計量文献学

計量文献学で用いられるパソコンソフト

浅井　勇夫*

計量文献学におけるランク頻度分布や年齢分布は，オンライン情報検索を利用して，収集可能である。この論文の目的は，得られた分布データを分析するパソコンソフトを開発することにある。ブラッドフォード分布，ロトカ分布，ジップ分布，寿命分布，そして成長分布を扱う。主要な機能として，スクリーン上でのデータの編集，カラー画面へのグラフ表示，複数の分布グラフの重ね合わせ表示，各数学モデルのパラメータの統計的推定を含む。計量文献学へのパソコンの利用は，情報の特性を表示したり，分析したりするのに効果的である。

1．はじめに

情報科学分野における情報処理技術は，大型コンピュータの発達とともに，飛躍的に発展しているが，それに比べると，情報現象を科学的に解明する研究は，あまり進展していない。それは，研究の第一歩である計量的なデータの収集が，困難なためである。計量文献学で利用される，測度の代表的なものとして，雑誌や著者の集中・分散を表わすランク頻度分布，論文の寿命や成長を表わす年齢分布，そして論文の引用文献から得られる引用・被引用文献数などがある。これらの測度の特徴は，いずれも連続量ではなく，離散的なデータであり，時間が時分秒でなく年単位である。したがって，計量文献分析を行なうには，数十年分のデータの蓄積が必要になり，データ収集が研究のネックになっていた。

しかし，オンライン情報検索を利用すれば，定量的なデータの収集問題は，ある程度解決する。莫大な文献を蓄積するデータベースの中から，研究に必要なカウント・データを，検索式を用いて求めることが可能である。オンライン情報検索で利用できる各種のデータベースは，計量文献学研究の重要なデータ源とみなすことができる。このように，データは容易に収集できるようになったが，その処理を手作業で行なえば，詳細な分析は不可能である。この論文では，計量文献学におけるデータ処理の問題を解決するために，従来よく研究されてきた，ランク頻度分布と年齢分布の分析を，最近，急速に発展・普及している，パソコンを用いて処理する，ソフトを開発することを目的とする。

2．ソフト開発

最近のパソコンの進歩はめざましく，漢字・グラフィック・カラー表示などの処理が，簡単にできるようになった。ソフト技術はハード技術と同様に，日進月歩しているため，機能を特定することは困難であるが，次のような条件を満たすソフトを開発した。

＊　あさい　いさお　大阪府立大学

2.1 ハード機器構成

① 16ビット本体（漢字ROM付）
PC—9801

② カラー専用高解像度ディスプレイ
PC—8853

③ 漢字ROM付プリンタ
PC—8822

④ 8インチフロッピーディスク
PC—8881

2.2 使用言語

N88—BASIC（86），一部マシン語（8086）

2.3 ソフト機能概略

① フロッピーディスクを使用して複数の分布データを記録・分析する。

② モノクロ用とカラー用のどちらのディスプレイにも対応可能にする。

③ 分析はジョブメニューやファンクションキーの選択により進行する。

④ データ作成や編集はスクリーン・エディットにより行なう。

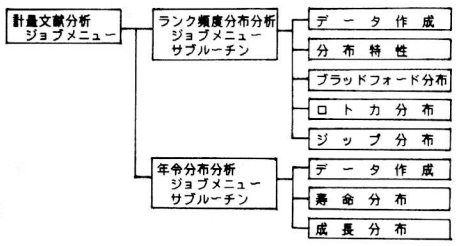
⑤ 入力データや出力結果などの作表は，画面やプリンタに出力する。

⑥ 分布グラフは640×400ドットの画面やプリンタに出力する。

2.4 プログラムの構成

全体は，11種類のプログラムで構成されているが，その関係を第1図に示す。ジョブメニュー

第1図 プログラムの構造



に示されるジョブ番号を入力すると，図の右側にある下位のプログラムが，チェイン・マージされて実行される。データは，ランダムファイル上に記録される。分析の際にはデータ番号で呼び出して使用する。

3．ランク頻度分布分析

計量文献学の分野でよく研究されている，ブラッドフォード分布やロトカ分布で用いるデータ構造は，すべてランク頻度とみなすことができる。累積をとるかとらないか，あるいはX軸やY軸にどのような測度をとるかによってさまざまな分布型が存在する。これらのデータ処理は，一括して扱うと便利である。

3.1 データ作成

ここでは，データの入力・追加・訂正・削除などを行なう。第2図は，データ作成時の画面を表わす。データ入力は，頻度と論文数の組データを，論文数の大きい順にキーボードから入力する。画面には，15個のデータしか表示できないが，ROLLキーを使ってデータを上下にスクロールさせれば，データを即時に画面上に表わすことができる。データを修正する場合には，カーソルを修正する箇所に移動させて，正しい数値を入力する。

ファンクションキーのリストが，作成画面の下から2行目に表示される。1と2は，ディスクからデータを呼出したり，記録したりする場合に使う。3は，データを訂正するときに用いる。4と5は，カーソルのある位置にデータを挿入したり，削除したりする。6の作表は，入力した頻度と論文数に対する累積を計算し，画面に表示する。7は，画面にある分布データをプリンタで印刷する。8の終了により，ランク頻度分布分析のジョブメニューに戻る。分布データは，125組まで入力可能である。頻度用・論文数用・結果出力用の，合計3セクタ（768バイト）単位で，ランダムファイルとしてセーブする。データ作成のプログラムは，一部分がマシン語になっているため，データのスクロール・挿入・削除などの処理は速い。

3.2 分布特性

ランク頻度分布は，情報が集中したり分散したりする現象を示している。その度合いを，単一指標を使って表わす試みがなされている。1977年にA.D. Prattが提案した測度は，経済学で使われているギニ係数に似ていることが，M.

第2図　ランク頻度分布分析のデータ作成画面

【1】データ作成　　DATBRA(11)　　【ランク頻度分布分析】
データ番号：　1　　雑誌(著者・語)総数：　326
組データ数：　24　　論　文　総　数　　：　1332
データ名　：　Bradford-34、Applied Geophysics

No	頻　度	論文数	累積頻度	累積論文	累積頻度%	累積論文%
1	1	93	1	93	0.31	6.98
2	1	86	2	179	0.61	13.44
3	1	56	3	235	0.92	17.64
4	1	48	4	283	1.23	21.25
5	1	46	5	329	1.53	24.70
6	1	35	6	364	1.84	27.33
7	1	28	7	392	2.15	29.43
8	1	20	8	412	2.45	30.93
9	1	17	9	429	2.76	32.21
10	4	16	13	493	3.99	37.01
11	1	15	14	508	4.29	38.14
12	5	14	19	578	5.83	43.39
13	1	12	20	590	6.13	44.29
14	2	11	22	612	6.75	45.95
15	5	10	27	662	8.28	49.70

1呼　出　2記　録　3名　前　4挿　入　5削　除　6作　表　7印　刷　8終　了
[ROLL UP&DOWN] データのスクロール　[→←↑↓] カーソル移動

P. Carpenter によって指摘されている[1]。ここでは，ギニ係数を計算し，生産性の高い順に累積した分布を，実スケールでグラフにかいたり，比率の表を作成したりする。異なった分布グラフを重ね合わせて表示できるので，分布に対する総合的な判断ができる。また，全部の分布データに関するギニ係数や比率の表を作成し，プリントすることも可能である。

3.3　ブラッドフォード分布

雑誌などの情報源の特性を，調べる際に使用する分布で，計量文献学の分野でいちばんよく研究されており，種々の数学モデルが提案されている。筆者は，その中の代表的な8種のモデルを統合する定式化を行ない，各モデルを位置づけるとともに，パラメータを統計的に推定する方法を提案した[2]。ブラッドフォード分布は，一般に，$y = a \log(x+c) + b$，と表わすことができる。従来のモデルは，パラメータの条件により，第1表のように5つのタイプに分類できる。また，8種のモデルと一般式とのパラメータの関係は，第2表のように要約できる。3つのパラメータが，すべて未知である第5番目の

第1表　各タイプのパラメータの条件

タイプ	パラメータ		
	a	b	c
1	未知数	1	0
2	未知数	未知数	0
3	$\frac{1}{\log(1+1/c)}$	$\frac{\log(1/c)}{\log(1+1/c)}$	未知数
4	未知数	$a \log(1/c)$	未知数
5	未知数	未知数	未知数

第2表　各モデルと一般式との関係

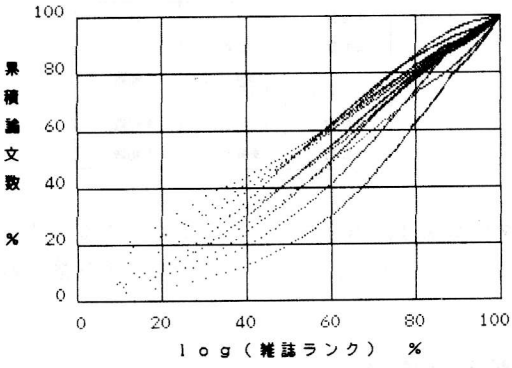
ﾀｲﾌﾟ	モ　デ　ル	パラメータ間の関係	提　案　者
1	$y = d \log x + 1$	$d = a$	Cole-62
2	$R(n) = k \log(n/s)$	$k = aR,\ s = N\ 10^{-b/a}$	Brookes-69
3	$y = \log(1 + zx)/\log(1 + z)$	$z = 1/c$	Leimkuhler-67
	$y = \log_r[(m + n)/m]$	$m = cN,\ r = 1 + 1/c$	Brookes-78
4	$y = p \log(1 + qx)$	$p = a,\ q = 1/c$	Fairthorne-69
	$R(n) = j \log(n/t + 1)$	$j = aR,\ t = cN$	Wilkinson-72
	$R(n) = f \log(1+gn)/\log(1+g)$	$f = aR \log(1 + 1/cN),\ g=1/cN$	Leimkuhler-77
5	$R(n) = h \log(n/u+1) + v$	$h=aR,\ u=cN,\ v=bR + aR \log c$	Haspers-76
	$y = a \log(x + c) + b$		Asai-81

タイプが観測データにもっとも適合することが，11種の分布データを使って確かめられた。

ここでは，分布データから各モデルのパラメータを統計的に推定し，表示できるようにした。雑誌の総数分のデータを処理するので，正確な値が得られる。全タイプのパラメータを推定するのに要する時間は，雑誌総数200で約10分である。パラメータなどの計算結果は，ファイルに記録するために一度計算するだけでよく，次回からは即時に使用できる。全部の分布データを使って，各パラメータごとの平均や分散を求めることができる。

ブラッドフォード分布のグラフは，普通片対数用紙に書かなければならないが，パソコンを使うと，グラフは数秒でプロットできる。また，任意の分布グラフを重ね合わせて，表示することもできる。第3図は，11種類の分布グラフを重ね合わせた例である。合計5,600の対数をとる点が，約2分間でプロットされる。その他の機能として，パラメータCの値をいろいろ変えて，グラフが変化する様子を調べることもできる。

第3図　11種のブラッドフォード分布の重ね合わせ



3.4　ロトカ分布

科学者の生産性を研究する場合に使い，生産性の測度として，発表論文数を用いる。統計学を基礎にした理論的な研究が，数多くなされている[3]。「成功が成功をもたらす」現象とみなして，確率モデルを使って定式化し，解析する場合が多い。

I.K.R. Raoは，5種類の分布データを使って，11種類の提案された理論モデルをカイ二乗検定し，負の二項分布が統計的にいちばん適合することを明らかにし，それを理論的にも導いている[4]。このプログラムでは，Raoが用いた11種の中から，比較的適合する3種類の理論モデルに関して，そのパラメータを推定し，カイ二乗検定できるようにした。ブラッドフォード分布は，雑誌総数分のデータがあるのに対して，ロトカ分布は，組データ数分の少量のデータしかないので，処理には工夫を要する。データ数によって，各理論モデルのパラメータが大きく変動するため，データを規準化することが必要である。分布グラフは，実スケールではプロットしにくいので，両軸は対数をとって表わした。

3.5　ジップ分布

キーワードの出現頻度を分析するときに使用する。Y軸の論文数に対して，累積をとる場合がブラッドフォード分布であり，累積をとらない場合がジップ分布である。あまり定式化されていないが，ジップ分布の数学モデルは，ブラッドフォード分布の数学モデルを，微分したもので代用することができる。しかし，キーワードの出現パターンと雑誌のそれとは，異なるかも知れないので，実データを使った統計的検定が必要である。ここでは，グラフ表示，重ね合わせ，ブラッドフォード分布モデルをもとにしたパラメータの推定を行なう。

4．年齢分布

計量文献学には，年単位に集計されたデータを用いる研究分野がある。寿命分布や成長分布がそれにあたり，X軸が時間軸を表わす。年齢をとるか年度をとるかによって，グラフ表示は異なるが，ここでは年齢分布として扱う。

4.1　データ作成

スクリーン・エディットにより，データの作成や編集を行なう。最初に，年齢か年度かの区別を入力し，つづいて時系列データを入力する。年齢分布の場合，年齢は発行年の差から求めるが，そのときに生ずる不合理な点を解消するた

めに，分布データを修正する方法が提案されている[5]。

4.2 寿命分布

資料の保管期間を設定する際に使う。文献の利用データは，貸出記録からも得られるが，最近では引用・被引用文献から集計したデータを用いる。この場合には，各年度の発行文献総数で，補正する必要がある。ここでは，寿命分布のグラフをプロットし，重ね合わせ表示する。理論モデルとしては，信頼性分析で用いられるワイブル分布を使用し，そのパラメータを推定する[6]。さらに半減期などの指標を求める。

4.3 成長分布

年度別に収集した分布を扱い，分野の成長や衰退などのマクロ的な分析をする。おもに分布グラフをえがく。

5．おわりに

現在までに研究されてきた，計量文献学の各種分布モデルを，パソコンを使って処理できるように試みた。グラフに表示したり，パラメータを推定したりする仕事は，パソコンで十分に遂行できることがわかった。高性能・低価格なパソコンの普及により，多くの人びとが計量文献学の研究に参加できるようになり，この分野の発展が期待される。オンライン情報検索時代には，ランク頻度分布や年齢分布以外の，新しい測度が必要であることを痛感する。

参考文献

1）Carpenter, M.P.：Similarity of Pratt's Measure of Class Concentration to the Gini Index, Journal of the American Society for Information Science, 30[2], 1979, p.p. 108～110

2）Asai, I.：A General Formulation of Bradford's Distribution：The Graph-Oriented Approach, Journal of the American Society for Information Science, 32[2], 1981, p.p. 113～119

3）小野寺夏生：Bibliometrics，情報管理，21[10]，1979，p.p. 782～802

4）Rao, I.K.R.：The Distribution of Scientific Productivity and Social Change, Journal of the American Society for Information Science, 31[2], 1980, p. p. 111～122

5）Asai, I.：Adjusted Age Distribution and Its Application to Impact Factor and Immediacy Index, Journal of the American Society for Information Science, 32[3], 1981, p.p. 172～174

6）松田武彦・浅井勇夫：経年的な文献価値の変化に関する研究，日本OR学会春季研究発表会アブストラクト集，日本OR学会，東京，1976，p.p. 8.9～90

質疑応答

質問　原田勝（京都大学）

生データの収集は，どのようにするのか。

回答　ここでは，データ処理の問題だけを扱ったので，生データの収集についてはふれない。

質問　上田（慶応義塾大学）

① 寿命分布の処理はどのようにするのか。

② 寿命分布においてハーフライフは求められるか。

③ 11種類のブラッドフォード分布の重ね合わせ図において，X軸を実スケールで書けないか。

回答　① グラフを表示し，ワイブル分布のパラメータを推定する。

② 求められる。

③ 実スケールにするのは簡単であるが，他の分布との比較ができなくなる。