然後下指令:

```
c:\xtemp>ttf2tfm wdm.ttf -e 0.90 m@ucwtex1@
```

過一會兒, 就會有 52 個明體 (m) 的 `tfm` 檔案產生, 選項 `-e` 的作用是把字形向中間壓扁, 一般字形廠商的字型相較於 cwTeX 字形, 會有寬度過寬的現象, 加此選項可以得到較美觀的排版效果。

接著, 我們須產生新的字形驅動程式, 你可以使用 cwTeX 內附的程式 `genfd.exe`, 指令如下:

```
c:\xtemp>genfd -d 0.90 m
```

以上的指令中, 選項 `-d` 是指要等比例縮小字形為原來的 90%, 這是因為 cwTeX 的字型較一般的商業字形小, 我們做如此的設定可以獲得較好的輸出效果。

當產生了新的 `tfm` 和 `fd` 檔案後, 最後我們要更新一下 DVIPDFM-CJK 的字型對應檔, 請在在 `c:\texmf\dvipdfm` 的子目錄下找到 `cid-x.map` 檔案, 在其中加入下列一行指令:

```
m@ucwtex1@ UniCNS-UCS2-H :0:wdm.ttf -e 0.90
```

其中, `UniCNS-UCS2-H` 是指 `\dvipdfm` 子目錄下的 字型對應檔 (font mapping)。繁體中文的 UNICODE 字形對應檔案是 `UniCNS-UCS2-H`, 如果你想嵌入非 UNICODE 的字形, 對應的檔案是 `ETen-B5-H`。:0: 這個數字表示, 你的字形的附檔名是 `.ttf`, 如果你的字型的附檔名是 `.ttc`, 你應該改為 `:1:`。請參見 `cid-x.map` 檔案內之簡單說明。由於技術上的的細節已經超出本文的範圍, 在此不贅述, 有興趣者請見 Adobe 公司的技術文件。

如果你的 TrueType 字形不是 UNICODE, 你就要把呼叫 `ttf2tfm` 指令中的 `@ucwtex1@`, 改為 `@final@`, `cid-x.map` 檔中的輸入也要做相同的更改。由於一些廠商的字型並不允許內嵌其字形, 這時候我們可以叫 DVIPDFM-CJK 不要內嵌, 指令是在字形名稱前加一

個 !, 例如, Window 系統的新細明體是不能內嵌的, 這時候你可以先照以上的步驟, 產生 tfm 與 fd 後, 然後在 cid-x.map 檔案中加入:

```
m@ucwtex1@ UniCNS-UCS2-H :1:!mingliu.ttc -e 0.90
```

這樣你的 PDF 文件就不會內嵌 Window 系統的新細明體, 這時候, 如果你把你產生的 PDF 檔案拿到其他的電腦上, 如果該電腦有裝新細明體, 就可以由 Acrobat Reader 開啓瀏覽, 如果該台電腦沒有裝相對應的字形, 這一台電腦就無法瀏覽這個 PDF 檔案。

4 字型授權問題

若使用商業字型創造 PDF 檔案, 要特別注意的是版權問題。基本上, 使用者並不能隨意內嵌中文字形在 PDF 檔案內, 因為內嵌在 PDF 檔案內的字型資料, 是直接提取字形檔案中的資料; 這些資料是有版權限制的。一般來說是屬於該創作公司的, 如果你要把 PDF 檔案放在網路上散播的話, 就是直接侵權的行為, 跟隨意散播某軟體公司的軟體, 是同樣的行徑。

目前為止, 商業的軟體字形, 只有文鼎公司允許內嵌該公司的字型, 至於華康字型, 由於華康公司在推廣其 DYNADOC, 這個產品在跟 PDF 爭食中文網路可攜式文件的市場, 短期內並不可能允許使用者內嵌其字形, 據我們所知, 在 Mac 作業系統下, 華康將開放使用者可以在 PDF 內嵌其字形, 不過允許內嵌的是 PostScript 的 CID 字形, 這種字形的只能由 Mac 下的 Adobe Type Management 讀取嵌入, 由於是內嵌 PostScript 字形的資料, 品質沒有話講, 但是其他的作業平台可能就沒有辦法享受到, 遑論是 T_EX 的使用者。再一次的提醒讀者, 請各位使用者不要任意內嵌字形, 免得觸法而不自知。關於字形的授權請詳讀相關文件。

一般而言, 每一家字形廠商, 都會在字形資料裡加註是否允許使用者進行某種行為, 這一些行為包括了:

- Only printing and previewing of the document is allowed.
- Everything is allowed.
- Embedded of this font is not allowed.
- Editing this document is allowed.

目前而言,如果你在使用 DVIPDFM-CJK 時看到了“printing and previewing only”的字樣時,就是不能內嵌的。所以你最好詳讀字形的使用授權說明,然後才能內嵌這一些字形,目前 `cwTeX` 的 UNICOD 字形也是在字形資料中採取這種加註的方式,如果你要使用這一些字形,請詳讀 `cwTeX` 相關授權的文件,就實際測試列印的結果,`cwTeX` 的 UNICOD 字型的品質,相較於知名字形廠商華康和文鼎的 UNICOD 字形,一點也不遜色。不過,如果你要的是高品質的輸出,解析度要大於 700dpi 的話,建議你還是用 `cwTeX` 的 PostScript 字形,用 PostScript 字形是高品質印前輸出的唯一選擇。

5 聲明

本文件所提的各公司和商品,皆為該註冊公司所有。本文論及的各公司及所屬產品,純為評析和說明的需要,並無侵權之意,特此聲明。本文件為 `cwTeX` 排版系統的一部份。