

用 CJK 宏文件包来编译中文 TeX 文本文件^{*}

姜 正 禄

(中山大学数学与计算科学学院, 广东 广州 510275)

摘 要: CJK 是一个专门用于处理中日韩泰等亚洲国家文字的 LaTeX 宏文件包。LaTeX 是一套出色的文本处理软件。揭示了 LaTeX 编译中文的原理, 介绍了如何用该宏文件包的格式来书写中文 TeX 文本文件, 便于 latex (或 pdflatex) 把 tex 文档直接编译成 dvi (或 pdf) 文件。与此同时, 还叙述了中文全真字库的安装。

关键词: LaTeX 排版系统; CJK

中图分类号: G642 0 **文献标识码:** A **文章编号:** 0529-6579 (2004) S1-0119-04

LaTeX 是一套出色的文本处理软件 (见文献 [1, 2]), 具有很强的处理科学文件尤其是数学文件的功能, 因而深受广大用户特别是学术界的朋友所喜爱。它是免费的软件, 安装在 Unix 和 Windows 等系统上的都有, 可以从许多大学的计算机上直接下载。CJK 是一个专门用于处理中日韩泰等亚洲国家文字的 LaTeX 宏文件包。CCT, 也称中文 LaTeX, 是一套处理中文的 LaTeX 排版系统 (见文 [3]), 其实它仅仅是一个扩展了 LaTeX 中 CJK 宏包的处理中文文本文件的软件。不过, 它需要该系统中的 cct, 先把中文的文本文件编译成 ASCII 码的 TeX 文件, 再用 latex 把该 ASCII 码的文件编译成 dvi 文件。而用 CJK 宏包的格式来书写的中文 TeX 文本文件可以被 latex 直接编译成 dvi 文件或被 pdflatex 直接编译成 pdf 文件; 编译时可使用第 1 类型或全真的字体, 生成的 pdf 文件更清楚、更漂亮; 在一个中文 TeX 文本文件中可以使用多种编码的文字, 如中文简体和繁体、日文、韩文等等。另外, CJK 的使用还比较简单, CJK 宏包里的 “CJK. doc” 等文档都有很好的说明。尽管如此, 作者认为通过揭示 LaTeX 编译中文的原理向广大用户介绍如何用该宏文件包的格式来书写中文 TeX 文本文件是很有必要的。很多用户往往会被 LaTeX 中的各种文件和字体搞得头昏脑胀。其实, LaTeX 用 CJK 宏包处理汉字的过程很简单。

假定带有 Redhat Linux 系统的计算机上已经安装好了 LaTeX, 还有 CJK 宏包和第 1 类型或全真的中文字库。首先用 gedit 编辑一个文件 “test. tex”, 里面有如下内容:

```
\documentclass{ article}
\usepackage{ CJK}
\begin{ document}
\begin{ CJK } { GBK } { song}
幸福伴您一生
\end{ CJK}
\end{ document}
```

值得提醒的是, 保存 TeX 文本文件时一定要用中文编码的格式。用户再在终端控制台上敲入命令行

```
latex test tex
```

latex 的程序就启动了, 如图 1 所示。图 1 提供了 latex 编译过程的信息, 它使用了 TeX 的 Web2C7. 3. 1 版本, 还调用了相关的 CJK 宏包中的文件, 并把中文 TeX 文本文件 “test. tex” 编译成 dvi 文件 “test. dvi”。Redhat Linux 系统中 LaTeX 和 TeX 其实是同一个程序。在有些系统下 latex 是一个脚本, 里面只有几行字。LaTeX 只是 TeX 的一种特殊 “格式”。LaTeX 中的条条框框比 TeX 中的多得多。

LaTeX 编译到 `\begin{ CJK } { GBK } { song}` 的时候开始发现用户使用了 CJK 宏包, 能把从现在开始一直到 `\end{ CJK }` 的中文变成了一些 TeX 的命令。这些命令根据不同的汉字转换成不同的字体。在终端控制台上敲入命令行

```
dvips -o test. ps test
```

就启动了 dvips 的程序, 如图 2 所示。它把 dvi 文件打印成打印机可以识别的 ps 文件。如图 2 中有 dvips 打印时用到的字体文件。LaTeX 发现 “幸福伴

* 收稿日期: 2004—02—13

基金项目: 国家自然科学基金资助项目; 留学归国人员基金资助项目

作者简介: 姜正禄 (1966 年生) , 男, 副教授。

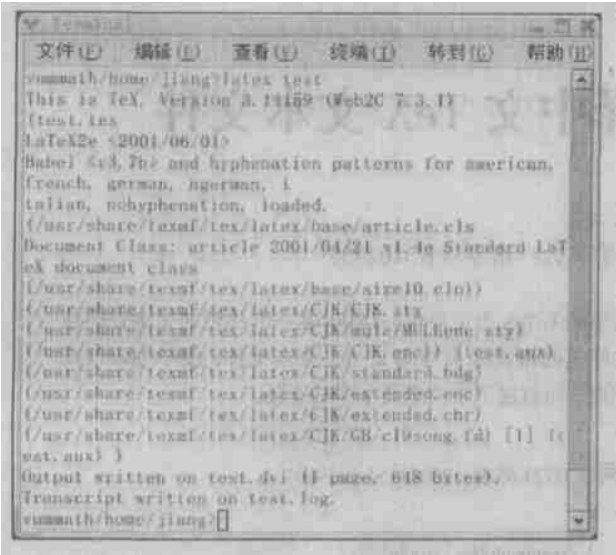


图 1 latex 编译过程

Fig 1 Procedure process of latex compilation



图 2 dvips 编译过程

Fig 2 Process of dvips compilation

您一生”，就会把它们要对应的字体文件名依次设置为 gbksong60、gbksong42、gbksong36、gbksong51、gbksong61 和 gbksong55，这就有利于 dvips 从那些与上述文件名相对应的 pfb 字体文件“gbksong60. pfb”、“gbksong42. pfb”、“gbksong36. pfb”、“gbksong51. pfb”、“gbksong61. pfb”和“gbksong55. pfb”里面分别能取到各自相应的字符。不过，图 2 表明 dvips 是从后向前查找字体文件的。

LaTeX 编译时用到的文件“l9song. fd”是描述字体映射方式的，图 1 表明它在目录/usr/share/texmf/tex/latex/CJK/GB/中。在终端控制台上敲入命令行

```
cat/ usr/ share/ texmf/ tex/ latex/ CJK/ GB/ c19song. fd
```

就会发现它里面有如下一行

```
/DeclareFontShape{ C19}{ song}{ m}{ n}{ <-> CJK * gbksong}
{ }
```

C19 和 song 分别表示 GBK 编码和宋体，m 和 n 分别表示字体的粗细中等（medium）和形状正体（upright）。LaTeX 用到这个字体描述文件时就会把汉字变成一系列有规则的能映射到字体文件名的代码。比如，LaTeX 把汉字“幸”变成了代码：“C19/song/m/n/10/60/3”。该代码前面是“C19/song/m/n/10/60”，它表示用 GBK 编码（C19）的宋体，粗细为中等（m），形状为正体（n），大小是 10pt，字集号码为 60，这些字符会被映射为 gbksong60；后面接着一个 ASCII 码字符是 3，它是汉字“幸”在这个字体文件“gbksong60. pfb”中的索引号，它的十进制数值正好是 51，由于索引号所对应的十进制数值为 0 的字是在其相应的字体文件的第 1 个位子上，所以“幸”就在该字体文件的第 52 个位子上。“福”字被变成的代码为“C19/song/m/n/10/42/4”，它在字体文件“gbksong42. pfb”的第 53 个位子上。从上面得到的文件“test. ps”中还可以找到其它 4 个字被变成的代码、所在的文件名、索引号及其十进制数值，如表 1 所示。

dvips 打印 dvi 文件“test. dvi”时会发现该 dvi 文件里面有上述那六个字体文件名，而且还有对应的索引号。它就马上去读取配置文件“config. ps”里面的字体映射文件列表中的文件“cjk. map”，该字体映射文件会引导 dvips 把这些字体的图像（glyph）从上述那六个字体文件里下载来，并放在文件“test. ps”中。

像处理英语 tex 文件一样，latex 通过使用 CJK 宏包的转化可以把中文 tex 文件编译成了 dvi 文件。同样，pdflatex 还能把中文 tex 文件编译成 pdf 文件。

下面我们介绍 CJK 宏包中两个新环境和一些新命令。两个新环境如下：

```
\begin{CJK}{编码名}{字族名}
...
\end{CJK}
\begin{CJK*}{编码名}{字族名}
...
\end{CJK*}
```

表 1 字体代码

Tah 1 Firt codes

字	被变成的代码	所在的文件名	索引号	十进制数值
伴	C19/song/m/n/10/36 \ 212	gbksong36. pfb	\ 212	138
您	C19/song/m/n/10/51 s	gbksong51. pfb	s	115
一	C19/song/m/n/10/61 \ 230	gbksong61. pfb	\ 230	152
生	C19/song/m/n/10/55 \)	gbksong55. pfb	\)	41

编码名是指编码方式的名称，共有GBK，GB，UTF8，JIS，GB5，GB5+，GBt 等等。字族名是指在指定编码下的字体类型的名称。我们可以不填字族名，如字族名没有填，那它还有个默认值，即为 song。我们也可以自己定义字族名，不过，对中文来说，通常定义字族名为 song 和 kai 等等，song 表示宋体，kai 表示楷体，便于编辑中文 tex 文件。

这两种环境的不同在于：在后一种环境下编译中文文档时 latex 会自动删除汉字后面不受保留的空格和新行，而在前一种环境下编译却没有这种功能。我们建议初学者采用后一种环境来编辑中文 tex 文件，并用命令 \CJKtilde 来改变当空格使用的符号~的长度。

\CJKenc{ 编码名 } 是一个改变编码方式的命令。比如说，前面的编码名是 GB，指的是汉字简体，要把它改为繁体，其对应的编码名是 GB5，就用如下命令 \CJKenc{BG5} 即可。

\CJKfamily{ 字族名 } 和 \CJKencfamily{ 编码名 } { 字族名 } 是两个改变字体类型的命令。前者在任何环境下都能改变字体类型，而后者只在环境中的字族名没有填的情况下才起到改变字体类型的作用。如前面的字族名为 song，要把它改为 kai，就要用如下命令：

\CJKfamily{ kai } 或 \CJKfamily{ } \CJKencfamily{ 编码名 } { kai }

如前面的字族名没有填，要把它改为 kai，就用命令：

\CJKfamily{ kai } 或 \CJKencfamily{ 编码名 } { kai }。

我们可以用 \fontshape{ n } 和 \fontshape{ sl } 分别表示选择正体和斜体，用 \fontseries{ m } 和 \fontseries{ bx } 分别选择标准体和粗体。

\CJKchar[GBK] { 197 } { 198 } 是一个用汉字本身带有的惟一的编码的值来获取汉字的命令。其中 GBK 是编码名，如要繁体字，就要用 GB5 作为它的编码名；197 是“牌”字的中文编码值中第一个字节的十进制数值，198 是相应的第二个字节的十进制数值。字的图像中文编码值还可用一个占两个字节的十六进制数来表示，比如说，“牌”字的图像中文编码值为 0xC5C6。因此，用 \CJKchar[GBK] { "C5 "C6 } 也能获取“牌”字。字的图像还有用 UTF8 格式的编码，UTF8 编码值是用一个占三个字节的十六进制数来表示的，还用“牌”字来说吧，其 UTF8 编码值为 0x724C。 \CJKindent 是一个采用“经典”的中文段落开头空格的命令。

有些命令还有选择项，在这里没有被列出来。

更多的命令和说明在 CJK 宏包里的 doc 文件中。下面我们费点功夫来安装中文全真字库。与英文全真字库的安装类似，中文全真字库的安装需要制作相应的 tfm 和 fd 文件以及字体映射文件。我们以 Windows 系统中的宋体全真字库“simsun.ttf”为例在 Redhat Linux 系统上进行安装。

首先用以下命令制作标准宋体和斜体宋体的 tfm 文件。

```
tfm2tfm simsun.ttf -P 3 -E 1 gbksong@UGBK@
tfm2tfm simsun.ttf -s 0.167 -P 3 -E 1 gbksongsl@UGBK@
```

其中需要一个存有 GBK 编码的文件“UGBK.sfd”。

再修改前面提到的字体描述文件“c19song.fd”如下，并保存覆盖原文件。

```
\def\fileversion{4.5.2}
\def\filedate{2003/03/28}
\ProvidesFile{c19song.fd}[ \filedate \space \fileversion]
\DeclareFontFamily{C19}{song}{\hyphenchar\font\m@ne}
\DeclareFontShape{C19}{song}{m}{n}{<->CJK*gbksong}{}
\DeclareFontShape{C19}{song}{bx}{n}{<->CJK*gbksong}{}
\DeclareFontShape{C19}{song}{m}{sl}{<->CJK*gbksongsl}{}
\DeclareFontShape{C19}{song}{bx}{sl}{<->CJK*gbksongsl}{}
\endinput
```

接着，编写一个包含如下两行的字体映射文件“ttfonts.map”：

```
gbksong@UGBK@ simsun.ttf Pld=3 Eid=1
gbksongsl@UGBK@ simsun.ttf Slant=0.167 Pld=3 Eid=1
```

并保存在目录“/usr/share/texmf/ttf2pk/”下。

然后，还要编写另外一个字体映射文件“gbksong.map”，该文件只包含类似如下两行的语句，共有 188 行。

```
gbksong??<gbksong??.enc<simsun.ttf
gbksongsl??<gbksong??.enc<simsun.ttf
```

并把该文件保存在目录“/usr/share/texmf/pdftex/config/”下。同时还要新建 94 个与文件“UGBK.sfd”中的汉字编码对应的文件“gbksong??.enc”，并把这些文件保存在目录“/usr/share/texmf/dvips/chinese/gbksong/”下。以上??代表 01，02，一直到 94，与 tfm 文件名相对应。

最后，还需打开一个保存在目录“/usr/share/texmf/pdftex/config/”下的文件“pdftex.cfg”，并加入“map + gbksong.map”。别忘了把字库文件“simsun.ttf”复制到如下目录：

/usr/share/texmf/fonts/truetype/chinese

并执行“texhash”来更新文件名数据库，到此，中文全真字库安装完成。这时，就可以用 pdf_lat_ex 把 TeX 文档直接编译成 pdf 文件。

根据以上原理，借用陈向阳 (chenxy@sun.ihep.ac.cn) 的 GBK→UTF8 编码表，He Qiang (obase@163.net) 首先在修改了 FreeType 软件包中的文件“t1asm.c”、“ttf2pfb.c”和“type1afm.c”的基础上编写了一个制作第 1 类型、t_fm 和 a_fm 的中文字库及其字体映射文件的源程序“gbkfonts.c”，后来该程序由 Wang Yin (wang_y01@mails.tsinghua.edu.cn) 作了修改。通过网络检索，我们可以找到该程序被编译后得到的执行文件，这个执行文件使用起来很方便。

有关 LaTeX 的参考文献很多，这里没有一一列出。大家看过本文提到的文献之后就会对由 Stanford 大学 Donald E. Knuth 发明的 TeX 的历史和未来及其使用有了全面的了解。目前，CCT 还是很

好的中文 LaTeX 排版系统之一。不过，据统计，由 Werner Lemberg (wl@gnu.org) 设计的 CJK 宏包是现在最受中国用户欢迎的编译中文文档的 TeX 排版系统，可从他的主页 (<http://cjk.ffi.org>) 上下载，也可从 CTAN (<http://tex-archive/language/chinese/CJK/>) 上找到。值得高兴的是，中文 Turbo Linux 系统已经自带 CJK 宏包和中文字库，用起来就很方便了。希望大家能从中受益！

致谢：作者非常感谢胡国权博士对本文提出过宝贵的建议。

参考文献:

- [1] GOOSSENS M, MILLELBACH F, SAMARIN A. The LaTeX Companion[M]. Boston: Addison-Wesley Press, 1994.
- [2] TOBIAS O. The Not So Short Introduction to LaTeX[EB/OL]. Available online from CTAN: <http://tex-archive/info/lshort/english/lshort.pdf>, 1997.
- [3] 罗振东, 葛向阳. 排版软件 LaTeX 简明手册[M]. 北京: 电子工业出版社, 2001.

Compilation of Chinese TeX Text Files by Use of CJK Packages

JIANG Zheng lu

(Department of Mathematics, Sun Yat sen University, Guangzhou 510275, China)

Abstract: CJK is a LaTeX macro package used to deal with Asian fonts such as Chinese, Japanese, Korean and Thai. LaTeX is an outstanding typesetting system. In this paper, we show the principle of LaTeX compilation of Chinese fonts and introduce how to write a Chinese TeX text file by use of the format of the package, in order that latex (or pdf_lat_ex) directly compiles the TeX file into a dvi (or pdf) file. Also, the installation of Chinese TrueType fonts is described.

Key words: LaTeX typesetting system; CJK