LaTeX's Index Processing

Maurizio Zocchi

This note suggests an easy way to process a LaTeX index. As known, index files produced by the LaTeX system involve a lot of manual work to obtain the final form of an index.

In order to do that operation automatically we need a program able to: (1) sort the index's entries alphabetically, (2) remove possible duplicate keys, (3) collect all page numbers related to each entry.

Such a program will permit a simple two-step method to correctly construct a document's index. When LaTeX processes a document, it can produce index files, which may be adjusted by the program mentioned above. During the next processing, LaTeX will find a ready-to-use index file.

Obviously, sorting of an index's entries can be performed by a standard utility routine, but data must be provided with a specified form. Sort routines can accept specification on key positions, lengths and other information to 'direct' the process, but in the simplest case there is one key, as long as the entire record. In this case, the index's lines have different lengths so the sort routine is not directly applicable. The index is a normal ASCII file and consists of several lines in the form

```
\indexentry{title}{page no.}
```

and processing the variable length key '`title`' may lead to wrong results, because of the criteria by which characters are sorted. A way to avoid this is to fix the initial key position and length to standard values. Of course, this solution occupies much memory space and limits the generality of the application.

Another solution is to estimate the maximum key length and then fill all the smaller title record keys with blanks; the page numbers must also be adjusted with leading zeros. In this way, the sort routine will compare the data keys exactly, without any misplacement.

Furthermore, accented letters need particular care, because every accent is composed by calling the `\accent` primitive or an alignment command and their presence in the index can cause an incorrect result of the sort.

It is supposed here that accents are written in the index in their unexpanded form, so it is possible to invert the position of the characters composing every accent to partially preserve the alphabetical order. For example, the sequence '`\'E`' becomes '`E\'`'; this is particularly important when the accent appears in the first position of the key.

After the sort it is necessary to merge all the page numbers related to each title key, and then remove blanks and zeroes. For example, two entries that appear in the index as

```
\indexentry{Guess     }{0020}
\indexentry{Guess     }{0122}
```

need to be 'compacted' into the form

```
\indexentry{Guess}{20, 122}
```

that is ready to be processed by LaTeX. Of course, all accent control sequences are inverted again and the result is the initial situation.

Everything described above is implemented under MS-DOS and VAX/VMS. Let me define more precisely what kind of programs were developed.

Under the VAX/VMS system there was developed a program called INDTEX that can execute the three steps mentioned, provided that sorting is on system charge, too. The sort routine is called within the program.

Under MS-DOS a Pascal program was developed for each step of the process; these programs are:

- INDTEX — performs 'expansion' of input lines.
- COMPACT — provides 'reformatting' of files after sort.

It seems important to remember here that the MS-DOS sort cannot handle files with sizes larger than 64KB; so INDTEX splits output into different files, which are sorted separately. Then, a MERGE program establishes the correct situation. However, users can deal with simple batch commands.

INDTEX was applied during the construction of a publishing house catalog of 65 pages. The catalog includes a title index, index of arguments, and author index. The index files were constructed by using the `\contentsline` form, and the line structure was expanded to include the volume code and the author's name. The final document was 97 pages.

The total time required was about 45 minutes for processing text and about an hour for sorting and adjusting the indexes.

In the future, ASCII accented characters may be used to replace a LaTeX command by a single character. Something will be done to improve the execution. Special thanks to A. Mattasoglio of Cilea and G. Canzii of TECOGRAF.

For further information or suggestions (which are welcome), please contact

M. Zocchi
c/o TECOGRAF
via Plinio 11
20100 Milano Italy