USES OF FULL TEXT DATABASE - AN EXPERIMENT WITH UNIX

A.A. Vaishnav, N.G. Bapat & V.N. Deo

ABSTRACT

Considerging the development in IT and availabilty of textual databases in electronic form, attempts needs to be made to know the techniques of using them. The study attempts to derive index terms from the text, assigns the subject headings to an index stored in a data file to get local documentation list, to provide CAS. Provides facilities for searching and retrieval of the text.

INTRODUCTION

The databases are the textual or numeric data in machinereadable form, which is processed for electronic dissemination.

The databases can be classified into following categories (Wagner & Landau, 1980).

1. Reference databases

Contain reference or secondary information that indentifies various primary information sources. The reference databases are of two types:

a) Bibliographic Databases: Contain bibliographic reference of citations with or wothout abstracts to the published literature sources such as journals, books, magazines, newspapers, reports, patents or theses. Bibliographic databases are most frequently used libraries.

b) Directory databases: Contain references, with or withot abstracts or summaries to people, organizations, grants, research projects, contracts, etc.

2. Sources databases

Source databases contain complete primary information. There are three basic types of source databases:

a) Numeric databases: Contain statistical or other numeric data. In some cases the numeric data may be statistically manipulated on-line to produce customised tables, graphs ratios, etc. Economic time series are a common type of information found in numeric databases. Numeric databases may contain textual data.

b) Directory databases: Contain hand book or dictionary information such as definitions, chemical nomenclature, physical properties, etc. There are relatively few of these dictionary databases.

c) Full text database: Usually contain the complete text of a document, Such as a court decision, a law or a newspaper or magazine article. They may contain numeric data as well.

Scope of the study:

Present study is limited with the uses of full text databases.

A Delphi study (Lancaster,1982) forecasted that by the year 2000 A.D. 50% of indexing and abstracting journals, 25% of books, 90% technical reports will be available in electronic form.

Now most of the journals and reference books (Encyclopedia Britanica, etc.) are available in electronic form. More than 600 full text databases are on CD-ROM. In these context it is the right time to know multifaceted uses of full text databases and the techniques to use them.

The full text databases can be used for automatic indexing, to prepare local documentation list as a part of CAS and for information retrieval.

Objectives:

The study was undertaken with a view-

1. To derive index terms from the full text database.
2. To prepare and retrieve documentation list with

the help of index term derived from the text.
3. To provide full text searching facility.

Sample:

The experiment of deriving index terms were carried out on five journal articles of the different subgects from science as well as social science disciplines viz.

1. Journal of the Helminthological Society of Washington.

2. Physiological Planetarium.

3. Library Science with a slant to documentation and Information studies.

4. Journal of the science of food and agriculture.

5. International Medical Research.

Methodology:

The infrastructure available to the authors is INTEL 486 multi-uses computer with OS SCO-UNIX. Due to unavailability of textual databases attempts were made to key in full text of the five articles from the journals included in the sample each in individual data file by using VI (visual instructor) editor.

A shell script with UNIX commands cat, tr, sort, uniq etc. was developed to get the list of most frequently occurring words in the text. By omitting structure words the most frequently occurring subject terms were chosen as index terms. The index terms derived from each article were assigned as key words to respective articles for preparing an index by using a BASIC programme. Provision was made to assign seven index terms to each article and the index terms were linked with bibliographical detals of the article. The index terms alongwith bibliographic details of the article were stored in a data file doc.pro. To search the index as well as the text the shell scripts were developed with UNIX tool grep. grep is the pattern searching utility. The name grep stands for global regular expression printer. It searches given pattern by an entire file and points out the lines containing them.

Discussion and results:

To get most frequently occurring index terms UNIX sheel script (Slide-1) was used. The cat command cats the data file. The output of this command is piped as an input to tr command. The tr command transfers every word separated by blank space to a new line. This output is piped as an input to sort command which sorts the received input (words) in alphabetical order.
The sorted output is piped as an input to the uniq command. The uniq command with -c option compresses each group of identical words into one line prefixed by a count. The output of this command is piped as an input to sort command. The sort command with -r option arranges uniquewords in reverse order of frequency. the output of this command is redirected as an input to the files namely f1, f2,.....fn.(one file for each article). The output of the file is piped as an input to the lp command which gives the printout of words appeared in text in reverse order of frequency. The most frequently occurring kernel terms (slide-1) were chosen as index terms.

A comparative study of the sample articles under study was made to find out compatibility of the headings assigned to the articles and derived from the articles as shown in table-1 (slide-2).

TABLE-1

COMPATIBILITY OF ASSIGNED TERMS WITH DERIVED TERMS :

Article	Assigned Term	Derived Term
1. Monoecocestus Centroovarium Sp.n.(cestoda: Anoploocephalidae) from Attwater's Pocket Gopher, Geomys attwateri from the san Antonio Area of Texas	Monoecocestus	Monoecocetus(9)
		Proglottids (8)
	Centrovarium	Centrovarium(3)
	Cestoda	Helmintus (3)
	Anoplocephalidae	Anoplocephalidae(4)
	Geomys	Geomys (8)
	Attwateri	Attwateri (3)
	Texas	Texas (7)
71% matching		
2. Polyphenolic auxin protectors in buds of Juvenille and adult chestnut	Auxin protectors	Auxin (13)
		protectors (3)
	Buds	Jevenille (15)
	Castanea sativa	C.Sativa (1)
	Castanea crenata	C.crenata (1)
	Catechin	Catechin (4)
	Chesnatin	Chesnatin (4)
	Chesnut	Chesnut (6)
	Crenatin	Crenatin (9)
	Cretanin	Cretanin (5)
	Polyphenols	Polyphenolic(1)
		Rooting (6)
82% matching.		
3. Authorship Trend and Solo Research in Bibliometrics: A bibliometic study.	Bibliometrics	Bibliometrics(29)
	Authorship	Authorship (12)
	Pattern	Pattern (9)
	Solo Research	Single (12)
		Research (10)
80% matching.		

4. Effect of	Dynamics testing	Dynamics	(5)
		testing	(3)
Stailing on	Water binding	Waterbinding	(3)
Viscoelastic	Capacity	capacity	(4)
Properties of	Amylose	Amylose	(5)
Pastes	Sensory	Sensory	(5)
Prepared from	Staling	Staling	(8)
Arabic Bred	Arabic bred	Arabic	(7)
		Bred	(17)
		Viscoelastic	(5)
		Pastes	(7)
		Gluten	(6)

100% matching.

5.Twice weekly	Thyroxine	Thyroxine	(30)
Dosing for	Primary	Primary	(6)
Thyroxine replace-	Hypothyroidisim	Hypothyroidisim	(8)
ment in Elderly	Elderly	Elderly	(9)
Patients with		Weekly	(15)
Primary		Patients	(15)
Hypothyroidism		Twice	(13)
		Dosing	(13)
		Intermittent	(11)
		Level	(9)
		Treatment	(8)
		Thyrotrophin	(8)
		Triodothy	(7)
		Therapy	(7)
		Thyroid	(6)
		Pep.net	(5)

100% matching.

Note: Figures in bracket indicate frequency of terms in the text.

To prepare a documentation list index terms derived from each article were assigned to respective articles for preparing the index. To store the index in a data file (doc.pro) a BASIC Programme was developed which has the facility to create, append, read and modify the data file. The data was keyed in with the screen section using BASIC programme (Slide 3 & 4).

To circulate the documentation list a print out of the data file doc.pro was taken with the help of BASIC programme as shown in slide-5.

Searching and Retrieval:

To search the information from the data file doc.pro or from the textual database of the articles the UNIX shell script with grep was used. To search the text pre and post truncation of the index terms is presupposed in grep. As grep finds strings, it does not care whether the string forms a complete word or just part of one word (slide-6). Hence the search expression can be given as-

grep chloroform art2.

This search expression will retrieve the lines with the words chlorophen, chlorophenol, di-chlorophenol, while the serch expression-

grep chrom* art2

will retrieve the lines with chromatography,

chromogenicity, chromatograms (Slide-6).

Searching multiple files simultaneously :

It is possible to search a term in multiple files. In this case each file is searched in succession. When matching lines are printed they are preceded with the file name to indicate which file contains which file (Slide-6). e.g.

grep cestoda art1 art2 art3 art4 art5.

Coordination of terms :

Though word indexing (uniterm) technique is used, coordination of any number of terms is possible in search expression (Silde-7).e.g.

grep "waterbinding capacity" art4

will retrieve the lines containing words waterbinding capacity.

Boolean Expressions :

It is possible to use boolean opertors (AND, OR, NOT) in search expression. To use AND operator the search expression is to be formed as-

grep thyroxine art5 | grep intermittent |
grep therapy | more

Here | (pipe) works as AND operator (Slide-7).

To use OR operator the search experssion can be given in two ways (Slide-7)-

1. fgrep monoecocestus.
> centroovarium
> attwateri art1 or
egrep 'monoecocestus | centrovarium | attwateri' art1.
(Slide-8)

Here single term on one line or the pipe within quote in search expression works as OR operator.

It is possible to frame a search expression with combination of operators.

e.g. grep chestnut art2 |grep adult | grep-v auxine.
here -v works as NOT operator (Slide-8).

Display of numbers of hits :

The search expression-
grep -c bibliometrics art2
displays number of postings i.e. number of times a term occurs in the text (Slide-8).

It is possible to store the terms to be searched in a

file. This term file can be matched with the database for searching (Slide-8). e.g. f1 is the file which

contains the terms to be searched viz.

intermittent
therapy
elderly

these terms can be matched with the search expression given in a shell-

fgrep -f f1 art5 | more

The -f is an option to collect the patterns from a file f1.

Conclusion :

Considering the developments in IT and its applications it is the right time to develop computer assisted indexes. Moreover, as the forecast of F.W. Lacanster is coming into vogue, it is the right time to develop techniques for full text searching of the mechanized databases hence the present study was taken up.

The experiments performed in this context with the available infrastructure shows that-

1. Though assigned indexing uses controlled vocabulary 71-100% (table-1) derived from the article fall under controlled vocabulary which means derived indexing system being faster than assigned indexing system can be used as a gap filling mechanism between the appearance of source article (on CD-ROM) and its appearance in indexing and abstracting journal.

2. The generated documentation list can be used to provide current awareness service.

3. With the help of search expressions written in grep it is possible to search given information in documentation list as well as in the source article.

To conclude, it can be said that automatic indexing and full text searching in UNIX environments is powerful and faster information retrieval technique.

References

1. Lancaster, F.W. and others, The Impact of Paperless Society on the Research Library of the future. Quoted by F.W. Lancaster. The Future of the Library in the Age of telecommunications in changing Information concepts and Technologies : AReader for the Professional Librarian. New York, Knowledge Industry Publications,1978.

2. Muster, J and others, UNIX Power utilities, New Delhi: BPB Publications, 1989.

3. Prasher, R.G., Index and indexing Systems, New Delhi: Medallian Press, 1989.

4. Prata, S., Advanced UNIX, A Programmer's Guide, New Delhi: FPB Publications, 1986.

5. Rijsbergen, C.J. Van, Information Retrieval, London: Butterworths, 1979.

6. Rowley, J.E., Abstracting and Indexing, London: Clive Bingley, 1982.

7. Rowley, J.E. and Turner, M.D., The Dissemination of Information, London : Andre Deutsch, 1978.

8. Tedd, L., Introduction to Computer-based Library Systems, New York : John Wiley, 1985.

9. Wagner, J. and Landau, R.n.: Non-bibliographic On-line Data Base Services. Journal of the American Society for Information Scicnce. 28(1) Jan, 1977, p.13-18.


```
Cat art | tr" " "/012" |sort | uniq -c |sort -r |
more
91
35 )
33 (
23 of
23 and
20 in
18 the
15 from
13 long
12 wide
12 by
11 m.
19 .
10 to
10 species
9 monoecocestas
8 Proqlottids
7 texas
8 qeomys
6 excretary
5 species
5 ovary
5 located
```

(Slide-1)

DOCUMENT PROFILE

SERIAL NUMBER :

KEYWORD1 :
KEYWORD2 :
KEYWORD3 :
KEYWORD4 :
KEYWORD5 :
KEYWORD6 :
KEYWORD7 :
AUTHOR1 :
AUTHOR2 :
TITLE1 :
TITLE2 :
TITLE3 :
JOURNAL :
YEAR :
VOL (Issue No.) :
PAGES (From - To) :

(Slide-3)

DOCUMENT PROFILE

SERIAL NUMBER : 1
KEYWORD1 : Monoecocestus
KEYWORD2 : Proglottids
KEYWORD3 : Texas
KEYWORD4 : Geomys
KEYWORD5 : Douthiff
KEYWORD6 : Anocephaloides
KEYWORD7 : Helminths
AUTHOR1 : Helminths
AUTHOR2 : and others
TITLE1 : monoecocestus Centro-
TITLE2 : varium sp.n.(Cestoda:
TITLE3 : Anoplocephalidae) from
JOURNAL : J.Helminthologial Soc.
YEAR : 1994
VOL (Issue No.) : 61 (1)
PAGES (From - To) : 61-63

(Slide-4)

DOCUMENTATION LIST

1. Monoecocestus Proglottids Texas Geomys Centroovarium Attwateri Douthitt
Dronen,N.O. & others : Monoecocestus Centroovarium
sp.n. (Cestoda : Anoplocephatidae) from Attwateri Pocket Gopher, Geomys attwateri, from the San Antonio Area of Texas.
J. Helminthological Society of Washinqton: 61(1) 1994, p.61-63.

2. Juvenile Auxine Crenatin Roqting Chestnut Cretanin Chesnatin
Mato, M.C. 7 others : Polyphenolic auxin Protectors in buds of Juvenile and adult chestnut.
Physioloqia Plantarum, 19(1) 1994; p.23-26.

3. Bread Arabic Viscoelastic Staling Water-binding Capacity Dynamic Testing
Toufili, I & others : Effect of stailing on Viscoelastic properties of pastes prepared from Arabic Bread.
J.Science of Food and Aqriculture 64(3) 1994; p.271-273.

4. Bibliometrics Single Authorship Pattern research Kabir, H:
Authorship Trend and Solo Research in Bibliometrics : A Bibliometric study.
Lib.Science with Slant to Documentation and Information studies 31(2) 1994; p.87-90.

5. Thyroxine Weekly Twice Dosing Intermittent Elderly Triiodothyronine
Taylor, J & others : Twice - Weekly Dosing for Thyroxine Replacement in Elderly Patients with Primary Hypothyroidism.
J.of International Medical Research, 22(5) 1994; p.273-77.

(Slide-5)

1. grep chestnut art2
Polyphenolic auxin protectors in buds of juvenile and adult chestnut.
Chestnut lends itself to a study of this problem since cutting from
and adult plants of chestnut was carried out in an attempt to explain
adult chestnut - (castanea sative x c.crehata close hv)
grown in the
All phese compounds were previusly found in chestnut - galls induced by chestnut.

2. grep chrom art2 | more
Were monitored by paper chromatoqraphy with butanol : ethnol :
water (40 : 10 : 2.2)
v/v/v/ as solvent and sprayed diazotised benzidine the chromoqenic positive.
residue was dissolved in 1 ml of methanol and chromatoqraphed on 3 mm paper.
co-chromatoqraphy with authentic markers in several solvents.
from the chromoqenic positive bands the mixture were incubated at 30 degree.
when paper chromatoqrams of the active fractions were sprayed with
diazotised benzidine and chromoqenically positive bands were detected at spectral and chromatoqraphic analyses since an authentic market was not
the elutes of the and chromoqenically positive bands tested for auxin.

3. grep cestoda art1 art2 art3 art4 art5
art1 : monoecocestus centroovarium sp.n. (cestoda:anocephalidae) from

(Slide-6)

4. grep 'water-binding capacity' art4
water-binding capacity was measured according to the procedure ofaccord with the sharp decrease in the water-binding capacity of breadin water-binding capacity increases the proportion of mobilewater in the

5. grep thyroxine art5 | grep intermittent | grep therapy | more
intermittent thyroxine therapy have examine young fit individuals to whom
may safely be given intermittent thyroxine therapy

6. fgrep ' monoecocestus <>
> centroovarium <>
> attwateri' art1
* species of monoecocestus bedard 1914 was found six species of monoecocestus
variabilis douthitt 1915 from c.dorsatum monoecocestus anoplocephaloides
Materials and methods : Six specimens of geomys attwateri were trapped
an undescribed species of monoecocestus
monoecocestus of centroovariium sp.n. description (based on type host : geomys attwateri tucker and schmidly 1981
of the ovary this species of monoecocestus known from north american
geomys breviceps most closely resemble m.centroovarium sp.n. in general monoecocestus

(Slide-7)

* Monoecoeestus centrovarium Sp.n. (cestoda annoplocephalidae) from attuvater's pocket gopher geomys attwateri from the san antonio area of Texas.

7. egrep 'monoecocestus | centroovarium | attwateri' art1
monoecocestus centroovarium sp.n. (cestoda : annoplocephalidae) from
attwater's pocket gopher geomys attwateri from the san antonio area of texas
species of monoecocestus bedard 1914 was found six species of monoecocestus
variabilis douthitt 1915 from c.dorsatum monoecocestus anoplocephaloides
Materials and methods : Six specimens of geomys attwateri were trapped
an undescribed species of monoecocestus
monoecocestus of cetroovarium sp.n. description (based on type host: geomys attwateri tucker and schmidly 1981
of the ovary for this species of monoecocestus known from north american
geomys breviceps most closely resemble m.centroovarium sp.n. in general monoecocestus

8. grep chestnut art2 |grep adult | grep -v auxine
and adult plants of chestunt was carried out in an attempts to explain

adult chestnut (castanea saliva x c.crenata clonehy) grown in the

9. grep -c bibliometrics art3
23
grep -c bibliometric art3
27

10. fgrep -f f6 art5 :
intermittent thyroxine therapy have examined young fit individuals to whom
may safely be given intermittent thyroxine therapy

(Slide-8)