CABIOS

Vol.8. no.2. 1992
Pages 196-198

REF-LIST: program to list references found in DOS text files of scientific manuscripts

John A.Byers

Abstract

A computer program, coded in QuickBASIC® language, is used to make a non-redundant list of references, in alphabetical order, that are found in a scientific manuscript (formatted as a DOS text file). The program can extract either direct or indirect citations as in the following: Byers (1983a, b, 1984) and Miller and Keen, 1960; Byers et al., 1984; Byers 1984) An IBM-compatible personal computer is required to run the executable program.

Introduction

Most scientific journals and books have a similar style for referring to citations of published research articles and reviews. Two basic ways are used to refer to citations and both use the author(s) and year of publication. The first way to reference is direct, for example: 'Miller and Keen (1960) summarized many reports of natural densities of attacking [bark beetles] and found them to vary from 5.9 to 23.2 per 0.1 m^2 but "always within certain limits".' The second way is indirect: 'The mechanisms that produce these attack distributions are virtually unknown, although acoustic/stridulatory (Rudinsky and Michael, 1973), olfactory (Byers, 1983a, b; Byers *et al.*, 1984) or a combination of mechanisms may be involved (Rudinsky *et al.*, 1976; Hedden and Gara, 1976)' (cf. Byers, 1984). During development of a research paper one must find these direct and indirect references and compare them to the 'References' or 'Literature cited' section. Missing references in the text or 'References' section must be added to the paper, while unneeded references must be deleted. Also, the spellings of the author names must be consistent and correct. Several comparisons may be required until the final version of the manuscript is accepted.

A program REF-LIST.BAS is reported here, coded in the BASIC programming language (QuickBASIC® 4.0 or 4.5, © 1982-1988, Microsoft Corp.), which when compiled to REF-LIST.EXE runs on IBM-compatible computers from the DOS command line. The program can take DOS text files of scientific manuscripts and make a non-redundant, alphabetical list of references in only a few seconds. The list then can be compared to the references cited section. The program can recognize both direct and indirect references that have various punctuation styles (, ; or space) as well as letters representing dates (as above).

The program, with little or no modification, can run on other computers such as Macintosh that have QuickBASIC installed. The source code can be easily modified for other computer brands using other versions of BASIC because of the restricted use of commands and simple output statements. The program supports both monochrome and color monitors and does not require graphics capabilities.

Algorithm

The program consists of (i) an input menu, (ii) a reference location and extraction algorithm, (iii) a binary search, aphabetical listing algorithm and (iv) a screen output and printing and/or text file saving facility (Figure 1). The input menu (part 1, Figure 1) requests the user for the name of the file with the manuscript in which to locate references. The file with the manuscript should normally be a DOS text file, though WordPerfect® (v. 4.2-5.1, WordPerfect Corp.) files can usually be processed directly. Almost all word-processing programs, including WordPerfect, have facilities for importing and exporting their manuscript files as DOS text files (ASCII text). The program also requests whether the subsequent reference list should be (i) printed, (ii) saved in a sequential file or (iii) both. The name of the output file in part (ii) must be entered.

The program opens the text file as a random access file and loads the text into memory in 32K portions as a string variable. The algorithm for locating references in the manuscript (part 2, Figure 1) searches sequentially through the text string for a pair of left and right parentheses and extracts a smaller text string between these characters. This text string is processed so that carriage returns (ASCII no. 13), line feeds (no. 10) and extra spaces are removed. A date value between 1700 and 2010 is searched for in the smaller text string; if none is found, then the string is not a citation. If the string is a citation, then it is determined whether the string contains only dates, or authors and dates. The dates, and if present the authors, are placed in separate strings for later use. If the smaller text string contains only dates, then authors are extracted from a preceding string of 50 characters in length. Three 'words' then are obtained by stepping back toward the beginning of the preceding text string

Department of Ecology, Lund University, 223 62 Lund, Sweden

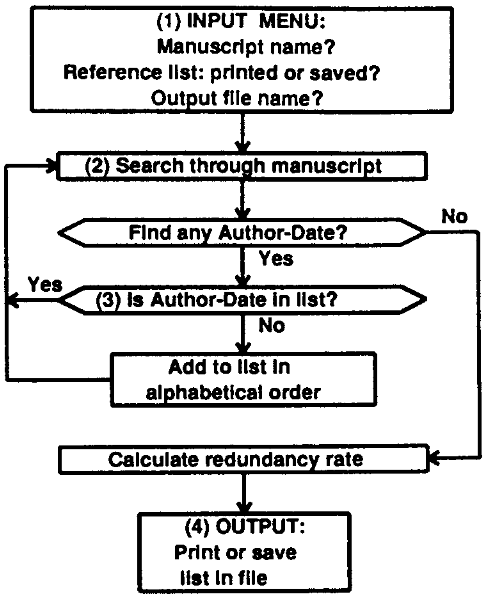
(1) INPUT MENU:
Manuscript name?
Reference list: printed or saved?
Output file name?
(2) Search through manuscript
Find any Author-Date?
No
Yes
Yes
(3) Is Author-Date in list?
No
Add to list in alphabetical order
Calculate redundancy rate
(4) OUTPUT:
Print or save
list in file

Fig. 1. Flowchart of REF-LIST program.

and using spaces as delimiters. The words are combined to form the author string, depending on whether the second word is 'and' or '&', or if two of the words are 'et' and 'al'.

The next step is to combine the author strings with the appropriate dates. The algorithm finds the first 'author word' and then adds possible 'et' and 'al' strings or 'and' or '&' strings as well as a second author. Then the first date is joined to the existing author string and this string is compared to a 'dictionary' of author–date strings. A bindary search algorithm is used to find any identical strings in the dictionary (part 3, Figure 1). If none exist, the current string is added in alphabetical order to the dictionary. The program proceeds to find any more dates, or single letters that indicate a date, and adds these to the author string and then tests for their presence in the dictionary. If another author string instead of a date is found next, subsequent dates are added to this author string rather than the previous author string.

When the entire file has been processed, the program lists the author–date strings from the dictionary on the screen, and either prints the list or saves it on disk (part 4, Figure 1). While the dictionary is being built, a running total of the number of unique (non-redundant) citations is made. Also, a running total of all citations is done. These values are reported, as well as an average redundancy rate. This rate is based on the ratio of the number of total citations to the number of unique citations. A value of 1 would indicate that all references were cited only once, while a larger value would mean that references were cited many times throughout the paper.

Implementation

It is expected that normally the program will be run as a compiled, executable file called REF-LIST.EXE. The user begins by entering the name REF-LIST at the DOS command line. A list of file names in the current directory is shown as an aid in selection of the file to be processed for references. The user must then enter a file name, including a possible path to another directory. Several options then are available as to whether the reference list will be output on either the (i) printer, (ii) saved in a file on disk, or (iii) both.

Table I. Output file of list references found in the text of this paper by the program REF-LIST

```
Byers 1983a
Byers 1983b
Byers 1984
Byers et al. 1984
Hedden and Gara 1976
J.A. 1984
Miller and Keen 1960
Rudinsky and Michael 1973
Rudinsky et al. 1976
CITES: Total = 15  Unique = 9  Redundancy(Total/Unique) = 1.666667
```

The references are located in the text of the Abstract and introcuction except for J. A. 1984 which is located in the References.

Assuming options (ii) or (iii) have been selected, the user enters the name of the file where the list will be saved on disk. If the file specified already contains code or text, the program asks if the file can be overwritten. The user may exit without changes or continue and replace the contents of the file with a subsequent list. Assuming no file exists, it is empty, or the file can be replaced, the program loads the DOS text file and makes a list of references in the file specified above. This file then can be looked at with the DOS TYPE command, imported into a word processor, or printed with the DOS command: COPY filename PRN.

The program is very fast both in QuickBASIC or compiled, taking < 1.5 s to process a 44K file (23 pages, double spaced, 12 pitch) containing 52 distinct references cited 83 times (a redundancy rate of 1.6). The program was used on a DOS text file of the present manuscript and the resulting list of author citations is shown in Table I. The program found all relevant citations. The nine citations (Table I) were cited 15 times, so that each reference was cited an average of 1.7 times. However, in some journals, such as *Computer Applications in the Biosciences*, the citations in the 'References' section have parentheses around the year date, which causes the program to list them erroneously (e.g., J.A. 1984, Table I). This problem can be avoided simply by using the word processor to remove temporarily the 'References' section. A DOS text file of this shortened version can then be saved with a word processor and REF-LIST used to list the references.

There are no commercial programs that will find references in a manuscript and make a non-redundant list of them. Thus the program here will aid scientists during publication of their research manuscripts by saving time and by improving the accuracy of the literature citations. The program, including QuickBASIC source code, executable file and text of the paper, is available from the author. Please send a DOS formatted floppy diskette, with mailing protection cover, and sufficient postage money for airmail.

Acknowledgements

I am grateful to Fredrik Schlyter who commented on the manuscript.

References

Byers,J.A. (1984) Nearest neighbor analysis and simulation of distribution patterns indicates an attack spacing mechanism in the bark beetle, *Ips typographus* (Coleoptera: Scolytidae). *Environ. Entomol.*, **13**, 1191–1200.

Circle No. 15 on Reader Enquiry Card