Library and Information Science No. 26 1988

出現頻度情報に基づく単語重みづけの原理

Some Principles of Weighting Methods Based on Word Frequencies for Automatic Indexing

海　野　　敏
Bin Umino

Résumé

Characteristics of the occurrence frequency of words in natural language texts have been used as an indicator for the selection of significant words in automatic indexing. This paper describes some general principles common to term weighting methods which use occurrence frequency measures.

For this purpose, nearly sixty weighting fomulas were collected from the documents published in the past thirty years. Then their theoretical characteristics were analyzed and compared with each other. As a result, these formulas were classified into following five categories.

1) absolute frequency measures
2) two kinds of relative frequency measures
3) word dispersion measures
4) 2-Poisson model proposed by Harter
5) information theory similar to the one proposed by Shannon

Various mathematical relations peculiar to the formulas of each category were found. These relations were well explained by a model consisting of two kinds of word sets, one of which is subsumed by the other; that is, the significance of a word depended on the degree of its maldistribution to the subsumed word set.

海野　敏：東京大学大学院教育学研究科博士課程，東京都文京区本郷 7-3-1
Bin Umino, Graduate School of Education, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo.
1989年1月21日受付

I. はじめに

単語の出現頻度情報の利用は，自動インデクシング研究の流れの中では古典的な手法であり，その試みは最も早くから始められ，しかも現在まで綿々と続けられている。単語の出現頻度情報は，自動インデクシングのさまざまな局面で利用されているが，もっとも頻繁に行われているのは「単語の重みづけ」における利用である。本研究の目的は，このような「出現頻度情報に基づいた単語重みづけ」の原理を明らかにすることである。

出現頻度情報に基づいて単語に重みづけをする手順は，おおよそ次の通りである。

(1) 何が「単語」であるかを定義する。
(2) 対象となるすべての文献中のすべての単語について，それぞれの出現頻度情報を計測する。
(3) それぞれの単語の重みを，出現頻度情報から算出する。

これらの作業の中でも，この手法の中心をなすのは，いうまでもなく(3)の重み算出のプロセスである。過去30年のあいだに，研究者によって提示されてきた重みの算出方法は数十にのぼっている。

ところが，これら従来提示されてきた重みの算出式をながめてみると，利用されている出現頻度情報も，式の成立ちも，実に多種多様である。一見すると，そこに共通する原理などは，とてもありそうに思われない。

しかし，これらの一見雑多な算出式の背後には，いわば暗黙の前提として，いずれの算出式にもあてはまるひとつの考え方が隠されている。本研究は，このような共通の考え方，すなわち「原理」を，従来提示されてきた多数の算出式を整理，分析することによって明らかにすることをねらいとしている。

本稿は6章から構成されている。II章では，本研究で論じようとしている「単語の重みづけ」とは何を目的とするどのような作業なのかを説明し，同時に，多種多様な手法を共通に論ずるためのいくつかのモデルを提示する。III章では，単語の重みを算出するために用いられる出現頻度情報を整理し，その基本的な数量に記号を与える。IV章では，従来提示されてきた60あまりの重みの算出式を分類，整理し，これらの式のあいだにある相互の関係を明らかにする。そしてV章では，算出式を3つのグループに分け，それぞれに含まれる式のふるまいが，いずれも「偏在性の数量化」という共通の原理に従っていることを説明する。VI章は，本稿のまとめである。

II. 重みづけの諸相

A. 情報検索システムのモデル

単語の重みづけという作業が，情報検索システムという全体的な眺めの中でどのような位置づけにあるかを説明するためには，システムを構成する多くの要素の複雑な関係を単純化して記述したモデルを作っておくとわかりやすい。そこで，A節では文献データベースを対象とした主題検索システムを，B節ではそこで行われるインデクシング作業を，ごく単純なモデルにして提示し，C節以降での考察の準備とする。

いま，文献データベースに含まれている文献すべてからなる集合 DB と，利用者がこの文献集合に対して与えるであろう質問すべてからなる集合 QR を考える。このとき検索システムの最も基本的な機能は，「与えられた質問 q ($q \in$ QR) に最も合致した文献の集合 D ($D \in 2^{DB}$) を出力すること」と表現することができる。この機能を実現するために，検索システム内では，一般に以下のような作業が行われている。

まず，文献 d ($d \in$ DB) は，システム内である手続きに従って表現され，システムが扱いやすいかたちに変換されている。単語の重みづけは文献の主題による検索のみに関係する作業なので，ここでは表現される対象として文献の内容だけを考えることにする。この表現アルゴリズムを μ とし，μ に従って表現された d を $\mu(d)$ と表す。一方，質問 q も，システム内ではある手続きに従って表現され，システムが扱いやすいかたちに変換される。この表現アルゴリズムを μ' とし，μ' に従って表現された q を $\mu'(q)$ と表す。

システムに質問 q が与えられると，まず q が μ' に従って表現されたのちに，DB に含まれるすべての文献 d について，$\mu(d)$ と $\mu'(q)$ の合致性の度合がある手続きに従って判断される。この判断アルゴリズムを ν とする。ν は，いいかえれば d の q に対するレレバンスを評価する処理手続きである。ただし，ν の操作対象となるのはあくまで $\mu(d)$ と $\mu'(q)$ であり，d と q ではない。

そして，判断結果に基づいて質問 q に最も合致した文献集合 D がある手続きに従って決定され，q を入力した利用者にある手続きに従って表示される。以上が，文献検索システムの最も基本的なふるまいのモデルである。

B. インデクシングのモデル

情報検索システムが操作の対象とする文献は，自然言語で表現されている限り，「単語の列」とみなすことができる。それでは単語とは何かという疑問が当然生じようが，この問題は本稿では扱わない。単語の定義は重要な問題ではあるが，研究者の多様な考え方の背後にある共通項を見つけ出す本研究の目的からすれば，それぞれの研究者が「単語」と呼んでいるものが単語であると定義しておけば十分である。

初めに，インデクシングが行われる「語彙空間」に関するいくつかの記号を決めておくことにする。文献を作成するにあたって使用される可能性のあるすべての単語の集合を NL と表すことにする。これは，自然言語で用いられる語彙の集合と同じと考えてもよいであろう。また，文献データベース中の文献を構成するすべての単語の集合，すなわちデータベースの使用語彙を WD と表す。さらに，文献 d を構成するすべての単語の集合，すなわち文献 d の使用語彙を W と表す。それぞれの単語の集合に含まれる個々の単語は w と表す。

インデクシングは，前節のモデルに照らせば，アルゴリズム μ にかかわる作業である。上記の記号を用いれば，インデクシングとは「アルゴリズム μ の一部として，NL の要素である単語と，DB の要素である文献のあいだの関係づけを，システム固有の規則に従って行うこと」であると説明することができる。このような解釈からすれば，インデクシングは，単語の側から見れば「単語に文献の集合を指示させる手続き」，つまり「NL から 2^{DB} への写像」であり，文献の側から見れば「文献に単語の集合を付与する手続き」，つまり「DB から 2^{NL} への写像」である。

また，「索引語」は，「インデクシングの結果，ひとつ以上の文献と対応関係の生じる単語」と解釈することができる。「索引語」の類義語として，「キーワード」，「ディスクリプタ」，「主題語」などの用語もあるが，本稿ではこれらを用いず，以後一貫して「索引語」を用いることにする。

ところで，単語集合 NL に含まれるすべての単語が，いずれも索引語となる可能性をもっているわけではない。索引語となるには何らかの条件が必要であり，その条件を満たした単語しか索引語にはなれないのである。そこで，「索引語となる条件を満たしている単語」をすでに文献に付与されている索引語とは区別して，「索引語候補」と呼ぶことにする。索引語候補は，DB に含まれている文献，および将来含まれるであろう文献の内容を表現する資格をもっている単語である。

ここで，さらにいくつかの記号を定めておく。すべての索引語候補の集合を IT，すべての索引語の集合を

IT′ と表す。文献 d を表現するために使われるすべての索引語の集合，すなわち文献 d に付与される索引語集合を T とする。また，集合 IT の要素である個々の索引語候補と，集合 IT′，集合 T の要素である個々の索引語は，どちらも t で表す。

これらの記号を使えば，インデクシングとは「DB の要素である d に対して，その内容を表現するために，IT の部分集合である T を付与すること」と表現することもできる。

C. 重みづけの4つの目的

従来，自動インデクシングの領域で，多くの研究者たちが行なってきた単語の重みづけの試みを，前節までのモデルに照らして整理すれば，その目的は次の4つの手続きのいずれかを自動化するための測度を手に入れることにあったとまとめることができる。

α: $w(w \in \mathrm{WD})$ が IT に含まれるかどうかを判定するアルゴリズム

β: $d(d \in \mathrm{DB})$ に対し，$t(t \in \mathrm{IT})$ が T に含まれるかどうかを判定するアルゴリズム

γ: $d(d \in \mathrm{DB})$ に対し，$w(w \in \mathrm{W})$ が T に含まれるかどうかを判定するアルゴリズム

δ: $d(d \in \mathrm{DB})$ に付与された $t(t \in \mathrm{T})$ に重みを与えるアルゴリズム

これらは，それぞれ文献の表現に関わる手続きであるから，いずれもアルゴリズム μ の構成要素と考えることができよう。

アルゴリズム α は，DB 中で用いられているある単語 w が索引語となる条件を満たしているかどうかを判定する手続きである。α が WD のすべての要素に対して行われれば，索引語候補の集合 IT の要素が確定する。あらかじめ確定された IT を，何らかの規則に従って配列すれば，いわゆるキーワード・リストと同等のものを生成することができるし，さらに何らかの規則に従って IT の要素間に関係づけを行えば，いわゆる件名標目表やシソーラスと同等のものを生成することができよう。このアルゴリズム自動化のために算出される単語の重みは，DB 中で用いられている特定の単語 w の索引語候補としてのふさわしさの程度であり，これを以降本稿では，「Aタイプの重み」と呼ぶことにする。

ところで，実際の自動インデクシング研究では，文献データベースがあらかじめいくつかの主題領域に区分されているとき，それぞれの主題領域について重要語，すなわち索引語候補を選定する際に，単語の重みづけが用いられることがある。特定の主題領域に含まれる文献中の単語すべての集合を WG とすると，WG⊂WD であり，このような単語の重みづけは，

α': $w(w \in \mathrm{WG})$ が IT に含まれるかどうかを判定するアルゴリズム

なるアルゴリズム α' を自動化するためのものである。本稿では，このような重みづけもAタイプとみなすことにする。

アルゴリズム β は，ある索引語候補 t を，特定の文献 d の索引語として付与するかどうかを判定する，いわば索引語の自動付与の手続きである。β は，IT の要素 t を対象にして行われるものであるから，β の実行のためには事前に α が実行されていなければならない。このアルゴリズム自動化のために算出される単語の重みは，特定の索引語候補 t の，特定の文献 d の索引語としてのふさわしさの程度であり，これを「Bタイプの重み」と呼ぶことにする。

アルゴリズム γ は，文献 d 中で用いられているある単語 w を，文献 d の索引語として抽出するかどうかを判定する，いわば索引語の自動抽出の手続きである。γ は，その手続きの中で単語 w が索引語候補としてふさわしいかどうかを同時に判定している。この意味で γ は α と β をその中に含めたアルゴリズムである。このアルゴリズム自動化のために算出される単語の重みは，特定の文献 d 中で用いられている特定の単語 w の，文献 d の索引語としてのふさわしさの程度であり，これを「Cタイプの重み」と呼ぶことにする。

アルゴリズム δ は，最もふつうに「索引語の重みづけ」と呼ばれているものであり，ある索引語 t に，文献 d の索引語としての適切さに応じて重みを与える手続きである。δ は，文献 d に付与された T の要素を対象にして行われるものであるから，δ の実行のためにはあらかじめ β か γ が実行されていなければならない。δ によって与えられた重みは，$\mu(d)$ の一部であり，通常 $\mu(d)$ と $\mu'(q)$ の合致性の度合を判断するときに利用される。したがって，δ は μ の構成要素であると同時に ν の構成要素である。このアルゴリズム自動化のために算出される単語の重みは，特定の索引語 t の，特定の文献 d の索引語としてのふさわしさの程度であり，これを「Dタイプの重み」と呼ぶことにする。

D. 重みづけの基本構造

前節で述べた4つのタイプの重みづけは，いままで混同されて論じられたことはあっても，区別を明確にした上で同時に論じられたことはない。実際，4つのアルゴリズムは別個のものなのであるから，4つの重みづけを一緒に扱うのは一見乱暴のように思われよう。それにもかかわらず本稿でこれらを同一のレベルで論じようとしているのは，すでに述べたように，これらの多様な重みづけを実現する数多くの手法に，通底する原理が存在しているからである。

さて，4つの重みづけを同じ土俵の上で論じるために，ここで単語の重みづけの基本的な構造を説明しておくことにする。いま，文献集合 Dy と，Dy に含まれる文献を構成するすべての単語の集合 Wy を考え，Dy の部分集合を Dx，Dx に含まれる文献を構成するすべての単語の集合を Wx と表すことにする。これらの記号を用いれば，単語の重みづけの基本的な構造は，「Wx⊂Wy なる関係をもつ単語集合に注目し，Wx の各要素に，それぞれが Wy 内で Wx の要素としてどの程度特徴的であるかを数値化して与えること」であると表現できる。

4つのタイプは，いずれもこの基本構造に即して解釈し直すことができる。まず，Aタイプの重みづけは，アルゴリズム α の自動化の場合，WD⊂NL なる関係をもつ2つの単語集合において，WD の各要素に数値を与える作業と解釈できる。アルゴリズム α' の自動化の場合は，WG⊂WD なる関係における作業である。

Bタイプの重みづけは，W の部分集合 $Wi=\{w|w\in W$ かつ $w\in IT'\}$ という単語集合を考えたとき，Wi⊂WD なる関係において，Wi の各要素に数値を与える作業と解釈できる。C タイプの重みづけは，W⊂WD なる関係において，W の各要素に数値を与える作業である。そしてDタイプの重みづけは，T⊂WD なる関係における作業と解釈できる。

このように，タイプにかかわらず単語の重みづけに共通していることは，それが「対象となる文献の集合を単語の集合とみなして，その特定の部分集合に注目し，この部分集合の各要素に数値を与える作業」であるという点である。いずれの重みづけも，包摂関係にある2つの単語集合を操作の対象としている点で同じである。ここで示した基本構造は，V章で再び論じることにする。

III. 単語の出現頻度情報

A. 文献空間と語彙空間

単語の重みづけは，各単語あるいは各索引語の重みを，単語の出現頻度情報に基づいたいくつかの数量を組み合わせて算出することによって実現されている。重みを求める数式の見かけ上の多様さにもかかわらず，そこで用いられている数量は基本的には共通であり，見かけ上の多様さはこれらの数量の組合せ方の多様さに過ぎない。本章では，これらの基本的な数量を整理して定義し，その表記法を定める。

基本的な数量を定義する前に，重みづけが行われる「文献空間」と「語彙空間」に関する記号を改めて定義する。まず，前章と同様に，文献検索システムが操作の対象とするすべての文献の集合を文献データベースと呼び，DB で表す。以降，単にデータベースといった場合にはこの DB を指すものとする。DB の要素である個々の文献は d_j で表す。すなわち，

$$DB=\{d_1, d_2, \cdots, d_j, \cdots\}$$

となる。データベースはしばしば下位の主題領域に区分されている。この下位の主題領域の文献の集合を「文献グループ」と呼び，DG_h で表す。このように，重みづけが行われる文献空間には，

$$d_j\in DG_h\subset DB$$

という関係が存在している。

語彙空間に関する記号としては，まず自然言語で用いられる語彙の集合を NL，データーベースの使用語彙を WD，文献グループ DG_h の使用語彙を WG_h，そして文献 d_j の使用語彙を W_j と表す。さらに，α の実行によって決定する，データベースの索引語候補の集合を IT，β または γ の実行によって決定する，データベースのすべての索引語の集合を IT′，同じく β または γ の実行によって決定する，文献 d_j に付与された索引語の集合を T_j と表す。また，W_j の部分集合 Wi_j を，

$$Wi_j=\{w|w\in W_j \text{ かつ } w\in IT'\}$$

と定義する。

以上の8つの記号は，添字を除けば前章と同じである。これらの単語集合のあいだには，次のような包含関係が成り立っている。

$$Wi_j\subset W_j\subset WG_h\subset WD\subset NL$$
$$T_j\subset IT'\subset IT\subset NL$$

WD の要素である個々の単語は w_j によって，IT の要素である個々の索引語候補は t_k によって表す。すなわち，

$$WD=\{w_1, w_2, \cdots, w_i, \cdots\}$$

$$\mathrm{IT}=\{t_1, t_2, \cdots, t_k, \cdots\}$$

である。

ところで，前章の4つのアルゴリズムは，すべての索引語候補が必ずデータベース中に出現することが仮定されている。重みづけの対象は，あくまでデータベース中のいずれかの文献に少なくとも1回は出現した単語である。同様に，アルゴリズム β, γ, δ では，ある文献に付与されるべき索引語は必ずその文献中に出現することも仮定されている。これらより，単語の重みづけにおいては，前述の関係に加えて次のような2つの包含関係が成立していることも明かである。

$$\mathrm{IT}\subset\mathrm{WD}$$

$$\mathrm{T}_i\subset\mathrm{W}_j$$

さて，前章では，単語の集合を論じるときに，同一の単語の異なった箇所の出現をそれぞれ別の要素として数えるか，同一の単語ならば何回出現していても1個と数えるかを問題とはしなかった。前者のように，同一の単語でも出現箇所が異なれば別の要素とみなして数える数え方における単語は，言語学では通常「トークン」と呼ばれている。これに対し，後者のように，同一の単語の出現は重複して数えず，いわば同一の単語のトークンをひとつにまとめて数える数え方における単語は「タイプ」と呼ばれている。また，タイプを単位として数えられた単語の数は「異なり語数」と呼ばれている。

単語の出現頻度を算定するときには，トークンとタイプのどちらを要素の単位とするかはきわめて重要である。そこで，本稿では以後，単に単語集合 X と表記した場合はトークンを単位とするものとし，タイプを単位とする場合は 〈X〉 と表記することで，この相違を明確にする。

前述の包含関係は，要素の単位をタイプにしても同様であるから，以下の関係が成り立つ。

$$\langle\mathrm{Wi}_j\rangle\subset\langle\mathrm{W}_j\rangle\subset\langle\mathrm{WG}_h\rangle\subset\langle\mathrm{WD}\rangle\subset\langle\mathrm{NL}\rangle$$

$$\langle\mathrm{T}_j\rangle\subset\langle\mathrm{IT}'\rangle\subset\langle\mathrm{IT}\rangle\subset\langle\mathrm{NL}\rangle$$

$$\langle\mathrm{IT}\rangle\subset\langle\mathrm{WD}\rangle$$

$$\langle\mathrm{T}_j\rangle\subset\langle\mathrm{W}_j\rangle$$

B. 基本的な数量の定義

はじめに，N, O_h, L, M, M' を，次のように定義する。

$$N=n\ (DB)$$

$$O_h=n\ (\mathrm{DG}_h)$$

$$L=n\ (\langle\mathrm{WD}\rangle)$$

$$M=n\ (\langle\mathrm{IT}\rangle)$$

$$M'=n\ (\langle\mathrm{IT}'\rangle)$$

ただし，$n(\mathrm{X})$ は集合 X の要素数である。N はデータベースの総文献数，O_h は文献グループ DG_h の総文献数，L はデータベース中の文献すべてで使用されている単語の異なり語数，M はデータベース中の索引語候補の異なり語数，そして M' はデータベース中の索引語の異なり語数をそれぞれ表している。

単語の重みを算出するにあたって，最も基本的な数量は「文献 d_j 内の単語 w_i の出現頻度」である。これを f_{ij} で表し，f_{ij} を累積することで，$sf_j, F_i, sF, F^{\sharp}_{ih}, sF^{\sharp}_h$ を次のように定義する。

$$sf_j=\sum_i f_{ij}$$

$$F_i=\sum_j f_{ij}$$

$$sF=\sum_j sf_j=\sum_i F_i=\sum_i\sum_j f_{ij}$$

$$F^{\sharp}_{ih}=\sum_j f_{ij}\quad (j\text{ は }d_j\in\mathrm{DG}_h\text{ を満たす})$$

$$sF^{\sharp}_h=\sum_j sF_j\quad (j\text{ は }d_j\in\mathrm{DG}_h\text{ を満たす})$$

また，$sf_j, sF, sF^{\sharp}_h$ は，次のようにも定義することができる。

$$sf_j=n\ (\mathrm{W}_j)$$

$$sF=n\ (\mathrm{WD})$$

$$sF^{\sharp}_h=n\ (\mathrm{WG}_h)$$

次に，g_{ij} を次のように定義する。

$$g_{ij}=\begin{cases}1 & (w_i\in\mathrm{W}_j)\\ 0 & (w_i\notin\mathrm{W}_j)\end{cases}$$

これは，単語 w_i の文献 d_j 内の出現を示す数である。この g_{ij} を累積することで，$sg_j, G_i, G^{\sharp}_{ih}$ を次のように定義する。

$$sg_j=\sum_i g_{ij}$$

$$G_i=\sum_j g_{ij}$$

$$G^{\sharp}_{ih}=\sum_j g_{ij}\quad (j\text{ は }d_j\in\mathrm{DG}_h\text{ を満たす})$$

sg_j は，次のようにも定義することができる。

$$sg_j=n(\langle\mathrm{W}_j\rangle)$$

索引語候補の出現頻度に関しては，まず「文献 d_j 内の索引語候補 t_k の出現頻度」を ϕ_{kj} で表し，ϕ_{kj} を累積することで $s\phi_j, \Phi_k$ を次のように定義する。

$$s\phi_j=\sum_k \phi_{kj}$$

$$\Phi_k=\sum_j f_{kj}$$

$s\phi_j$ は，次のようにも定義することができる。

$$s\phi_j=n(\mathrm{W}i_j)$$

ここで，$w_i=t_k$ の場合，$\phi_{kj}=f_{ij}$，$\Phi_k=F_i$ は成り立つが，$\mathrm{W}i_j\subset\mathrm{W}_j$ なので $s\phi_j\neq sf_j$ であることに注意してほしい。

さらに，索引語の出現頻度に関して q_{kj} を次のように定義する。

$$q_{kj}=\begin{cases}1 & (t_k\in \mathrm{T}_j)\\ 0 & (t_k\notin \mathrm{T}_j)\end{cases}$$

これは，文献 d_j に対する索引語 t_k の付与を示す数である。この q_{kj} を累積することで，sq_j，Q_k，sQ を次のように定義する。

$$sq_j=\sum_k q_{kj}$$

$$Q_k=\sum_j q_{kj}$$

$$sQ=\sum_j sq_j=\sum_k Q_k=\sum_k\sum_j q_{kj}$$

sq_j は，次のようにも定義することができる。

$$sq_j=n(\mathrm{T}_j)$$

ここで，$w_i=t_k$ の場合でも，一般には $q_{kj}\neq g_{ij}$，$sq_j\neq sg_j$，$Q_k\neq G_i$ であることに注意してほしい。

以上で定義した 22 個の数量が，単語の重みづけのための基本的な数量である。これらの数量は，すべて 0 以上の整数を値とする。また，それぞれの数量の具体的な意味は，第 1 表に示した通りである。

なお，これらの表記は添字を使っているものが多いが，いずれも添字を省略しても識別できるように定めてある。したがって，場合によっては，h, i, j, k などの添字は省略して表記し，数式表現を簡潔にする。

第 1 表　基本的な数量の表記法

記号	数量のもつ意味
N	文献データベースの総文献数
O_h	文献グループ DG_h の総文献数
L	文献データベース中の異なり語数
M	文献データベース中の索引語候補の異なり語数
M'	文献データベース中の索引語の異なり語数
f_{ij}	単語 w_i の文献 d_j 内の出現頻度
sf_j	文献 d_j の延べ語数
F_i	単語 w_i の文献データベース内の出現頻度
sF	文献データベースの延べ語数
$F^{\#}{}_{ih}$	単語 w_i の文献グループ DG_h 内の出現頻度
$sF^{\#}{}_h$	文献グループ DG_h の延べ語数
g_{ij}	単語 w_i の文献 d_j 内の出現を示す数（0 または 1）
sg_j	文献 d_j の異なり語数
G_i	文献データベース内で単語 w_i の出現している文献総数
$G^{\#}{}_{ih}$	文献グループ DG_h 内で単語 w_i の出現している文献総数
ϕ_{kj}	索引語候補 t_k の文献 d_j 内の出現頻度
$s\phi_j$	文献 d_j の延べ索引語候補数
Φ_k	索引語候補 t_k の文献データベース内の出現頻度
q_{kj}	文献 d_j に対する索引語 t_k の付与を示す数（0 または 1）
sq_j	文献 d_j に付与された索引語の総数，すなわち T_j の要素数
Q_k	文献データベース内で索引語 t_k が付与された文献総数
sQ	文献データベース内の全索引語の延べ付与数

C. 基本的な数量の相互関係

前節で定義した基本的な数量は，便宜的に次の 6 つにグループ分けすることができる。

（1） N, O_h

（2） L, M, M'

（3） f_{ij}, sf_j, F_i, sF, $F^{\#}{}_{ih}$, $sF^{\#}{}_h$

（4） g_{ij}, sg_j, G_i, $G^{\#}{}_{ih}$

（5） ϕ_{kj}, $s\phi_j$, Φ_k

（6） q_{ki}, sq_j, Q_k, sQ

(1) は文献集合の要素数に関する数量，(2) は基本的な単語集合の要素数に関する数量である。(3)，(4)，(5) はインデクシングの実行以前に確定できる数量であるのに対し，(6) はインデクシングが完了してはじめて確定できる数量である。また，(3) と (5) は単語のトークンを単位として数えた頻度であるのに対し，(2) と (4) は単語のタイプを単位として数えた頻度である。(6) は，索引語の付与を単位として数えた頻度である。

f_{ij}, sf_j, g_{ij}, sg_j, ϕ_{kj}, $s\phi_j$, q_{kj}, sq_j の 8 つの数量は，特定の文献内だけの頻度情報で確定できる数量であるのに対し，F_i, sF, G_i, Φ_k, Q_k, sQ の 6 つの数量は文献内だけの頻度情報では確定できない。そこで前者を「文献内情報」，後者を「文献間情報」と呼ぶことにする。また，これらのうち f_{ij}, g_{ij}, ϕ_{kj}, q_{kj} の 4 つの数量を「文献内出現頻度」，F_i, G_i, Φ_k, Q_k の 4 つの

数量を「データベース内出現頻度」と呼び，$F^{\sharp}_{ih}$ を「文献グループ内出現頻度」と呼ぶことにする。

N と O_h が文献を単位とした数量であるのは明白であるが，単語を単位とした数量の累積頻度である G_i，$G^{\sharp}_{ih}$，Q_k の3つの数量も，実は文献を単位としていると考えた方が理解しやすい。G_i はデータベース内で単語 w_i の出現している文献総数，$G^{\sharp}_{ih}$ は文献グループ DG_h 内で単語 w_i の出現している文献総数，そして Q_k はデータベース内で索引語 t_k が付与された文献総数である。これら5つの数量を「文献頻度」と呼ぶことにする。

ここにあげた22の数量以外にも，さらに sG，$sG^{\sharp}_h$，$s\Phi$ などいくつかの数量を，他の数量とパラレルに定義することは可能である。しかし，ここに定義した以上の数量は本稿で考察する単語の重みづけの手法においては使われていないので，煩雑さを避けるためにあえて定義しなかった。

IV. 重みの算出方法

本章では，いままで研究者が単語の重みづけのために提示してきた多様な算出式を，5つのグループに分けて説明する。説明にあたっては，算出式の基本的な成立ちと数値のふるまいに注目し，特に，異なった算出式のあいだにどのような関係があるのかを検討する。

説明の順番は，おおよそ単純な手法から複雑な手法である。発表された年代に関しては順番を考慮していないが，概して単純な手法ほど早くから提示されていたという傾向は見られる。

なお，基本的な数量から算出された単語の重みを，次のように表記することにする。

I_{ij}：文献 d_j 内の単語 w_i の重み
I_{ih}：文献グループ DG_h 内の単語 w_i の重み
I_{kj}：文献 d_j に対する索引語候補 t_k の重み

また，重みが文献や文献グループが特定されない場合には，I_i，I_k という表記を用いる。

A. 基本的な数量の単純な組み合せによる方法

まず，最も単純な重みづけの式として，次のものを想定することができる。

$$I_{ij}=g_{ij} \qquad (A.\ 1)$$

$$I_{kj}=q_{ki} \qquad (A.\ 2)$$

これらは，単語が当該の文献に出現しているか否か，索引語が当該の文献に与えられているか否か，つまり出現・非出現の情報のみを用いた重みづけである。

(A. 1) は，Sparck-Jones が，Dタイプの重みづけの最も単純なかたちとして提示している[2]。(A. 2) もDタイプであるが，特に誰かによって示されたものではない。たとえば，ブール演算を用いた単純な検索システムで，「特定の索引語 t を含む / 含まない」という命題の論理積結合によって文献と質問の両者を表現し，それらのマッチングを行うようなシステムでは，(A. 2) のような重みづけが行われていると考えられる。

(A. 1) は，Sparck-Jones によって次のように修正されている[2]。

$$I_{ij}=\frac{g_{ij}}{sg_j}=\frac{1}{sg_j} \qquad (A.\ 3)$$

g_{ij} が1に置き換えられているのは，そもそも文献中に出現していない単語は，初めから重みづけの対象とはならないからである。

単語の出現頻度に基づいた重みづけをはじめて提案したのは H.P. Luhn である[3][4][5]。Luhn は，単語のキーワードとしてのふさわしさを「解像力」(resolving power) と呼んでいるが，これはCタイプの重みと解釈できる。彼は単語を出現頻度の多い順に並べ，この順位にともなう解像力の増減を，モデル化したグラフで示している[3]。このグラフは全体として左右対称な山形をなしており，これに従えば，Luhn の主張は次のような関係式の提示であったと解釈できる。

$$f_{xj}\geq m \text{ のとき，} f_{xj}>f_{yj}\Rightarrow I_{xj}<I_{yj}$$
$$f_{xj}<m \text{ のとき，} f_{xj}<f_{yj}\Rightarrow I_{xj}<I_{yj} \qquad (A.\ 4)$$

ただし，m は単語の重みが最大になる f_{ij} の値，「$X\Rightarrow Y$」は命題 X が命題 Y の必要条件であることを表すものである。

Sparck-Jones は，Luhn の考え方をもとにした次のような式を，(A. 1) と並べて最も単純なかたちの評価式として提示している[2]。

$$I_{ij}=f_{ij} \qquad (A.\ 5)$$

彼女は，これを D タイプの重みづけとして示しているが，文献に出現しているすべての単語について適用でき，かつ単語間に差異を与えるので，Cタイプの重みづけともいえる。同一の式は，Sager と Lockemann[6]，Noreault ら[7] によっても示されている。また，彼女は (A.5) の代替案として，同時に次の式も提示している[2]。

$$I_{ij}=\log f_{ij} \qquad (A.\ 6)$$

ただし，以後特にことわらない限り log は e を底とするものとする。

Sparck-Jones が文献の異なり語数を考慮して (A. 1) を (A. 3) のかたちに修正したのと同じように，Sagar と Lockeman は文献の延べ語数を考慮して (A. 5) を次のかたちに修正している[6]。

$$I_{ij}=\frac{f_{ij}}{sf_j} \qquad (A.\ 7)$$

この式で求められる値は，文献の延べ語数に対する単語の出現頻度の割合，すなわち「単語の文献内相対出現頻度」である。ある数量の変動の影響を排除するためにその数量で割り算することを，その数量による「標準化」と表現すると，(A. 7) は (A. 5) を文献の延べ語数によって標準化した式であり，同じように (A.3) は (A. 1) を文献の異なり語数によって標準化した式と表現できる。

Sparck-Jones が対数を用いて (A. 5) を (A. 6) のように修正したのと同様に，Noreault らも対数を用いて (A. 7) を次のように修正している[7]。

$$I_{ij}=\frac{f_{ij}}{\log sf_j} \qquad (A.\ 8)$$

以上に示した8つの重み算出式は，いずれも文献内情報のみを組み合わせたものである。(A.1) から (A.3) までの算出式が単語のタイプを単位として数えた頻度に基づいているのに対し，(A.4) から (A.8) までの算出式は単語のトークンを単位として数えた頻度に基づいている。

さて，Sparck-Jones は，単語の文献内情報を用いた最も単純なかたちの算出式 (A. 1)，(A. 5) を，それぞれ (A. 3)，(A. 7) のかたちに修正したのと同時に，次のかたちへの修正を行なっている[2]。

$$I_{ij}=\frac{1}{G_i} \qquad (A.\ 9)$$

$$I_{ij}=\frac{f_{ij}}{F_i} \qquad (A.10)$$

これらは，(A. 1)，(A. 5) を単語のデータベース内出現頻度によって標準化したものと解釈できる。さらに彼女は，これらの式をそれぞれ (A. 3)，(A. 7) と掛け合わせて，次の算出式を提示している[2]。

$$I_{ij}=\frac{1}{sg_j\cdot G_i} \qquad (A.11)$$

$$I_{ij}=\frac{f_{ij}^2}{sf_j\cdot F_i} \qquad (A.12)$$

これら4つの式で求められる重みのタイプはDだが，いずれもCタイプの性格をもっている。

ところで，Sager と Lockemann は，Sparck-Jones が提示したものとして次の式を紹介している[6]。

$$I_{kj}=\frac{1}{Q_k} \qquad (A.13)$$

しかし，Sparck-Jones が示したのは実際には (A. 9) であるから，これは彼らが G_i と Q_k を混同したための誤解ではないかと思われる。彼らは，Sparck-Jones を引用しつつ，実は異なった評価式を示したのだと考えるべきであろう。同様の誤解は，Noreault らにも見られる[7]。彼らは次の式を Sparck-Jones の提示したものとして紹介している。

$$I_{kj}=\frac{1}{sg_j\cdot Q_k} \qquad (A.14)$$

これは，やはり G_i と Q_k を混同したため，(A.11) の式を誤解したものだろう。

Noreault らは，(A. 9) 以降の式ですでに用いられているいくつかの文献間情報を組み合わせた，次の4つの評価式も提示している[7]。

$$I_{kj}=\frac{1}{\log (sg_j\cdot Q_k)} \qquad (A.15)$$

$$I_{ij}=\frac{f_{ij}}{\log F_i} \qquad (A.16)$$

$$I_{ij}=\frac{f_{ij}}{sf_j\cdot F_i} \qquad (A.17)$$

$$I_{ij}=\frac{f_{ij}}{\log (sf_j\cdot F_i)} \qquad (A.18)$$

いずれも意図された重みのタイプは D だが，(A.13)～(A.15) は B タイプ，(A.16)～(A.18) は C タイプの重みとも考えられる。(A.17) は (A12) と同じ3つの数量を同じように組み合わせたものだが，分子が2乗されていない点のみ相違している。(A.15)，(A.16)，(A.18) は，それぞれ (A.14)，(A.10)，(A.17) の分母の値を対数値に修正したものである。

ところで，(A.16)～(A.18) の3つの式が単語のトークンを単位とした頻度情報のみに基づいているのに対し，(A.14) と (A.15) は単語のタイプを単位とした頻度情報と索引語の付与を単位とした頻度情報が取り混ぜて使われている。索引語の付与を単位とした頻度情報のみを用いて，

$$I_{kj}=\frac{1}{sq_j\cdot Q_k} \qquad (A.14)'$$

というかたちの算出式も構成できるはずであるが，この式を提示している研究者は見あたらない。これは，おそらくいずれの研究者も，基本的な数量のグループ (4) とグループ (6) を意識して区別していないためと思われる。

Sager と Lockemann は，(A. 5) と次の4つの式を，クィーンズ大学の QUIC/LAW システムと IBM の STAIRS システムで実験的に使用された，D タイプの重みの算出式として紹介している[6]。

$$I_{kj}=\phi_{kj}\frac{\Phi_k}{Q_k} \qquad (A.19)$$

$$I_{kj}=\phi_{kj}{}^2\frac{1}{Q_k} \qquad (A.20)$$

$$I_{kj}=\phi_{kj}{}^2\frac{\Phi_k}{Q_k{}^2} \qquad (A.21)$$

$$I_{kj}=\phi_{kj}\frac{Q_i}{\Phi_k-\phi_{kj}} \qquad (A.22)$$

Sager らの表記では f_{ij} と ϕ_{kj}, F_i と Φ_i が区別されていないが，ここでは明らかに索引語の文献内での出現頻度が問題となっているので，ϕ_{kj}, Φ_k を表記に用いた。(A.20) と (A.21) で数量がわざわざ2乗されているのは，算出される数量の次元を1次元に統一するためであろう。これらの式では，測度の次元は頻度情報と同次元に統一されている。

加藤緑らは，対象とする文献集合があらかじめいくつかの主題分野に分類されているようなシステムにおいて，キーワードを自動的に決定するための“数量的に表わされた語の重要度基準”[8]を提示しているが，これはタイプ A の重みづけに相当する作業である[8][9]。いま，rO_h を「文献グループ DG_h の総文献数の，データベースの総文献数に対する割合」，すなわち $rO_h=O_h/N$ と定義する。このとき，彼らが示したのは次のような3つの算出式である。

$$I_{ih}=F^{\#}{}_{ih} \qquad (A.23)$$

$$I_{ih}=\frac{F^{\#}{}_{ih}}{F_i} \qquad (A.24)$$

$$I_{ih}=\frac{F^{\#}{}_{ih}}{rO_h\cdot F_i} \qquad (A.25)$$

(A.24) は，(A.23) をデータベース内出現頻度で標準化した式である。(A.25) は，文献グループの大きさの及ぼす影響を排除するために，(A.24) の分母を rO_h で標準化した式である。

以上，(A.9)～(A.25) の 17 の算出式は，文献内情報に文献間情報を組み合わせたかたちをしている。これらは見かけはばらばらであるが，その構造はよく似通っている。どのように似通っているかは，次章で明らかにする。

B. 2つの相対出現頻度を用いた手法

はじめに，rf_{ij}, rF_i と $F^{\#}{}_{ih}$, rq_{kj} と rQ_k を次のように定義する。

$$rf_{ij}=\frac{f_{ij}}{sf_j}, \qquad rF_i=\frac{F_i}{sF}$$

$$rF^{\#}{}_{ih}=\frac{F^{\#}{}_{ih}}{sF^{\#}{}_h}$$

$$rq_{kj}=\frac{1}{sq_j}, \qquad rQ_k=\frac{Q_k}{sQ}$$

これらの数量のもつ意味は次の通りである。

- rf_{ij} ： 単語 w_i の文献 d_j 内の相対出現頻度
- rF_i ： 単語 w_i のデータベース内の相対出現頻度
- $rF^{\#}{}_{ih}$： 単語 w_i の文献グループ DG_h 内の相対出現頻度
- rq_{kj} ： 文献 d_j に付与された索引語集合 T_j の要素数の逆数
- rQ_k ： 索引語 t_k のデータベース内の相対付与頻度

本節では，rq_{kj} と rQ_k も含めて，この5つの数量を「相対出現頻度」と呼ぶことにする。また，rF_i と $rF^{\#}{}_{ih}$ を「文献間相対出現頻度」と呼ぶ。

H. P. Edmudson と R. E. Wyllys は，文献の主題を指示するものとしての単語の価値は，文献内相対出現頻度と文献間相対出現頻度の対比によって明らかになると主張し，Cタイプの重みを求める算出式として次の4つを提示している[10]。

$$I_{ij}=rf_{ij}-rF^{\#}{}_{ih} \qquad (B.\ 1)$$

$$I_{ij}=\frac{rf_{ij}}{rF^{\#}{}_{ih}} \qquad (B.\ 2)$$

$$I_{ij}=\frac{rf_{ij}}{rf_{ij}+rF^{\#}{}_{ih}} \qquad (B.\ 3)$$

$$I_{ij}=\log\frac{rf_{ij}}{rF^{\#}{}_{ih}} \qquad (B.\ 4)$$

いずれも文献内相対出現頻度と文献間相対出現頻度の対

比を数量化しようとしたものだが，（B．1）がその差に基づいているのに対して，（B．2)～(B．4）はその比に基づいている。(B．3)は，操作をしやすくするために，式の値が1より大きくならないように（B．2）を修正したものであろう。(B．4)は，（B．2）の対数値をとって修正したものである。

さらに Edmundson らは，単語の出現に関して「文献－文献データベース」の関係を「文献－文献グループ」の関係とまったく同じ次元で論じ，rF_i と $rF^{\sharp}_{ih}$ を記号の上では区別せず，上の4つの式に対応する次の4つの評価式を同一の式で表現している[10]。

$$I_{ij}=rf_{ij}-rF_i \qquad (B．5)$$

$$I_{ij}=\frac{rf_{ij}}{rF_i} \qquad (B．6)$$

$$I_{ij}=\frac{rf_{ij}}{rf_{ij}+rF_i} \qquad (B．7)$$

$$I_{ij}=\log\frac{rf_{ij}}{rF_i} \qquad (B．8)$$

(B.5)～(B.7）と同じかたちの算出式は，F.J. Damerau によっても提示されている[11]。

後藤，細野らは，漢字の出現頻度特性に基づいて，特定の主題分野に関連の深い漢字を，主題分野とは関連の薄い一般的な漢字から識別して抽出するための方法を提示している[12][13][14]。彼らがいわゆる単語ではなく漢字を対象にしたのは，当時日本語文の機械処理において，単語を切り出すことがかなり困難であったためで，実際，彼らは重要漢字の抽出を索引語候補の抽出と同等のプロセスとみなして分析を行なっている。そこで，ここでは漢字を単語の一種とみなし，彼らの重要漢字抽出の手法をAタイプの重みづけとして説明する。

後藤らは，特定の主題分野の重要漢字，すなわちある文献グループを特徴づける漢字を識別するための測度として，次の2つの算出式から求められる数値を提案している。

$$I_{ih}=rF^{\sharp}_{ih}-rF_i \qquad (B．9)$$

$$I_{ih}=\frac{rF^{\sharp}_{ih}-rF_i}{rF_i} \qquad (B.10)$$

彼らは $rF^{\sharp}_{ih}$ を「分野内出現率」，rF_i を「平均出現率」と呼んでいるが，これは明らかに2つの相対出現頻度を用いた評価式である。(B.10）は（B.9）を「平均出現率」で標準化したものである。彼らは（B.9）から求められる値を「重要度」，(B.10）から求められる値を「重要率」と呼んでいる。

一方，田中と岡坂[15]は，データベース中の専門用語を自動抽出するために，ブラウン大学英単語頻度辞書[16]を利用している。専門用語の抽出はアルゴリズム α に相当するので，ここでは彼らの提示した式を，Aタイプの重みの算出式として説明する。ブラウン大学英単語頻度辞書は，15の分野から抽出した異なり語数約5万，延べ語数約100万のサンプルデータを用いて，英単語の頻度情報を分析したものである。田中らは，この辞書における「各単語の出現頻度の延べ語数に対する割合」を求めて単語の評価に用いているが，これは特定の主題に限定しない場合の単語の相対出現頻度，あるいはすべての主題を含む自然言語の語彙NLにおける単語の仮想的な相対出現頻度と考えることができる。

単語 w_i の自然言語の語彙 NL における仮想の相対出現頻度を rF^*_i と表すと，彼らが示した算出式は次の4つである。

$$I_i=rF_i-rF^*_i \qquad (B.11)$$

$$I_i=2\cdot rF_i-rF^*_i \qquad (B.12)$$

$$I_i=\frac{rF_i-rF^*_i}{rF_i} \qquad (B\cdot13)$$

$$I_i=\frac{s(rF_i-rF^*_i)^2}{rF_i} \qquad (B.14)$$

ただし，

$$s=\begin{cases}-1 & (rF_i-rF^*_i<0)\\ 1 & (rF_i-rF^*_i>0)\end{cases}$$

である。

（B.12）は，（B.11）を修正し，2つの相対出現頻度のうちデータベース内出現頻度により大きな重みをつけたものである。(B.14）は，田中らによれば，式から求められる値が広く分布するように（B.13）を改良し，利用しやすくしたものである。

さて，(B.1)～(B.14）は，いずれも単語のトークンを単位とした相対出現頻度に基づいた評価式であるが，これらに対し，Sager と Lockemann は，単語のタイプを単位とした相対出現頻度に基づいた次の2つの算出式を，Dタイプの重みづけの式として提示している[6]。

$$I_{kj}=rq_{kj}-rQ_k \qquad (B.15)$$

$$I_{kj}=\frac{rq_{kj}}{rQ_k} \qquad (B.16)$$

これらは，それぞれ（B.5）と（B.6）に対応した式である。

ここまでに列挙した16の算出式は，いずれも2つの相対出現頻度のみを組み合わせた式である。これらに対し，以下に説明する4つの算出式は，2つの相対出現頻度を組み合わせて求めた値を，さらに分布のちらばりの特性値によって標準化するかたちをしている。分布のちらばりの特性値には，平均偏差，四分位範囲，ジニ係数などもあるが，ここで用いられているのは，最も一般的な分散と標準偏差である。

J. W. Carroll と R. Roeloffs は，文献の内容を最もよく特徴づける単語をキーワードと呼び，文献からキーワードを自動的に選択するためのCタイプの重みづけの式を5つ提示し，これらを比較している[17]。第1の式は，Sparck-Jones が示した（A.5）と同等であり，彼らはこれを‘word count’による方法と呼んでいる。第2，第3の式は，Edmundson と Wyllys が示した（B.5），（B.6）と同等であり，彼らはこれらをそれぞれ‘frequency difference’による方法，‘frequency ratio’による方法と呼んでいる。Carroll らがその次に示したのは，次のような式である。

$$I_{ij}=\frac{sF\cdot rf_{ij}-sF\cdot rF_i}{\sqrt{sF\cdot rF_i}} \qquad (B.17)$$

この式の分母は，$sF\cdot rf_{ij}$ の分布をポアソン分布であると仮定したときの，$sF\cdot rf_{ij}$ の分布の標準偏差の値である。

Carroll と Roeloffs は，sF_i はデータベースに固有の定数であるから，実際に計算するには（B.17）を修正した次のかたちの算出式でもよいと主張している。

$$I_{ij}=\frac{rf_{ij}-rF_i}{\sqrt{rF_i}} \qquad (B.18)$$

これらが彼らの示した第4の式であり，彼らはこれを‘Poisson standard deviate’による方法と呼んでいる。

（B.17）では，ポアソン分布の仮定から標準偏差を求めているが，Carroll らが示した第5の式では，rf_{ij} の分布の標準偏差が，標準偏差の基本的な定義から求められている。まず，rf_{ij} の分布の不偏分散 $r\sigma_i^2$ を次の式から定義する。

$$r\sigma_i^2=\frac{1}{N-1}\sum_j (rf_{ij}-r\bar{f}_i)^2$$

ただし，$r\bar{f}_i$ は w_i の文献内相対出現頻度 rf_{ij} の平均値であり，次の式から求められる。

$$r\bar{f}_i=\frac{1}{N}\sum_j rf_{ij}$$

Carroll らが示した第5の式は，次の通りである。

$$I_{ij}=\frac{rf_{ij}-rF_i}{r\sigma_i} \qquad (B.19)$$

彼らはこの式を‘standard deviate’による方法と呼んでいる。

Carroll と Roeloffs の評価式は，いずれも単語のトークンを単位とした頻度に基づいているのに対し，Sager と Lockemann は，単語のタイプを単位とした頻度に基づいて（B.18）を修正して次の式を提示している[6]。

$$I_{kj}=\frac{rq_{kj}-rQ_k}{\sqrt{rQ_k}} \qquad (B.20)$$

重みづけのタイプは D である。

これら4つの算出式は，いずれも2つの相対出現頻度の差をちらばりの特性値で標準化している点で，同じ構造をもっている。（B.17）～（B.19）は（B.5）を修正したもの，（B.20）は（B.15）を修正したものと説明できる。

C. ちらばりの特性値を用いた方法

前節の最後に説明した方法は，分布のちらばりの特性値を重みの標準化に用いたものであったが，本節では，ちらばりの特性値，あるいはそれに相当する値そのものを重みとして使用する方法を説明する。

S. F. Dennis は，自動インデクシングのシステムにおいて，文献中の「内容語」(content word) を「非内容語」(noncontend word) から識別するための手法として，すなわちAタイプの重みの算出式として，次のようなかたちの式を提示している[18]。

$$I_i=\frac{F_i}{r\bar{f}_i^2/r\sigma_i^2} \qquad (C.1)$$

ただし，$r\bar{f}_i$，$r\sigma_i^2$ の定義は，（B.19）と同様である。Dennis の説明によれば，この式より算出される値は，“それぞれの文献に対する単語の出現のふぞろいさ”[18]の程度を反映するものである。

一方，Stone と Rubinoff は，文献中の「専門語」(speciality word) を「非専門語」(non-speciality word) から識別する手法として，すなわち同じくAタイプの重みづけの式として，次の式を提示している[19]。

$$I_i=\frac{\sigma_i^2}{F_i} \qquad (C.2)$$

ただし，σ_i^2 は w_i の文献内出現頻度 f_{ij} の分布の分散を

表している。Stone らは，この分散の値を求める式を示していないが，不偏分散の定義に従えば次のようになろう。

$$\sigma_i^2=\frac{1}{N-1}\sum_j\left(f_{ij}-\frac{F_i}{N}\right)^2$$

Stone らは，(C.2) から求められる数量を，“分布が，ちらばりに関してポアソン分布から離れている程度を測る測度”[19] であると説明している。なぜなら，f_{ij} の分布がポアソン分布であるならば，分散と平均が等しいことより，分散は F_i に比例するからである。彼らはこの式を，Dennis の式 (C.1) の代替案として示している。

竹内，岩坪，西野は，文献の自動分類のための第1段階に“すでに正しく分類されている文献データを使ってキーワードを抽出”[20]する作業を位置づけ，キーワードを抽出するための指標として「単語の局在性を示す指標」を提案している。これはAタイプの重みに相当するものであり，その算出式は次の通りである。

$$I_i=\frac{1}{\Omega-1}\sum_h(1-rG^{\#}{}_{ih})^2 \qquad (C.3)$$

ただし，Ω は DB に含まれる DG の数，すなわち文献データベース中の総文献グループ数である。また，$rG^{\#}{}_{ih}$ は，文献グループ DG_h 内で単語 w_i の出現している文献総数を，最大値が1になるように $G^{\#}{}_{ih}$ の最大値で標準化したものであり，次の式で求められる。

$$rG^{\#}{}_{ih}=\frac{G^{\#}{}_{ih}}{\max_h G^{\#}{}_{ih}}$$

(C.3) は，分散そのものではないが，分散と同じ考え方から導かれた式である。分散が平均からの偏差の平方の平均であるのに対し，この式では1からの偏差の平方の平均を求めている。したがって，この式から求められる数量は，$rG^{\#}{}_{ih}$ のちらばりの程度を測る数量であるとみなすことができる。

次に，長尾，水谷らが提示した，カイ2乗を用いたAタイプの重みづけの手法を説明する[21)22)]。長尾らは，“文献内容をよく表し，検索する際に「見出し語」として使用できるような特徴のある単語”[21]のことを「重要語」と呼び，この重要語をその他の「一般語」から区別して抽出するための指標に，カイ2乗の値を用いている。カイ2乗は，本来は，期待値からの観測値の乖離度を測る値であるが，分布の平均値を期待値とみなせば，分布の平均値からのちらばりの程度を測るものともみなすことができる。

長尾，落合，水谷が示した「文献から重要語を抽出するためのカイ2乗」は，次の式で求められる[21]。

$$I_i=\sum_j\frac{(rf_{ij}-rF_i)^2}{rF_i} \qquad (C.4)$$

これは，それぞれの文献における w_i の文献内相対出現頻度が，データベース内相対出現頻度からどの程度離れているかを示すカイ2乗である。

一方，長尾，水谷，池田が示した「文献グループから重要語を抽出するためのカイ2乗」は，すでにデータベースがいくつかの分野，すなわち文献グループに分類されていることを前提とするもので，次の2つ式で求められる[22]。

$$I_i=\frac{\sum_h(F^{\#}{}_{ih}-rF_i\cdot sF^{\#}{}_h)^2}{rF_i\cdot sF^{\#}{}_h} \qquad (C.5)$$

$$I_i=\frac{\sum_h(rF^{\#}{}_{ih}-rF_i)^2}{rF_i} \qquad (C.6)$$

前者は，それぞれの文献グループにおける w_i の文献グループ内出現頻度が，データベース内相対出現頻度からどの程度離れているかを示すカイ2乗である。これに対し，後者は前者から w_i の文献グループ内出現頻度の大きさの影響を除いたものである。

後藤，細野らは，前節で紹介したように，特定の主題分野の重要漢字を識別して抽出するために，2つの相対出現頻度を用いた重みの算出式を提示しているが，さらに，長尾，水谷らの (C.5)，(C.6) とまったく同等の次のような算出式もあわせて提示している[13]。

$$I_i=\sum_h\frac{(F^{\#}{}_{ih}-sF^{\#}{}_h\cdot rF_i)^2}{sF^{\#}{}_h\cdot rF_i} \qquad (C.5)'$$

$$I_i=\sum_h\frac{(rF^{\#}{}_{ih}-rF_i)^2}{rF_i} \qquad (C.6)'$$

後藤らは，これら2つの値を「出現偏差度」と呼んでいる。ただし，彼らの論文中には，これらがカイ2乗と同等であるという説明はない。

D. 2-ポアソン・モデルに基づく方法

S.P. Harter は，文献中の単語の分布を，「2-ポアソン・モデル」と名付けられた独自の分布モデルによって説明することを試み，これに基づいて単語の重みづけを行う方法を提案している[23)24)]。本節では，このモデルと重みづけの方法を説明する。

Harter のモデルは，第1に，特定の単語に関してデータベース中の文献が，(1) その単語が表現している内

容を特に主題として扱っている文献の集合と，(2) 特に主題扱いしていない文献の集合の2つに分類できることを仮定している。彼は前者をクラスⅠ，後者をクラスⅡと呼んでいる。そして第2に，このどちらの集合においても，その単語の文献内出現頻度はある平均値をもったポアソン分布に従うことを仮定している。これらの仮定より，Harter は，特定の単語の 文献内出現頻度の分布を，2つのポアソン分布を組み合わせた次のような式によってモデル化している。

$$Pr(f_i=x)$$
$$=\pi\frac{e^{-m_{1i}}\cdot m_{1i}{}^{x}}{x!}+(1-\pi)\frac{e^{-m_{2i}}\cdot m_{2i}{}^{x}}{x!}$$

ただし，$Pr(f_i=x)$ は，単語 w_i の文献内出現頻度が x である文献の 文献総数に対する割合，すなわち単語 w_i の文献内出現頻度が x である確率を表している。さらに m_{1i} と m_{2i} は，それぞれ単語 w_i のクラスⅠ，クラスⅡにおける文献内出現頻度の平均値であり，π は，クラスⅠに属する文献の文献総数に対する割合，すなわち文献がクラスⅠに属する確率を表している。また，$m_{1i}\geqq m_{2i}$ である。Harter の 2-ポアソン・モデルとは，この式によって表現される分布モデルである。

このモデルに基づいて，「キーワード」(keyword) を「非キーワード」(non-speciality word) から識別する手法，すなわちAタイプの重みづけの方法として Harter が提案したのは，次のようなかたちの式である[23]。

$$I_i=\frac{m_{1i}-m_{2i}}{\sqrt{m_{1i}+m_{2i}}} \qquad (D.\ 1)$$

Harter の説明によれば，この数値は，クラスⅠとクラスⅡの文献内出現頻度の分布の平均の差を，その分散の和の平方根で除したものであり，2つのクラスのへだたりの大きさを測る測度である。

実際に (D.1) の式から重みを求めるには，まず m_{1i} と m_{2i} の値を算出しなければならない。Harter は，これらの値を観測された単語の出現頻度から求める方法を，2-ポアソン・モデルの積率母関数から導いて説明している。

E. Shannon の情報量の概念を用いた方法

「情報量」という概念は，C.E. Shannon と N. Wiener によって確立された「情報理論」において確立されたものである[25]。Shannon らの情報理論における「情報量」は，ごく簡潔にいえば，偶然性を伴う事象のあいまいさの程度を，事象の生起確率に基づいて数量化したものである。文献中に特定の単語が出現する事象を確率事象と考えれば，単語の重みづけに情報量の概念を応用することが可能である。本節では，情報量の考え方を用いた重みづけの方法を説明する。

S.E. Robertson は，情報検索システムにおいてすでに付与された索引語の重みづけ，すなわちDタイプの重みを求める次のような算出式を提案している[26]。

$$I_{ik}=-\log_2\frac{Q_k}{N}$$
$$=\log_2 N-\log_2 Q_k \qquad (E.\ 1)$$

Q_k/N は「データベースの総文献数に対する索引語 t_k が付与された文献総数の割合」であるが，Robertson はこれを「データベースからランダムに文献をひとつ取り出したとき，その文献が索引語 t_k を付与されている確率」とみなしている。そして，この確率に基づき，索引語の重みを「データベースからランダムに文献をひとつ取り出したとき，その文献が索引語 t_k を付与されていることを知ったときに与えられる情報量」として算出したのがこの式である。これは，いいかえれば，文献 d_j に索引語 t_k が付与されるという事象の自己情報量である。

Robertson は，(E.1) は Sparck-Jones が提示したものを修正した式であると説明し，Sparck-Jones が作成したオリジナルの式として次のものを示している[26]。

$$I_{kj}=\log_2 N-\log_2 Q_k+1 \qquad (E.\ 2)$$

Robertson は，$+1$ は式の値が 0 にならないようにするための値であると説明している。しかし，Sparck-Jones が実際に提示した式は，これとは若干異なった次のような式である[27]。

$$I_{kj}=[\log_2 N]-[\log_2 Q_k]+1 \qquad (E.\ 3)$$

ただし，$[X]$ は X の小数点以下を切り上げて整数化した値を表している。Sparck-Jones が小数点以下を切り上げた値を用いたのは，単に計算の便宜を図るためであったと思われる。

G. Salon と M.J. McGill は，索引語の自動抽出と自動重みづけの手法のひとつとして上述の Sparck-Jones の論文を引用している[28]。しかし，彼らが示したのも Sparck-Jones のオリジナルの式ではなく，やはり (E.2) のかたちの式である。さらに彼らは，(E.2) を次のように修正した算出式を提案している。

$$I_{kj}=\phi_{kj}\cdot(\log_2 N-\log_2 Q_k+1) \qquad (E.\ 4)$$

$t_k=w_i$ のとき，ϕ_{kj} は f_{ij} と同じであるから，この式は (E.2) を (A.5) と組み合わせた式と考えることができる。彼らは (E.2), (E.4) による方法を，'inverse document frequency' による方法と呼んでいる。

Noreault らは，すでに紹介したようにDタイプの重みの算出式をいくつか示しているが，その中のひとつに次のような式が含まれている[7]。

$$I_{ij}=f_{ij}\cdot\log\frac{sF}{F_i} \qquad (E.\ 5)$$

この式がどのようにして導かれたのかについては何の記述もないが，ここにも情報量の考え方が含まれていると解釈することができる。(E.5) は次のように変形できる。

$$I_{ij}=f_{ij}\cdot-\log\frac{F_i}{sF} \qquad (E.\ 5)'$$

F_i/sF は「データベース内の延べ語数に対する w_i の出現頻度の割合」であるが，これは「データベースからランダムに単語をひとつ取り出したとき，その単語が w_i である確率」とみなすことができる。したがって (E.5) は，単語の文献内出現頻度に，データベース内で単語 w_i が出現するという事象の自己情報量を乗じたものと説明することができる。

G. Salton の 1975 年の著作には，索引語決定のための頻度情報に基づいた手法がいくつか紹介されているが，その中のひとつは情報量の考え方をもとにしたものであり，彼はこの手法を「シグナル-ノイズ算出法 (signal-noise calculation)」と呼んでいる[29]。Salton によれば，これは S.F. Dennis がはじめに提案した手法ということだが，Dennis の論文にこの手法の説明は見あたらない。したがってこの手法に関する以下の説明は，Salton の 75 年の著作[29]，および Salton と McGill の著作[28]の記述によるものである。

Salon は，まず特定のデータベースにおける単語 w_i の「ノイズ」NZ_i を次のように定義している。

$$NZ_i=\sum_j\frac{f_{ij}}{F_i}\log\frac{F_i}{f_{ij}}$$

Salton 自身の説明によれば，ノイズは，データベース内での単語の出現のかたよりの大きさに反して変化する数量である。このノイズより，さらに単語 w_i の「シグナル」を次のように定義している。

$$SG_i=\log F_i-NZ_i$$

Salton は，これらの数量が情報理論に基づいて導かれたものだと説明しているが，具体的にどのような確率事象系を想定して導いたのかについては説明を与えていない。そこで，次章E節では，これらの数量がどのような事象系におけるどのような情報量に相当するのかを推定し，検討を加えることにする。

Salton は，これらの2つの数量を用いて，Aタイプの重みづけする式として，75年の著作では次の2つの式を提示している[29]。

$$I_i=\frac{SG_i}{NZ_i} \qquad (E.\ 6)$$

$$I_i=\frac{SG_i}{NZ_i}\cdot SG_i \qquad (E.\ 7)$$

また，Satlon と McGillの 著作では，同じく シグナルを用いた次の算出式を提示している[28]。

$$I_{ij}=f_{ij}\cdot SG_i \qquad (E.\ 8)$$

(E.6) と (E.7) は，どちらもノイズに対するシグナルの比を求めたものである。これに対し，(E.8) は，シグナルの考え方に (A.5) を組み合わせたものと解釈できる。

V. 単語の偏在性に基づく算出方法の解釈

A. 偏在性の原理

前章で説明した 60 余りの重み算出式 (第2表を参照) について，そこで用いられている基本的な数量が重みの大小にどのように影響しているかをそれぞれ調べてみると，いくつかの算出式ごとに共通の数量関係を見いだすことができる。本章では，算出式のそれぞれのグループごとにこのような共通の数量関係を抽出し，さらにそれらすべてを包括するロジックとして「偏在性の原理」と呼びうる考え方が存在することを明らかにする。

Ⅱ章D節では，さまざまな目的のもとに行われる単語の重みづけが，実は「特定の単語集合 Wy の特定の部分集合 Wx の各要素に対して数値を与える」という同一の構造をそなえていることを説明した。前章で示した多様な方法も，例外なくこの構造をそなえている。それぞれの方法がどのような単語集合のどのような部分集合を対象とした重みづけを行なっているかは，第3表の「Wx - Wy」関係の欄に示した通りである。

この Wx - Wy の記号を用いれば，重みづけにおける偏在性の原理とは，次のように表現される考え方である。

出現頻度情報に基づく単語重みづけの原理

第2表　単語の重みの算出式一覧

A. 基本的な数量の単純な組み合せによる方法

$$I_{ij}=g_{ij} \quad (A.\ 1)$$

$$I_{kj}=q_{kj} \quad (A.\ 2)$$

$$I_{ij}=\frac{g_{ij}}{sg_j}=\frac{1}{sg_j} \quad (A.\ 3)$$

$$I_{ij}=f_{ij} \quad (A.\ 5)$$

$$I_{ij}=\log f_{ij} \quad (A.\ 6)$$

$$I_{ij}=\frac{f_{ij}}{sf_i} \quad (A.\ 7)$$

$$I_{ij}=\frac{f_{ij}}{\log sf_j} \quad (A.\ 8)$$

$$I_{ij}=\frac{1}{G_i} \quad (A.\ 9)$$

$$I_{ij}=\frac{f_{ij}}{F_i} \quad (A.10)$$

$$I_{ij}=\frac{1}{sg_j\cdot G_i} \quad (A.11)$$

$$I_{ij}=\frac{f_{ij}^2}{sf_j\cdot F_i} \quad (A.12)$$

$$I_{kj}=\frac{1}{Q_k} \quad (A.13)$$

$$I_{kj}=\frac{1}{sg_j\cdot Q_k} \quad (A.14)$$

$$I_{kj}=\frac{1}{\log(sg_j\cdot Q_k)} \quad (A.15)$$

$$I_{ij}=\frac{f_{ij}}{\log F_i} \quad (A.16)$$

$$I_{ij}=\frac{f_{ij}}{sf_j\cdot F_i} \quad (A.17)$$

$$I_{ij}=\frac{f_{ij}}{\log(sf_j\cdot F_i)} \quad (A.18)$$

$$I_{kj}=\frac{1}{sq_j\cdot Q_k} \quad (A.14)'$$

$$I_{kj}=\phi_{kj}\frac{\Phi_k}{Q_k} \quad (A.19)$$

$$I_{kj}=\phi_{kj}^2\frac{1}{Q_k} \quad (A.20)$$

$$I_{kj}=\phi_{kj}^2\frac{\Phi_k}{Q_k^2} \quad (A.21)$$

$$I_{kj}=\phi_{kj}\frac{Q_i}{\Phi_k-\phi_{kj}} \quad (A.22)$$

$$I_{ih}=F_{ih}{}^{\#} \quad (A.23)$$

$$I_{ih}=\frac{F_{ih}{}^{\#}}{F_i} \quad (A.24)$$

$$I_{ih}=\frac{F_{ih}{}^{\#}}{rO_h\cdot F_i} \quad (A.25)$$

B. 2つの相対出現頻度を用いた方法

$$I_{ij}=rf_{ij}-rF^{\#}{}_{ih} \quad (B.\ 1)$$

$$I_{ij}=\frac{rf_{ij}}{rF^{\#}{}_{ih}} \quad (B.\ 2)$$

$$I_{ij}=\frac{rf_{ij}}{rf_{ij}+rF^{\#}{}_{ih}} \quad (B.\ 3)$$

$$I_{ij}=\log\frac{rf_{ij}}{rF^{\#}{}_{ih}} \quad (B.\ 4)$$

$$I_{ij}=rf_{ij}-rF_i \quad (B.\ 5)$$

$$I_{ij}=\frac{rf_{ij}}{rF_i} \quad (B.\ 6)$$

$$I_{ij}=\frac{rf_{ij}}{rf_{ij}+rF_i} \quad (B.\ 7)$$

$$I_{ij}=\log\frac{rf_{ij}}{rF_i} \quad (B.\ 8)$$

$$I_{ih}=rF^{\#}{}_{ih}-rF_i \quad (B.\ 9)$$

$$I_{ih}=\frac{rF^{\#}{}_{ih}-rF_i}{rF_i} \quad (B.10)$$

$$I_i=rF_i-rF^*{}_i \quad (B.11)$$

$$I_i=2\cdot rF_i-rF^*{}_i \quad (B.12)$$

$$I_i=\frac{rF_i-rF^*{}_i}{rF_i} \quad (B.13)$$

$$I_i=\frac{s(rF_i-rF^*{}_i)^2}{rF_i} \quad (B.14)$$

$$I_{kj}=rq_{kj}-rQ_k \quad (B.15)$$

$$I_{kj}=\frac{rq_{kj}}{rQ_k} \quad (B.16)$$

$$I_{ij}=\frac{sF_i\cdot rF_{ij}-sF_i\cdot rF_i}{\sqrt{sF_i\cdot rF_i}} \quad (B.17)$$

$$I_{ij}=\frac{rf_{ij}-rF_i}{\sqrt{rF_i}} \quad (B.18)$$

$$I_{ij}=\frac{rf_{ij}-rF_i}{r\sigma_i} \quad (B.19)$$

$$I_{kj}=\frac{rq_{kj}-rQ_k}{\sqrt{rQ_k}} \quad (B.20)$$

第2表 つづき

C. ちらばりの特性値を用いた方法

$$I_i=\frac{F_i}{\bar{rf_i}^2/r\sigma_i^2} \quad (C.\ 1)$$

$$I_i=\frac{\sigma_i^2}{F_i} \quad (C.\ 2)$$

$$I_i=\frac{1}{\Omega-1}\sum_h(1-rG_i^{\#}{}_h)^2 \quad (C.\ 3)$$

$$I_i=\sum_j\frac{(rf_{ij}-rF_i)^2}{rF_i} \quad (C.\ 4)$$

$$I_i=\frac{\sum_h(F^{\#}{}_{ih}-rF\cdot sF^{\#}{}_h)^2}{rF_i\cdot sF^{\#}{}_h} \quad (C.\ 5)$$

$$I_i=\frac{\sum_h(rF^{\#}{}_{ih}-rF_i)^2}{rF_i} \quad (C.\ 6)$$

$$I_i=\sum_h\frac{(F^{\#}{}_{ih}-sF^{\#}{}_h\cdot rF_i)^2}{sF^{\#}{}_h\cdot rF_i} \quad (C.\ 5)'$$

$$I_i=\sum_h\frac{(rF^{\#}{}_{ih}-rF_i)^2}{rF_i} \quad (C.\ 6)'$$

D. 2-ポアソン分布モデルに基づく方法

$$I_i=\frac{m_{1i}-m_{2i}}{\sqrt{m_{1i}+m_2}} \quad (D.\ 1)$$

E. Shannon の情報量の概念を用いた方法

$$I_{ik}=-\log_2\frac{Q_k}{N}$$

$$=\log_2 N-\log_2 Q_k \quad (E.\ 1)$$

$$I_{kj}=\log_2 N-\log_2 Q_k+1 \quad (E.\ 2)$$

$$I_{kj}=[\log_2 N]-[\log_2 Q_k]+1 \quad (E.\ 3)$$

$$I_{kj}=\phi_{kj}\cdot(\log_2 N-\log_2 Q_k+1) \quad (E.\ 4)$$

$$I_{ij}=f_{ij}\cdot\log\frac{sF}{F_i} \quad (E.\ 5)$$

$$I_{ij}=f_{ij}\cdot-\log\frac{F_i}{sF} \quad (E.\ 5)'$$

$$I_i=\frac{SG_i}{NZ_i} \quad (E.\ 6)$$

$$I_i=\frac{SG_i}{NZ_i}\cdot SG_i \quad (E.\ 7)$$

$$I_{ij}=f_{ij}\cdot SG_i \quad (E.\ 8)$$

「特定の単語 x ($x\in$ Wx) に対して与える重みは，x が Wy において Wx へ偏って存在している程度に応じた値である」

偏在性の原理に従えば，たとえば W-WD 関係に注目した場合，特定の文献中に出現している特定の単語に与えられる重み，「その単語が文献データベース中でその文献に偏って出現している程度を数量化した値」ということになろう。次節以降では，それぞれの重みの算出式の中に表われているどのような数量関係に，この原理を見いだすことができるかを説明する。

ところで，Ⅲ章B節では，集合 X の要素数を n(X) と表記した。本章ではこれに加えて，単語集合 X 中の特定の要素 x の数を x(X) と表記することにする。この記号を用いれば，Ⅲ章で定義した基本的な数量のいくつかは，次のよにに表現することもできる。

$f=w$(W)　　$F=\omega$(WD)　　$F^{\#}=w$(WG)

$g=w(\langle$W$\rangle)$　　$G=\omega(\langle$WD$\rangle)$　　$G^{\#}=w(\langle$WG$\rangle)$

$\phi=t$(Wi)　　$\Phi=t$(IT)

$q=t$(T)　　$Q=t$(IT′)

ただし，添字は省略している。

B. 3つの原始的な数量関係

まず，基本的な数量の単純な組み合せによる重みの算出式 (A. 1)～(A. 25) に共通に見いだされる数量関係は，次3のつである。

① 単語 x の重みは，x(Wx) の増加に伴って単調に増加

② 単語 x の重みは，n(Wx) の増加に伴って単調に減少

③ 単語 x の重みは，x(Wy) の増加に伴って単調に減少

これら3つの値は，いずれも単語 x の Wx への偏りの程度に応じて変化する値である。x(Wy)$=x'$(Wy) のとき x(Wx)$>x'$ (Wx) ならば，x の方が x' よりも Wx

に偏って多く存在していることは明かである。また，$x(Wx)=x(Wx')$ のとき $n(Wx)>n(Wx')$ ならば，x は相対的に Wx よりも Wx′ に偏って存在しているといえる。さらに，$x(Wx)=x'(Wx)$ のとき $x(Wy)>x'(Wy)$ ならば，x' は x よりも相対的に Wx に偏って存在しているといえる。

① から ③ の数量関係を含んだ重みづけの一般式は，たとえば次のようになろう。

$$I=\frac{x(Wx)}{n(Wx)\cdot x(Wy)}$$

この式では，I の値は $x(Wx)$ の値に比例し，$n(Wx)$ および $x(Wy)$ の値に反比例している。前章の算出式の中では，(A.17) と (A.14)′ がこの一般式とまったく同等である。この式を W-WD 関係に注目して組み立てれば (A.17) になり，T-IT′ 関係に注目すれば (A.14) になる。

その他の算出式も，① から ③ の数量関係のいくつか，あるいはすべてを含んでいる。いくつかの算出式では，式を組み立てる値に対数値や平方値が用いられているが，基本的な数量関係は変化していない。対数値が用いられるのは，その値の重みの変化に対する影響が低く見積られているためであり，反対に平方値が用いられているのはその値の影響が高く見積られているためであると解釈できる。

C. 2つの相対出現頻度の比較

2つの相対出現頻度を用いた重みの算出式 (B.1)～(B.20) は，特定の単語の Wx における相対出現頻度を Wy における相対出現頻度と比較し，前者が後者よりもどの程度大きいかを数量化しているという点で，すべて同一の構造をそなえている。たとえば(B.1)～(B.4) では W-WG 関係において，(B.5)～(B.8) では W-WD 関係において，(B.9)～(B.19) では WG-WD 関係において，それぞれ2つの相対出現頻度の比較が行われている。

単語 x の Wx における相対出現頻度が Wy における相対出現頻度よりも大きいということは，x が Wy 全体に出現している割合に比べて，その部分集合 Wx に出現している割合が大きいということである。これは「x の存在が Wx に偏っている」というのと同じ意味に解釈できる。すなわち，2つの相対出現頻度の比較によって求められる値は，単語の偏在性の測度とみなすことのできる値である。

(B.1)～(B.20) の算出式に見いだされる共通の数量関係は次の2つである。

④ 単語 x の重みは，$\frac{x(Wx)}{n(Wx)}$ の増加に伴って単調に増加

⑤ 単語 x の重みは，$\frac{x(Wy)}{n(Wy)}$ の増加に伴って単調に減少

これらの数量関係を含んだ重みづけの一般式は，たとえば次のようになろう。

$$I=\frac{x(Wx)\cdot n(Wy)}{n(Wx)\cdot x(Wy)}$$

$$I=\frac{x(Wx)}{n(Wx)}-\frac{x(Wy)}{n(Wy)}$$

前者は，2つの相対出現頻度の比をとることで比較を行なった式であり，(B.2)，(B.6)，(B.16) がこれと同等である。また，(B.3)，(B.4)，(B.7)，(B.8) は，これを修正した式と解釈できる。

後者は，2つの相対出現頻度の差をとることで比較を行なった式であり，(B.1)，(B.5)，(B.9)，(B.15) がこれと同等である。また，(B.10)，(B.12)～(B.14) は，これを修正した式と解釈できる。(B.17)～(B.20) は，いずれもこの式から求められる値を，ちらばりの特性値で標準化したものである。

④ と ⑤ に示した数量関係は，そのかたちを見れば明らかなように ①～③ の数量関係の十分条件である。したがって，上記の2つの式も ①～③ の関係を満たしている。この意味で，2つの相対出現頻度を用いた重みづけは，基本的な数量の単純な組み合せによる重みづけの考え方をその中に含んだ方法であるということもできよう。

D. ちらばりの特性値または偏りの測度

算出式 (C.1)～(C.6) は，分布のちらばりの特性値の大きさに応じて単語に重みを与えているという点で同じ構造をそなえている。これらの算出式に共通の数量関係は，次の2つである。

⑥ 単語 x の重みは，$x(Wx_j)$ の分布のちらばりの程度の増加に伴って単調に増加

⑦ 単語 x の重みは，$\frac{x(Wx_j)}{n(Wx_j)}$ の分布のちらばりの程度に伴って単調に増加

ここに示されている分布のちらばりの程度は，次に説明するように，どちらも単語 x の Wy 中での偏りの程度に応じて変化する値である。

いま，単語集合 Wy が，複数の部分集合

$$Wx_1, Wx_2, Wx_3, \cdots, Wx_j, \cdots$$

に分けられているとする。仮に単語 x がそれぞれの Wx_j に均一に出現しているとするならば，x は Wx の大きさに応じて Wx 中に含まれるはずであるから，$x(Wx_j)/n(Wx_j)$ の値はその平均値の近くにかたまるはずである。逆に x の出現がいずれかの Wx_j に偏っていれば，$x(Wx_j)/n(Wx_j)$ の分布は，そのちらばりの程度が大きくなるはずである。このことより，$x(Wx_j)/n(Wx_j)$ の分布のちらばりの特性値は，x がいずれかの Wx_j に偏って出現している程度を示す値であると解釈することができる。同様に，それぞれの部分集合 Wx_j の大きさがほぼ等しいことを仮定すれば，$x(Wx_j)$ の分布のちらばりの特性値を x の偏在性の測度とみなすことが可能である。

(C.2), (C.5) にはいずれも W-WD 関係における⑥の数量関係が認められるが，ちらばりの特性値として前者は分散を，後者はカイ2乗を用いている。また (C.1), (C.4), (C.6) には W-WD 関係における，(C.3) は WG-WD 関係における⑦の数量関係が認められる。ちらばりの特性値として (C.1) と (C.3) では分散を，(C.4) と (C.6) ではカイ2乗を用いている。

(C.1)～(C.6) は，x の偏りの測度としてちらばりの特性値を代用している重みづけであるが，これに対して，(D.1) は x の偏りの測度を独自の分布モデルに基づいて求める重みづけである。(D.1) から求められる値は，単語の出現が 2-ポワソン・モデルに従っている場合，クラスⅠの文献内出現頻度の平均がクラスⅡの文献内出現頻度よりもどの程度大きいかを示す値である。この値は，単語がクラスⅡよりもクラスⅠの文献集合に多く出現するほど大きな値を示すから，「単語がクラスⅠに偏って出現している程度」と解釈することができる。ここでは，単語の偏在性が，2つのクラスのへだたりというかたちで数量化されているといえよう。

ちらばりの特性値を用いた重みづけと，2-ポアソン分布に基づく重みづけは，特定の Wx を想定せずに単語の偏在性が数量化されているという点で，前節までの方法とは異なっている。これらの重みづけで数量化されているのは，「x が Wy においていずれかの Wx に偏って存在している程度」あるいは「x が Wy においていくつかの Wx に偏って存在している程度」である。これらの場合には，重みづけは特定の Wx の要素ではなく，Wy のすべての要素を対象として行われる。

E. 自己情報量と平均情報量

算出式 (E.1)～(E.5) には，次のような共通の数量関係を見いだすことができる。

⑧ 単語 x の重みは，「単語 x が Wx 中に出現する」という確率事象の自己情報量の増加に伴って単調に増加

(E.1) は，T-IT′ 関係において単語 x が Wx 中に出現する事象の自己情報量を求める式であり，(E.2) と (E.3) はこれを修正したかたち，(E.4) はこれに①の数量関係を組み合わせたかたちと解釈することができる。同様に (E.5) は，W-WD 関係において単語 x が Wx 中に出現する事象の自己情報量を求める式に，①の数量関係を組み合わせた式と解釈できる。

「自己情報量」は，単語の偏在性という考え方からは導きにくい概念である。しかし，式の成り立ちを見ると，そこには偏在性の原理に基づいた，すでに指摘した数量関係が存在していることがわかる。(E.1)～(E.3) には③の数量関係が，(E.4) には①と③の数量関係が存在している。また (E.5) の式には，W-WD 関係における①～③の数量関係を，(A.12), (A.17), (A.18) と同じように見いだすことができる。

さて，Salton は，(E.6)～(E.8) の算出式で用いられているシグナルおよびノイズと呼ばれる数量が，情報理論に基づいて導かれたものだと説明しているが，すでに述べたようにこれがどのように導かれたのかについては説明していない。そこで，これらの算出式の意味を「平均情報量」すなわち「エントロピー」の考え方に基づいて推定すると，以下のようになるであろう。

まず，F_i 個の単語 w_i の出現を F_i 個の事象からなる排反な事象系と考えたとき，各文献への出現の状態が不明なときは，各事象の確率は $1/F_i$ で近似される。したがって，この事象系の平均情報量は $\log F_i$ となる。一方，単語 w_i の各文献への出現の状態がわかっているときは，単語 w_i が N 個の文献に出現する過程を，N 個の事象からなる排反な事象系であると考えることができる。このとき，単語 w_i が文献 d_i に出現するという確率は f_{ij}/F_i で近似されるので，この事象系の平均情報量

は次の式で求められる。

$$-\sum_j \frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i}$$

この式はノイズを求める式とまったく同等であるから，ノイズNZはこの事象系のエントロピーとみなすことができる。すると，ノイズが「単語 w_i の各文献への出現状態」を知った後の状況の平均情報量，すなわちエントロピーとみなせるのに対し，シグナルは「単語 w_i の各文献への出現状態」を知る前と知った後の状況のエントロピーの差であり，「単語 w_i の各文献への出現状態」を知ったときに得られる情報量に -1 を乗じた値と解釈することができる。

以上より，(E.6)～(E.8) には次のような共通の数量関係を見いだすことができる。

⑨ 単語 x の重みは，単語 x の Wx への出現状態に関するエントロピーの増加に伴って単調に減少

第3表　各算出式にみられる数量関係

式番号	Wx-Wy関係	数量関係 ① ② ③	④ ⑤	⑥ ⑦	⑧ ⑨
(A. 1)	⟨W⟩-⟨WD⟩	○			
(A. 2)	T-IT′	○			
(A. 3)	⟨W⟩-⟨WD⟩	○ ○			
(A. 5)	W-WD	○			
(A. 6)	W-WD	○			
(A. 7)	W-WD	○ ○			
(A. 8)	W-WD	○ ○			
(A. 9)	⟨W⟩-⟨WD⟩	○ ○			
(A.10)	W-WD	○ ○			
(A.11)	⟨W⟩-⟨WD⟩	○ ○ ○			
(A.12)	W-WD	○ ○ ○			
(A.13)	T-IT′	○ ○			
(A.14)′	T-IT′	○ ○ ○			
(A.15)′	T-IT′	○ ○ ○			
(A.16)	W-WD	○ ○			
(A.17)	W-WD	○ ○ ○			
(A.18)	W-WD	○ ○ ○			
(A.19)	Wi-IT′	○ ○			
(A.20)	Wi-IT′	○ ○			
(A.21)	Wi-IT′	○ ○			
(A.22)	Wi-IT	○ ○			
(A.23)	WG-WD	○			
(A.24)	WG-WD	○ ○			
(A.25)	WG-WD	○ ○			
(B. 1)	W-WG		○ ○		
(B. 2)	W-WG		○ ○		
(B. 3)	W-WG		○ ○		
(B. 4)	W-WG		○ ○		
(B. 5)	W-WD		○ ○		
(B. 6)	W-WD		○ ○		
(B. 7)	W-WD		○ ○		
(B. 8)	W-WD		○ ○		
(B. 9)	WG-WD		○ ○		
(B.10)	WG-WD		○ ○		
(B.11)	WD-NL		○ ○		
(B.12)	WD-NL		○ ○		
(B.13)	WD-NL		○ ○		
(B.14)	WD-NL		○ ○		
(B.15)	T-IT′		○ ○		
(B.16)	T-IT′		○ ○		
(B.17)	W-WD		○ ○		
(B.18)	W-WD		○ ○		
(B.19)	W-WD		○ ○		
(B.20)	T-IT′		○ ○		
(C. 1)	W-WD			○	
(C. 2)	W-WD			○	
(C. 3)	WG-WD			○	
(C. 4)	W-WD			○	
(C. 5)	W-WD			○	
(C. 6)	W-WD			○	
(D. 1)	W-WD				
(E. 1)	T-IT′	○			○
(E. 2)	T-IT′	○			○
(E. 3)	T-IT′	○			○
(E. 4)	T-IT′	○ ○			○
(E. 5)	W-WD	○ ○ ○			○
(E. 6)	W-WD				○
(E. 7)	W-WD				○
(E. 8)	W-WD	○ ○ ○			○

エントロピーは，各事象の生起確率が等しくなるとき最大値をとる。すなわち，事象の生起がまったくでたらめで予測がつかないとき，最大となるのがエントロピーである。事象の生起のばらつきの程度が大きくなるほどエントロピーは小さくなる。⑨におけるエントロピーの値は，単語の出現がいずれかの Wx に偏っていればいるほど小さくなる。したがって（E.6)～(E.8）は，x の偏りの測度としてエントロピーを用いた重みづけであると解釈することができる。

なお，⑨におけるエントロピーは，特定の Wx とは無関係に求められる値であるから，（E.6）と（E.7）によって求められる値は，Wy のすべての要素についての重みである。この点で，これらの方法は前節の方法と同じである。これに対して（E.8）は，シグナルに x(Wx) を乗じることにより，特定の Wx の要素についての重みづけの式に修正されている。

ところで，シグナルと呼ばれる数量のもつ意味に関する上述の推測が正しいとすれば，Salton の提示した式には若干の誤解が含まれているように思われる。SG_i を求める式の右辺の第1項は，たしかに「F_i 個の事象からなる排反な事象系」の平均情報量と推測できるのだが，はたして F_i 個の単語 w_i の出現を F_i 個の事象からなる事象系と考えてよいものだろうか。単語の出現を確率事象系と考えるのならば，F_i 個の単語 w_i の出現は，「単語 w_i に関する F_i 回の試行」と考える方が自然ではないだろうか。

もしもそう考えるのならば，ノイズを求める式と同じように，N 個の事象からなる排反な事象系を想定するのが妥当である。N 個の事象からなる排反な事象系を想定すれば，「単語 w_i の各文献への出現状態」を知る前の平均情報量は $\log N$ となり，SG_i を求める式は次のようになるはずである。

$$SG_i = \log N - NZ_i$$

Salton の示した SG_i を求める式には，確率事象の仮定に関して一貫性が欠けているように思われる。

VI. おわりに

本研究の目的は，出現頻度情報に基づいた単語重みづけの原理を明らかにすることであった。II章では，まず，単語の重みづけが情報検索システムのどのような機能と関係づけられるのかを明らかにするために，情報検索とインデクシングを簡単なモデルにして表現し，これに基づいて重みづけには4つの相があることを説明した。III章では，重みづけが行われる文献空間と語彙空間を定義し，その上で単語の重みを算出するために用いられる基本的な数量を定義した。以上は，多様な重みの算出方法を，同じ枠組みで記述するための準備であった。

IV章では，従来提示されてきた重みの算出式を，(1) 基本的な数量の単純な組み合せによる方法，(2) 2つの相対出現頻度を用いた方法，(3) ちらばりの特性値を用いた方法，(4) 2-ポアソン・モデルに基づく方法，(5) Shannon の情報量の概念を用いた方法の5つのグループに分類して説明した。説明した算出式は，合わせて60余りであった。

そしてV章では，算出式を3つのグループに分け，そこに共通に見いだされる数量関係を抽出し，そこに示されている数量のふるまいがいずれも偏在性の原理に従っていることを明らかにした。偏在性の原理とは，「特定の単語に与える重みは，その単語が特定の単語集合に偏って出現している程度に応じた値である」という考え方であった。

ところで，前章までの説明で明らかなように，単語の重みづけにおける偏在性の原理は文献を単語の集合とみなして扱うことを前提とする考え方である。ここでは，単語の集まりが本来もっているさまざまな構造，あるいは体系はすべて捨象されている。すなわち，単語の集合のもっている語彙構造，文のもっている統語構造，テキストのもっている意味構造は，偏在性の原理に従う方法では考慮されることはない。この原理に従う限り，文献は単語の並びでさえなく，順序さえもたない単語の寄せあつめとして扱われる。

従来多くの研究者が試みてきた出現頻度情報に基づく単語の重みづけに，このような限界があることを認識することは重要である。限界を認識した上で，どこまで実用的かつ効果的なシステムが構築可能かを探ることが，今後の研究の大きな課題である。

なお，この研究にあたり，ご指導いただいた東京大学教育学部の長澤雅男教授に感謝の意を表したい。

1) この章で示すモデルの作成にあたっては，以下の文献を参考にした。
有川節夫，武谷竣一，“検索システムの数学モデル”，情報管理，Vol. 21, No. 11, p. 865-879 (1979).
中井浩，“インデクシングの数学モデル”，情報管理，Vol. 21, No. 12, p. 947-955 (1979).

伊藤哲郎，情報検索システム，昭晃堂，1986, p. 139-154.
2) Sparck-Jones, K. "Index Term Weighting", Information Storage and Retrieval, Vol. 9, p. 619-633 (1973).
3) Luhn, H. P. "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, Vol. 2, No. 2, p. 159-165 (1958).
4) Luhn, H. P. "A New Method of Recording and Searching Information", American Documentation, Vol. 4, No. 1, p. 14-16 (1953).
5) Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information", IBM Journal of Research and Development, Vol. 1, No. 4, p. 309-317 (1957).
6) Sager, W. K. H.; Lockemann, P. C. "Classfication of Ranking Algorithms", International Forum on Information and Documentation, Vol. 1, No. 4, p. 41-46.
7) Noreault, T. et al. A Performance Evaluation of Similarity Measure, Document Term Weighting Schemes and Representations in a Boolean Environment, 〈Oddy, R. N. eds. Information Retrieval Research, London, Butterworths, 1977〉 p. 57-76.
8) 加藤緑，里見研一，石坂哲郎，"分野別用語集のための語の選定方法に関する実験的な検討"，第7回情報科学技術研究集会発表論文集，p. 319-326 (1970).
9) 加藤緑，石坂哲郎，植島勇夫，"文献キーワードの自動抽出に関する実験的な検討"，第8回情報科学技術研究集会発表論文集，p. 143-151 (1971).
10) Edmundson, H. P.; Wyllys, R. E. "Automatic Abstracting and Indexing - Survey and Recommendations", Communication of the ACM, Vol. 4, No. 5, p. 226-234 (1961).
11) Damerau, F. J. "An Experiment in Automatic Indexing", American Documentation, Vol. 16, No. 4, p. 283-289 (1965).
12) 後藤智範 et al. "電気工学分野における重要漢字の調査" '83年度三田図書館・情報学会研究大会要旨，p. 19-22 (1983).
13) 後藤智範 et al. "出現頻度に基づく重要漢字と主題分野との関連"，第21回情報科学技術研究集会発表論文集，p. 209-215 (1984).
14) 細野公男 et al. "漢字の出現頻度情報を用いた日本語文献の自動分類"，自然言語処理研究会資料，No. 47-7 (1985).
15) 田中康仁，岡坂良雄，"専門用語の自動抽出—英単語頻度辞書を用いて—"，自然言語処理研究会資料，No. 29-4 (1982).
16) Kucera, H.; Francis, W. N. Computational Analysis of Present-Day American English, Providence, Brown University Press, 1967.
17) Carroll, J. M.; Roelloffs, R. "Computer Selection of Keywords Using Word-Frequency Analysis", American Documetation, Vol. 20, p. 227-233 (1969).
18) Dennis, S. F. The Design and Testing of a Fully Automated Indexing-Searching System for Documents Consisting of Expository Test, 〈Schecter, G. ed. Information Retrieval - a Critical Review, Washington, D. C., Thompson Book Co., 1967〉 p. 67-94.
19) Stone, D. C.; Rubinoff, M. "Statistical Generation of a Technical Vocabulary", American Documentation, Vol. 19, No. 4, p. 411-412 (1968).
20) 竹内晴彦，岩坪秀一，西野博二，"多変量解析によるキーワードの自動抽出と文献の自動分類"，自然言語処理研究会資料，No. 54-2 (1986).
21) 長尾真，落合和博，水谷幹男，"日本語文献検索におけるカイ2乗を使った重要語自動抽出"，昭和49年度電子通信学会全国大会，p. 1600 (1974).
22) 長尾真，水谷幹男，池田浩之，"日本語文献における重要語の自動抽出"，情報処理，Vol. 17, No. 2, p. 110-117 (1976).
23) Harter, S.P. "A Probabilistic Approach to Automatic Keyword Indexing—Part I. On the Distribution of Speciality Words in a Technical Literature", Journal of the American Society for Information Science, Vol. 26, No. 4, p. 197-206 (1975).
24) Harter, S. P. "A Probabilistic Approach to Automatic Keyword Indexing—Part II. An Algorithm for Probabilistic Indexing", Journal of the American Society for Information Science, Vol. 26, No. 5, p. 280-289 (1975).
25) Shannon, C. E. The Mathematical Theory of Communication, 〈Shannon, C. E.; Weaver, W. The Mathematical Theory of Communication, Urbana, University of Illinois Press, 1964〉 p. 29-125.
26) Robertson, S. E. "Specificity and Weighted Retrieval", Journal of Documentation, Vol. 30, No. 1, p. 41-46 (1974).
27) Sparck-Jones, K. "A Statistical Interpretation of Term Specificity and its Application in Retrieval", Journal of Documentation, Vol. 28, No. 1, p. 11-21 (1972).
28) Salton, G.; McGill, M. J. Introduction to Modern Information Retrieval, New York, McGraw-Hill International Book Company, 1983, p. 59-71.
29) Salton, G. Dynamic Introduction and Library Processing. London, Prentice-Hall, 1975, p. 79-97.