Library and Information Science No. 28 1990

索引語間の関連性を考慮した情報検索モデル

A Term Dependence Model in Information Retrieval

谷 口 祥 一
Shoichi Taniguchi

Résumé

In most information retrieval systems or models, the assumption is normally made that index terms assigned to the documents of a collection occur independently of each other. So as to improve the retrieval effectiveness of systems, there is a need to take dependencies between certain index term pairs into account.

As the similarity measure between a query and a document is important in quantitative retrieval, two measures, which reflect directly the relationships between index terms when they are given by pairwise correlations, are proposed in this paper. One of the proposed measures is an extension of the cosine function model. This measure is based on oblique coordinates whose degree of angle between axes corresponds to the pairwise correlation between index terms, in contrast to the conventional cosine function measure based on rectangular coordinates. The other measure is an extension of the extended Boolean model, which was proposed by G. Salton et al. Using these measures, we need no assumption of term independence.

Retrieval experiments to evaluate the proposed measures was performed on a test collection of 623 document records and 5 queries, in a weighted mode, in which index terms assigned to the document record were weighted, and in an unweighted mode. The experiment showed following results: 1) it is useful to incorporate term dependencies into the similarity measures; and 2) the proposed measures, however, did not have much better effectiveness than conventional ones.

谷口祥一: 図書館情報大学助手, 茨城県つくば市春日 1-2
Shoichi Taniguchi: University of Library and Information Science, 1-2 Kasuga, Tsukuba-shi, Ibaraki-ken.
1991年2月23日受付

I. はじめに

情報検索の領域における重要な問題の1つに，索引語間の関連性の取扱いに関わる問題がある。従来の主たる検索システムや検索モデルでは，質問を構成する語が検索対象となる各文献に索引語として付与されているか否かの点にのみ基づき，当該文献の質問に対する適合蓋然性評価すなわち適合文献である確からしさの度合い決定が行われていた。このように索引語間の関連性を考慮しない，換言すれば索引語間の独立性が仮定されている場合には，検索に用いた語の同義語や関連語により索引づけされた文献が検索されない結果となる点は，周知の事柄である。ある事例においては単純な共出現状況を調べても 30% 近くの索引語の組に正の相関関係が観察されており[1]，索引語間の関連性を何らかの方法により考慮することが検索効率を上げるためには必要不可欠な事柄となる。

現段階で実用化されているブール型の検索システムでは，シソーラスのシステム内組み込み，および検索時のシソーラス参照による同義語，上・下位語，類義語等の表示または検索語への追加等をもってこの問題に対処してきた。一方，質問と各文献間の類似性を何らかの類似尺度を用いて定量化し，検索をおこなおうとする，いわゆる定量的検索の場合には，この問題に対する多様な接近が可能である。これまでに提案されたものには，1) ベクトル型モデルに基づくもの，2) 確率型モデルに基づく木従属モデル (tree dependence model)[2),3)] や BLE モデル (Bahadur-Lazarsfeld expansion model)[4),5)]，あるいはファジィ事象の確率概念を確率型モデルに適用・展開したもの[6)]，3) 集合論型モデルに基づくもの[7)] などがある。これらはすべて索引語間の関連性が数値化された関連度をもって与えられていることを前提としたものである。

本稿では，これらと同様な考え方に基づき，索引語間の関連性を文献の類似度計測に反映できるよう既存の検索モデルに対して拡張を試みる。具体的には，1) ベクトル型モデルに属するコサイン関数モデル (cosine function model) に拡張を施したもの，および 2) 拡張ブール型モデル (extended Boolean model) に拡張を施したものの2つを提案し，実験によりその有効性の検証を試みる。

II. コサイン関数モデルの拡張

A. コサイン関数モデル

コサイン関数モデルが属するベクトル型モデルとは，文献および質問の内容・主題を索引語のベクトルをもって表現するものである。これにより，索引語間の論理関係までを含めて質問設定を行う必要はなくなり，また質問に対して各文献のとる類似度に基づき順位づけされた出力が容易にえられることになる。

具体的には，検索対象となる文献集合 $D=\{d_1, d_2, \ldots, d_m\}$，それに含まれる各文献 d_i の索引づけおよび質問設定に用いられる索引語集合 $T=\{t_1, t_2, \ldots, t_n\}$ が与えられたとき，ベクトル型モデルとはベクトル表現された質問 $q=(q_1, q_2, \ldots, q_n)^T$ を文献への索引づけ関数 $X: D\times T\rightarrow[-\infty, +\infty]$，または表現を換えて行列 $X=(x_{ij})$ に基づき変換を行い，応答ベクトル $r=(r_1, r_2, \ldots, r_m)^T$ をえようとするものと定義される。この検索過程を行列算表現すれば $r=Xq$ と表せ，個々の文献のレベルで表せば質問-文献ベクトル間の内積 $r_i=\sum_j x_{ij}q_j$ となることがわかる。これが同モデルの最も簡単な表現である。

なお，多くの場合に質問中の各索引語の重み q_j を

$0 \leqq q_j \leqq 1$, 文献への索引づけ関数 X を $X: D \times T \to [0, 1]$, すなわち $0 \leqq x_{ij} \leqq 1$ に便宜的に仮定するため，本稿でもこれらの条件を前提にして議論を進めることにする。さらには，質問に対する類似度計測値を表す応答ベクトル中の成分 r_i の値を区間 [0, 1] に収めるべく正規化をおこなうことが多く，よって $r=CXq$, C: 正規化係数対角行列, $c_{ij}=0$ $(i \neq j)$ となる。この正規化には例えば $c_{ii}=1/[\sum_j x_{ij}^2 \cdot \sum_j q_j^2]^{1/2}$ や $c_{ii}=1/[\sum_j x_{ij}+\sum_j q_j-\sum_j x_{ij}q_j]$ など，多数のものが提案されている。この正規化処理において前者の式を採用したものが，本稿で索引語間関連度の組み入れを試みるコサイン関数モデルである[8]。ここに改めてその式を記す。

$$r_i=\frac{\sum_j x_{ij}q_j}{[\sum_j x_{ij}^2]^{1/2} \cdot [\sum_j q_j^2]^{1/2}} = \frac{(x_i, q)}{|x_i||q|}=\cos\theta_{(x_i,q)} \qquad (1)$$

ここで x_i は文献 d_i に対応する文献ベクトルを表している。この式の表す意味をみてみると，各索引語が基底ベクトルをなす n 次元座標系において，質問および文献ベクトルを各々の成分に従い設定し，その2つのベクトルのなす角をもって質問-文献間の類似度としていることがわかる。$0 \leqq x_{ij}, q_j \leqq 1$ により，質問 q と任意の文献 d_i とのなす角 $\theta_{(x_i,q)}$ は $0 \leqq \theta_{(x_i,q)} \leqq \pi/2$ となり，対応するコサイン関数の値 $\cos\theta_{(x_i,q)}$ は $0 \leqq \cos\theta_{(x_i,q)} \leqq 1$ となる。2つのベクトルが重なり合うとき $(\theta_{(x_i,q)}=0)$ には類似度 $\cos\theta_{(x_i,q)}=1$ であり，逆に垂直となるとき $(\theta_{(x_i,q)}=\pi/2)$ には類似度 $\cos\theta_{(x_i,q)}=0$ となる。

B. コサイン関数モデルへの索引語間関連度の組み入れ

まず，索引語間の関連性が数値化して与えられている，換言すれば索引語間の関連度を与える関数 $Y: T \times T \to [0, 1]$ が与えられ，それに基づく索引語間関連行列 $Y=(y_{jk})$ がえられているものと仮定する。2つの索引語 t_j と t_k とが同義語関係にある場合には関連度 $y_{jk}=1$ とし，関連がない場合には関連度 $y_{jk}=0$ となるものと仮定する。このとき関連行列 Y は反射律 $y_{jj}=1$ を満たし，さらに反意語関係や上・下位語の区別を捨象した同義語・関連語関係のみを表現するものであるならば対称律 $y_{jk}=y_{kj}$ を満たすと考えるのが妥当であろう。

これにより，前節の先頭で定義したベクトル型モデルにおいては，この関連行列 Y を用いて検索処理を $r=CXYq$ と展開することができ，索引語間の関連度を文献の類似度計測に反映させることが可能となる。これを個々の文献のレベルで表せば，$r_i=c_{ii} \cdot \sum_{j,k} x_{ij}y_{jk}q_k$ となる。

同様にコサイン関数モデルに対して索引語間関連行列 Y を組み入れることを次に試みる。コサイン関数モデルもベクトル型モデルの1つであるため，$r=CXYq$ と展開することで基本的には十分であるが，その正規化処理も含めて n 次元座標系に即した解釈を試みてみよう。前節の終わりでみたように，コサイン関数モデルは各索引語が基底ベクトルをなす n 次元の座標系における質問ベクトルと文献ベクトルとのなす角を当該文献の類似度とするものであった。このとき無条件に前提としていた座標系は直交系であり，基底ベクトルは互いに直交する，換言すれば線形独立とされていた。このような視点を最初に提出したのは Wong と Raghavan であった[9),10)]。そこで，この前提とされていた直交座標系をその1つの特殊例として包含するような，索引語間関連度に対応した角度で基底ベクトルが設定されている斜交座標系に質問-文献間類似度の計測空間を拡張して考えることで，索引語間の関連性をコサイン関数モデルに組み入れることができよう。そして検索処理では，設定された斜交座標系上に質問および文献ベクトルを配置し，対応する直交座標系に各々のベクトルの正射影をとることにより，それらがなす角を基底ベクトルの交角を踏まえた値で計測することが可能となろう。その結果，索引語間関連性を反映した検索処理結果がえられることになる。この処理の定式化を以下でおこなう。

任意の n 次元座標系 $\Gamma=(O; e_1, e_2, \ldots, e_n)$ に関して，任意のベクトル a, b が

$$a=a_1e_1+a_2e_2+\cdots+a_ne_n$$
$$b=b_1e_1+b_2e_2+\cdots+b_ne_n$$

$|e_1|=|e_2|=\cdots=|e_n|=1$, a_j, b_j: スカラー，であるとする。内積 (a, b) は内積の分配律を用いて

$$(a, b)=\sum_{j,k} a_jb_k(e_j, e_k)$$

となる。ここで基底ベクトル e_j と e_k のなす角を $\omega_{(j,k)}$ とすると，$(e_j, e_k)=|e_j||e_k|\cos\omega_{(j,k)}=\cos\omega_{(j,k)}$ となることより，

$$(a, b)=\sum_{j,k} a_jb_k\cos\omega_{(j,k)}$$

がえられる。これより，$(a, a)=|a|^2$ に基づき

$$|a|=[\sum_{j,k} a_ja_k\cos\omega_{(j,k)}]^{1/2}$$

をえる。よって，ベクトル a, b のなす角を $\theta_{(a,b)}$ で表すと，次のものをえる。

$$\cos\theta_{(a,b)}=\frac{(a,\,b)}{|a||b|}$$

$$=\frac{\sum_{j,k} a_j b_k \cos\omega_{(j,k)}}{[\sum_{j,k} a_j a_k \cos\omega_{(j,k)}]^{1/2}\cdot[\sum_{j,k} b_j b_k \cos\omega_{(j,k)}]^{1/2}}$$

続いて任意の基底ベクトル e_j と e_k のなす角 $\omega_{(j,k)}$ を決定する。これは索引語 t_j, t_k 間の与えられた関連度 y_{jk} に対応しなければならない。関連度 y_{jk} は，1) $0\leqq y_{jk}\leqq 1$ で，かつ 2) 反射率および対称律を満たすものとする。いま，ベクトル a が a_j 成分のみ 1，他は 0 であり，ベクトル b が b_k 成分のみ 1，他は 0 である場合 ($|a|=|b|=1$) を想定してみると，ベクトル a と b のなす角は索引語 t_j, t_k 間の関連度 y_{jk} に等しくならなければならない。これより，

$$\cos\theta_{(a,b)}=\cos\omega_{(j,k)}/1\cdot 1=y_{jk}$$

となる。設定しようとする座標系の条件，1) 索引語 t_j, t_k 間に関連がない ($y_{jk}=0$) ならば，e_j と e_k は垂直 ($\omega_{(j,k)}=\pi/2$)，2) 索引語 t_j, t_k は交換可能 ($y_{jk}=1$) ならば，e_j と e_k は重なり合う ($\omega_{(j,k)}=0$) により，e_j と e_k のなす角は $0\leqq\omega_{(j,k)}\leqq\pi/2$，対応するコサイン関数の値は $0\leqq\cos\omega_{(j,k)}\leqq 1$ となることより次式をえる。

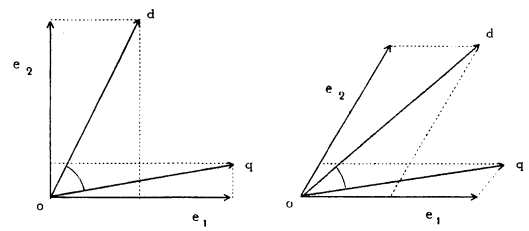
$$\omega_{(j,k)}=\cos^{-1} y_{jk} \qquad (2)$$

索引語間関連行列 Y が与えられれば，上式により e_1, e_2, …, e_n で生成される n 次元の斜交座標系が一意に設定されることになる。以上のことより，設定された座標系で質問ベクトル q と文献 d_i を表す文献ベクトル x_i とのなす角度，すなわちその類似度は，(1) 式に換えて次式で算出することになる。

$$r_i=\frac{\sum_{j,k} x_{ij} y_{jk} q_k}{[\sum_{j,k} x_{ij} y_{jk} x_{ik}]^{1/2}\cdot[\sum_{j,k} q_j y_{jk} q_k]^{1/2}} \qquad (3)$$

これが求める，索引語間関連度を組み入れた類似度計測式である。第 1 図中の左側に従来のコサイン関数モデルを，右側に本稿で提案したその拡張モデルを，各々 2 次元の場合について図示する。

次に，(3) 式で示される質問-文献ベクトル間の角度計測をおこなう n 次元斜交座標系が生成されるための必要十分条件を考えると，それは次の定理で示されるものとなる。



第 1 図　コサイン関数モデルへの索引語間関連度の組み入れ (2 次元の場合)
$q=(1.0,\ 0.2)$, $d=(0.5,\ 1.0)$ の例

[定理]　n 次元斜交座標系 (すべての値が正となる象限のみの座標系に限定) を生成するためには，基底ベクトル間の角度に関して次の関係が成立していなければならない。

$$\max_l |\omega_{(j,l)}-\omega_{(l,k)}|\leqq\omega_{(j,k)}\leqq\min_l (\omega_{(j,l)}+\omega_{(l,k)}) \qquad (4)$$

$$0\leqq\omega_{(j,k)},\ \omega_{(j,l)},\ \omega_{(l,k)}\leqq\pi/2$$

(証明)　n 次元斜交座標系が成立するためには，任意の 3 つの基底ベクトル e_j, e_k, e_l の各頂点を結んだ線分により三角形が形成される，すなわち次の三角不等式が成立しなければならない。

$$|e_k-e_l|\sim|e_l-e_j|\leqq|e_j-e_k|\leqq|e_k-e_l|+|e_l-e_j|$$

$$0\leqq|e_j-e_k|,\ |e_k-e_l|,\ |e_l-e_j|\leqq\sqrt{2}$$

ここにおいて，例えば $|e_j-e_k|=2\cdot\sin(\omega_{(j,k)}/2)$ であるので，同様に他の線分の長さをすべて基底ベクトル間の角度の関係に置き換えることができる。よって，それを簡約化することにより (4) 式がえられる。(証明終)

さらに索引語間関連行列 Y が関係

$$\max_l |y_{jl}-y_{lk}|\leqq y_{jk}\leqq\min_l (y_{jl}+y_{lk})$$

を満たすときには，(2) 式により，その必要条件として (4) 式を導くことができる。

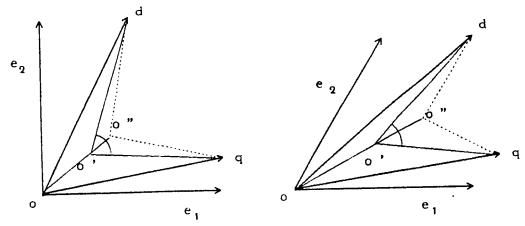
以上で提案したコサイン関数モデルの拡張は，Wong と Raghavan の提出した視点を踏襲し，展開したものであるが，その後彼らは索引語の論理和・論理積にそれぞれ対応する直交座標系の設定をおこなうベクトル型モデルを新たに考案し，その座標系上で質問-文献間の類似度計測をおこなう検索モデルを提案している[11)-14)]。また，比較的類似したモデル化に Deerwester らが用いた singular-value decomposition と呼ばれる手法があり，これは固有値解析によってえられる大きな固有値に

対応した固有ベクトルによる直交座標系を設定し，その縮小された次元において質問-文献間の本来の類似度を近似しようとするものである[15]。

C. コサイン関数モデルへの論理演算の導入

ベクトル型モデルは，それに属するコサイン関数モデルを含めて，論理演算を排除することにより，次の2点を回避できることを大きな特徴としている。1) 検索者にとり通常それほど明確でない，質問中の索引語間の論理関係（論理和・積関係）までを含めて検索質問を構成しなければならない困難さ。および，2) 論理形式をとり設定された質問に対する2値論理に基づく検索処理から生じる検索結果集合の極端な限定性（部分的に適合する文献すべてを切り捨てたり，あるいはほとんど不適合と考えられる文献を大量に検索したりする結果となること）。しかしながら，場合によっては論理形式をとる質問に即した論理演算処理が要求されることもあろう。また，この点がより重要であると考えられるが，ベクトル型の拡張によってブール型，ファジィ型，確率型等，他の検索モデルとの接合・統合を図るためには，その準備作業としてベクトル型モデルへの論理演算導入の試みは意義のある事柄となる。本節ではコサイン関数モデルに対する論理演算の適用法の1つを示し，その論理演算処理においては前節で提案した索引語間関連度の組み入れがそのまま適用できることを確認する。

まず，従来のコサイン関数モデルすなわち (1) 式で表されるモデルを論理和検索をおこなうものと仮定しよう。すると，質問-文献ベクトル間のなす角がその文献の類似度を表すものとされることより，論理積演算ではそのなす角が検索者により設定された論理積の度合いに応じて大きくなるものと考えられる。これにより，例えば2値論理における論理積と同等な強さの論理積演算としたり，あるいはより論理和に近い値となる論理積演算とすることも可能となる。そこで第2図中の左側に2次元の場合を図示したように，必要とする論理積の度合いに応じて質問および文献ベクトルを設定する基点（図中の O'）を移動し，その移動先の地点における質問-文献ベクトル間のなす角を計測することが考えられよう。これにより，質問-文献ベクトル間のなす角は原点 O における角度を最小値として，質問-文献ベクトルが垂直となる地点（図中の O''）における角，すなわち $\pi/2$ まで，変化させることが可能となる。この処理を定式化したものを次に示す。



第2図　コサイン関数モデルへの論理積演算の導入（2次元の場合）
$q=(1.0,\ 0.2)$, $d=(0.5,\ 1.0)$ の例

点 O' は，文献 d_i を表す文献ベクトル x_i と質問ベクトル q とのベクトル和を指す点 $O_{(x_i+q)}$ と原点 O の2点を通る直線上にとられた線分 $\overline{OO''}$ 上を動くものとする。ここで点 O'' は前述したように，質問-文献ベクトルが互いに垂直となる点を指している。このとき，点 $O'=(o_1', o_2', \ldots, o_n')$ における質問-文献ベクトル間のなす角のコサイン関数値 $\cos\theta_{o'}$ は次の式で表され，これが論理積演算による類似度 $r_i(\mathrm{AND})$ ともなる。

$$r_i(\mathrm{AND})=\cos\theta_{o'}=\frac{\sum_j (x_{ij}-o_j')(q_j-o_j')}{[\sum_j (x_{ij}-o_j')^2]^{1/2}\cdot[\sum_j (q_j-o_j')^2]^{1/2}}$$

さらに点 O' の位置を制御する変数，すなわち論理積の度合いを制御するパラメータ p を導入すると，上記の条件より $o_j'=(x_{ij}+q_j)p$ であり，よって前式は p の関数として表すことができる。

$$r_i(\mathrm{AND})=(1/\delta_i)\cdot[\sum_j (x_{ij}-(x_{ij}+q_j)p)(q_j-(x_{ij}+q_j)p)] \quad (5)$$

$$\delta_i=[\sum_j (x_{ij}-(x_{ij}+q_j)p)^2]^{1/2}\cdot[\sum_j (q_j-(x_{ij}+q_j)p)^2]^{1/2}$$

次に，点 O' における質問-文献ベクトル間のなす角 $\theta_{o'}$ は $\theta_o<\theta_{o'}<\pi/2=\theta_{o''}$ とするため，p のとりうる値は次のものとなる。

$$0<p<(1-\beta_i)/2$$

$$\beta_i=[\sum_j (x_{ij}-q_j)^2/\sum_j (x_{ij}+q_j)^2]^{1/2}$$

ここで現れた β_i は，質問ベクトルと文献ベクトルとの和と差の長さの比を指している。$p=0$ のときには，(5) 式は論理和演算をおこなうものと仮定した (1) 式と同値となる。よって，(1) 式により計測される類似度を $r_i(\mathrm{OR})$ と表せば，$r_i(\mathrm{AND})=r_i(\mathrm{OR})$ となり，(5) 式は

論理積の効力はもたないことがわかる。$p=(1-\beta_i)/2$ のときには，$r_i(\mathrm{AND})=0$ となり，すべての場合に質問-文献間の類似度は0となる。従って，p はその2つの値の間を動き，p の値が0に近ければ論理和演算の結果に近い類似度がえられ，逆に p の値が大きくなるほどより強い論理積演算の結果がえられる。論理和と論理積が組み合わされた複合質問に対しては，論理和の質問部分については(1)式を，論理積の質問部分については上記(5)式を適用し，質問全体に再帰的に適用させていけばよいことになる。

以上でコサイン関数モデルに対する論理演算導入の1つの可能性を試みたわけであるが，上記で定式化した論理積質問に対する類似度計測式についても，前節でコサイン関数モデルに索引語間関連度を組み入れたのと同一の方法で索引語間関連度を類似度計測式に組み入れることが可能である。2次元の場合について図式化したものを第2図中の右側に示しておく。これを式表現すると，(5)式は次式のように書き換えられる。

$$r_i(\mathrm{AND})=(1/\delta_{i(1)}\delta_{i(2)})\cdot \left[\sum_{j,k}(x_{ij}-(x_{ij}+q_j)p)(q_k-(x_{ik}+q_k)p)y_{jk}\right] \tag{6}$$

$$\delta_{i(1)}=\left[\sum_{j,k}(x_{ij}-(x_{ij}+q_j)p)(x_{ik}-(x_{ik}+q_k)p)y_{jk}\right]^{1/2}$$

$$\delta_{i(2)}=\left[\sum_{j,k}(q_j-(x_{ij}+q_j)p)(q_k-(x_{ik}+q_k)p)y_{jk}\right]^{1/2}$$

斜交座標系に即した，論理積の度合いを決定するパラメータ p がとりうる範囲を求めることは(5)式の場合と同様の手法で可能である。斜交系の場合に p がとりうる上限値 $(1-\beta_i')/2$ は前述の直交系の場合との関係において $(1-\beta_i)/2\leqq(1-\beta_i')/2$ となるが，その差は座標系が表す索引語間の関連度を示しており，論理積演算には依存しない値と考えるべきである。従って，索引語間関連度を論理積演算に組み入れた(6)式の場合においても，p のとりうる範囲は(5)式について示したものと同じになる。

D. 索引語間の高次関連度

索引語間関連度の組み入れを試みた前節までのモデル化においては，考慮すべき索引語間の関連性とは任意の2つの索引語間の直接的な関連性のみを想定していた。しかしながら，2つの索引語間の直接的な関連度に加えて，他の索引語を媒介にした2次的な結合による関連度，さらにはそれら2次結合による索引語を媒介にした3次結合の関連度など，高次の索引語間関連度を質問-文献間の類似度計測に反映させることが有用な場合も考えられる。そこで本節では索引語間の高次関連度を算出し，その値を質問-文献間の類似度計測に反映させることを，これまでになされた議論を踏まえて検討してみよう。

これまでのベクトル型モデルの範疇では，索引語間の直接的な関連度を表す索引語間関連行列 Y が与えられれば，その巾乗をとることで高次関連度を表す関連行列がえられるものと議論されてきた。すなわち，2次結合による関連度を表す関連行列はもとの関連行列 Y を2乗することでえられ，同様に n 次結合による関連行列はもとの関連行列を n 乗することでえられるものとされている。これによりすべての高次関連度を含めた索引語間関連行列 Y' をえるには，収束させるためパラメータ λ $(0\leqq\lambda<1)$ を導入し，巾零 $\lim_{n\to\infty}(\lambda Y)^n=0$ の仮定上でのみ，$Y'=I+\lambda Y+(\lambda Y)^2+(\lambda Y)^3+\cdots=(I-\lambda Y)^{-1}$ で計算されるものとされてきた（I: 単位行列）[16),17)]。しかしながらこの方法では，パラメータ λ の決定という困難な問題が発生し，また実際の計算量も大きな制約となる。

そこで索引語間関連行列 Y は，その要素が区間 [0, 1] の値であること，および反射律・対称律を満たすことを前提とすれば，$Y=\Psi\Lambda\Psi^{-1}$, (Λ: 固有値を対角要素とした対角行列，Ψ: 各固有値に対応する固有ベクトルをならべた直交行列）に分解することができる。これにより関連行列 Y の n 乗は $Y^n=\Psi\Lambda^n\Psi^{-1}$ となり，容易に求めることが可能となる。しかしこのような計算をおこなったとしても，すべての高次関連度を含めた関連行列 Y' を求めるには，パラメータ λ を必要とする。

従って，より制約の緩い計算の容易な演算の導入を検討すると，ファジィ行列積をはじめとする max-* 演算の適用にいたる（$Y^n=Y^{n-1}\bigcirc Y$, $n>1$)。max-* 演算には，ファジィ行列積である max-min 演算 $[y^n{}_{jk}=\max_l(y^{n-1}{}_{jl}\wedge y_{lk})]$ や，max-積演算 $[y^n{}_{jk}=\max_l(y^{n-1}{}_{jl}\cdot y_{lk})]$, max-限界積演算 $[y^n{}_{jk}=\max_l(0\vee(y^{n-1}{}_{jl}+y_{lk}-1))]$ など，"*" には区間 [0, 1] で定義される多数の t-norm をあてはめることができる。これら max-* 演算による巾乗を求めると，索引語間関連行列 Y は反射律および対称律を満たすことより，次の関係が導かれる[18)]。

$$Y\leqq Y^2\leqq Y^3\leqq\cdots\leqq Y^{n-1}=Y^n=Y^{n+1}=\cdots$$

これより，$Y'=\bigvee_{j=1}^{\infty}Y^j=Y^{n-1}=\hat{Y}$ がえられ，行列 $\hat{Y}$ は推移的閉包と呼ばれるものである。従って，max-* 演算

のもとでは索引語間のすべての高次関連度を組み入れるにも推移的閉包$\hat{Y}$を用いれば十分であり，かつ行列Yの$(n-1)$乗以下で収束することがわかる。なお，上に例示した3つの max-* 演算を用いて導かれた推移的閉包相互の関係は，限界積≦積≦min を反映して，$\hat{Y}$(max-限界積)≦$\hat{Y}$(max-積)≦$\hat{Y}$(max-min) の関係となる[19]。

また，えられた推移的閉包 $\hat{Y}$ はそれ自身次に示す max-* 推移律を満たす推移行列をなしている[18]。

$$y_{jk} \geqq \max_{l} (y_{jl} * y_{lk}) \Leftrightarrow Y \supseteq Y \bigcirc Y$$

ゆえに推移的閉包自身においては，2つの索引語間の関連度は他のいかなる索引語を介した関連度よりも大きくなる点が理解される。

与えられた索引語間関連行列 Y が max-* 推移律を満たすよう設定されている場合，あるいは max-* 推移律を満たすべきであるとの考えに基づき，報告されている行列置換操作[20]を用いて非推移行列から推移行列をつくりだした場合には，次の関係を満たし，

$$Y = Y^2 = Y^3 = \cdots = Y^{n-1} = Y^n = Y^{n+1} = \cdots = \hat{Y} = Y'$$

関連行列 Y 自身が推移的閉包をなす[18]。これは Y における索引語の関連度が2次結合以上のすべての高次関連度と同値であることを意味している。よって，この場合には関連行列 Y を質問-文献間の類似度計測に組み入れることで，すべての高次関連度を考慮したことになる。

以上でみたように，max-*演算に基づき索引語間関連行列の巾乗を求めることで高次関連度が算出できるものと仮定すれば，その結果えられる高次関連度を含んだ関連行列を，質問-文献間の類似度計測に直接適用できることがわかる。例えば，(3)式中の関連度 y_{jk} に推移的閉包 $\hat{Y}$ の要素 $\hat{y}_{jk}$ を適用すれば，すべての高次関連度を質問-文献間の類似度計測に反映させることが可能となる。あるいは，推移的閉包に達しない時点までの，関連行列Yの任意の巾乗までででファジィ和演算をとったものを用いることも可能であろう。

III. 拡張ブール型モデルの拡張

A. 拡張ブール型モデル

拡張ブール型モデルとは Salton らにより提案された検索モデルであり[21]，その最大の特徴は各論理演算子に付与された論理和/積の結合の強さを指示するパラメータ p の値を変化させることにより，同一式がベクトル型として機能したり，ブール型として機能したり，あるいはその中間の性質を有するものとして機能したりする点にある。同モデルは次に示す論理和演算をおこなう式 r_i(OR) および論理積演算をおこなう式 r_i(AND) の2つから構成されている。

$$r_i(\mathrm{OR}) = \left[\frac{\sum_j x_{ij}{}^p q_j{}^p}{\sum_j q_j{}^p} \right]^{1/p}$$

$$r_i(\mathrm{AND}) = 1 - \left[\frac{\sum_j (1-x_{ij})^p q_j{}^p}{\sum_j q_j{}^p} \right]^{1/p} \qquad (7)$$

$$1 \leqq p < \infty$$

$p=1$ のときには，$r_i(\mathrm{OR}) = r_i(\mathrm{AND}) = \sum_j x_{ij} q_j / \sum_j q_j$ となり，論理演算子の効力はなくなる。これは質問-文献ベクトル間の内積を質問ベクトルを用いて正規化したものに他ならず，ベクトル型モデルの1つと考えられる。一方，$p \to \infty$ のときには，ロピタルの定理を用いることにより $r_i(\mathrm{OR}) = \max_j x_{ij}$，$r_i(\mathrm{AND}) = \min_j x_{ij}$ がえられ(共に $q_1 = q_2 = \cdots = q_n = 1$ の場合)，ファジィ型の結果となる。さらに文献に付与されたすべての索引語の重みが2値 ($x_{ij} = \{0, 1\}$) であるときにはブール型と一致する。パラメータ p が上記以外の値 ($1 < p < \infty$) であるときには，その値に応じてベクトル型とファジィ型もしくはブール型との中間的な特性を有する式がえられる。この関係をまとめると次のものとなる。

$$\begin{array}{ccccc} r_i(\mathrm{AND}) & \leqq & r_i(\mathrm{AND}) & \leqq & r_i(\mathrm{AND}) \\ (p \to \infty) & & (1 < p < \infty) & & (p = 1) \\ = r_i(\mathrm{OR}) & \leqq & r_i(\mathrm{OR}) & \leqq & r_i(\mathrm{OR}) \\ (p = 1) & & (1 < p < \infty) & & (p \to \infty) \end{array}$$

なお，Salton らは (7)式のように論理和演算をおこなう式と論理積演算をおこなう式とを定式化しているが，これら2式は双対な関係にあり，パラメータ p を $-\infty < p \leqq 1$ とした場合には論理和演算をおこなう式と論理積演算をおこなう式とが相互に交換される関係にある。また，論理和と論理積が組み合わされた複合質問に対しては，コサイン関数モデルに論理演算を導入した場合と同様，上記 (7)式を再帰的に適用させて用いることになる。

B. 拡張ブール型モデルへの索引語間関連度の組み入れ

前節において解説した拡張ブール型モデルに索引語間の関連性を組み入れ，さらに展開を図ることは，同モデルが備えている柔軟性と相俟ってその有効性を倍加させ

ることになろう。前章においてコサイン関数モデルに対して索引語間関連度を組み入れ拡張したのと同様に，本節では拡張ブール型モデルに対する索引語間関連度の組み入れを試みる。ここでは索引語間関連度の組み入れをおこなった拡張ブール型モデルの1つを最初に示す。

$$r_i(\mathrm{OR})=\left[\frac{\sum_k \delta_{ik}{}^p q_k{}^p}{\sum_k q_k{}^p}\right]^{1/p}$$

$$r_i(\mathrm{AND})=1-\left[\frac{\sum_k (1-\delta_{ik})^p q_k{}^p}{\sum_k q_k{}^p}\right]^{1/p} \qquad (8)$$

$$\delta_{ik}=\left[\frac{\sum_j x_{ij}{}^p y_{jk}{}^p}{\sum_{\substack{j\\ y_{jk}\neq 0}} x_{ij}{}^p}\right]^{1/p}$$

$$1\leqq p<\infty$$

まず上式において $p=1$ の場合をみてみると，

$$r_i(\mathrm{OR})=r_i(\mathrm{AND})=\frac{\sum_k \delta_{ik} q_k}{\sum_k q_k}\ ; \qquad \delta_{ik}=\frac{\sum_j x_{ij} y_{jk}}{\sum_{\substack{j\\ y_{jk}\neq 0}} x_{ij}}$$

となる。さらに各文献への索引語の付与にあたって重みづけがなされていない，すなわちすべての付与された索引語の重みを $x_{ij}=1$ と仮定すると，δ_{ik} は質問語 q_k に関連するすべての索引語間関連度の平均値となることがわかる。一方，$p\to\infty$ のときには δ_{ik} は次のものとなり，

$$\delta_{ik}=\lim_{p\to\infty}\left[\frac{\sum_j x_{ij}{}^p y_{jk}{}^p}{\sum_{\substack{j\\ y_{jk}\neq 0}} x_{ij}{}^p}\right]^{1/p}=\frac{\max_j x_{ij} y_{jk}}{\max_{\substack{j\\ y_{jk}\neq 0}} x_{ij}}$$

論理和演算をおこなう式は，

$$r_i(\mathrm{OR})=\lim_{p\to\infty}\left[\frac{\sum_k \delta_{ik}{}^p q_k{}^p}{\sum_k q_k{}^p}\right]^{1/p}=\frac{\max_k \delta_{ik} q_k}{\max_k q_k}$$

となる。ここですべての質問語の重みを $q_1=q_2=\cdots=q_n=1$ とし，かつ $y_{ik}\neq 0$ となる範囲内で $\max_j x_{ij}=1$ と仮定すれば，論理和演算の値は $r_i(\mathrm{OR})=\max_k \max_j x_{ij} y_{jk}$ により決定される。同様に $p\to\infty$ のとき，論理積演算をおこなう式は次のものとなる。

$$r_i(\mathrm{AND})=1-\frac{\max_k (1-\delta_{ik}) q_k}{\max_k q_k}$$

論理和演算式の場合と同様，すべての質問語の重みを $q_1=q_2=\cdots=q_n=1$ とし，かつ $y_{jk}\neq 0$ となる範囲内で $\max_j x_{ij}=1$ と仮定すれば，論理積演算の値は $r_i(\mathrm{AND})=\min_k \max_j x_{ij} y_{jk}$ により決定される。以上のことより，拡張ブール型モデルに索引語間関連度を組み入れ拡張した上記 (8) 式は拡張ブール型の特性をそのまま継承していることがわかる。

なお，上記の (8) 式は索引語間関連度を最も包括的に組み入れたものであるが，より限定した索引語間関連度のみを組み入れることも考えられよう。その一例としては各質問語 q_k に対する値が最大となるもののみを採用する，すなわち $\delta_{ik}=\max_j x_{ij} y_{jk}$ とするなど，その他多数のものが考えられよう。

IV. 実　　験

A. 実験用文献集合

コサイン関数モデルおよび拡張ブール型モデルに対して索引語間関連度を組み入れ拡張した，本稿で提案したモデルの有効性を検証するため，以下の実験用文献集合を作成し，検索実験をおこなった。図書館情報学を対象領域とする *Library and Information Science Abstracts* から比較的抄録データの長いレコードを無作為抽出し，抽出された 623 文献レコードをもって検索対象文献集合を作成した。実験には各レコード中の標題および抄録データのみを使用した。レコード当たりの平均語数は 102.0 語であり，標題のみの平均語数 9.0 語，抄録のみの平均語数 92.0 語との構成であった。索引語の抽出・付与，およびその重みの設定は以下の手順ですべて機械的に処理した。

1)　ストップワード 253 語の除去。同ストップワードは，van Rijsbergen が列挙した語[22]に若干のものを追加して作成した。

2)　複数形および過去・過去分詞形の語尾処理によるまとめあげ。当該処理後の索引語彙数は 5,209 語であった。

3)　各索引語の出現文献数による絞り込み。検索実験を容易にする目的で，各索引語の出現文献数が 6 以上 63 以下のもののみ抽出した。出現文献数が文献集合全体の 1～10% の索引語に限定したのは，Salton らによる実験結果に根拠を置いている[23]。最終的に残った索引語彙数は 1,007 語であり，文献レコード当たりの平均索引語付与数は 26.0 語，最大 58 語，最小 8 語となった。

4) 個々の文献レコードに付与された各索引語の重みづけ。2種類の重みづけを採用したが，どちらも a) 各文献内での当該索引語の出現頻度数，b) 文献集合全体内での出現文献数，c) 重みのとりうる範囲を区間 [0, 1] におさめるための正規化，の3要素をすべて含んだ手法とした。1つは次式を用いて重みを計算したものである。

$$x'_{ij}=f_{ij}\cdot\log(M/g_j)$$
$$x_{ij}=x'_{ij}/[\sum_j x'^2_{ij}]^{1/2} \qquad (9)$$

ここで，f_{ij}: 文献 d_i における索引語 t_j の出現頻度数，g_j: 文献集合全体内での索引語 t_j の出現する文献数，M: 総文献数，である。当該重みづけ法を用いた結果，最大 0.9058，最小 0.0359 の重みが付与され，0.7 以上の重みは 57 回，0.5 以上 0.7 未満の重みは 317 回，0.3 以上 0.5 未満の重みは 1,158 回，文献の索引語に対して付与された。他方の重みづけは，Salton らによる実験において比較的良好な結果をえたものであり，前記の式とはいわば正規化の方法が異なるものである[24]。

$$g'_j=\log(M/g_j)$$
$$x_{ij}=(g'_j/\max_j g'_j)\cdot(f_{ij}/\max_j f_{ij}) \qquad (10)$$

当該重みづけによると，最大 1.0，最小 0.0428 の重みが付与され，0.7 以上の重みは 521 回，0.5 以上 0.7 未満の重みは 989 回，0.3 以上 0.5 未満の重みは 2,385 回，索引語に対して付与された。

また，実験に用いる索引語間関連行列 Y，すなわち索引語間関連度 y_{jk} は単純な共出現値をもって代用することとし，具体的には以下の2式を採用した。1つは索引語の重みを考慮せず，文献を単位とした共出現値を求めるものであり，他は索引語の重みを考慮した式のうちの1つであり，質問-文献間の類似度計測に用いたコサイン関数式を共出現値を求めるためにそのまま用いたものである。

$$y_{jk}=g_{jk}/[g_j+g_k-g_{jk}] \qquad (11)$$

g_{jk}: 索引語 t_j と索引語 t_k が共に出現する文献数。

$$y_{jk}=\sum_i x_{ij}\cdot x_{ik}/[\sum_i x_{ij}^2\cdot\sum_i x_{ik}^2]^{1/2} \qquad (12)$$

(11)式を用いて計算した共出現値の分布は，$y_{jj}=1$ を除いて 0.7 以上の共出現値3組，0.5 以上 0.7 未満は4組，0.3 以上 0.5 未満は 17 組，0.1 以上 0.3 未満は 2,738 組であった。同様に (12) 式を用いて計算した共出現値の分布は，重みづけの手法に依存し，重みづけを (9) 式でおこなった場合は，$y_{jj}=1$ を除いて 0.7 以上の共出現値5組，0.5 以上 0.7 未満は 31 組，0.3 以上 0.5 未満は 697 組，0.1 以上 0.3 未満は 26,460 組であった。また重みづけを (10) 式でおこなった場合には，共出現値の分布は，$y_{jj}=1$ を除いて 0.7 以上の共出現値 3組，0.5 以上 0.7 未満は 107 組，0.3 以上 0.5 未満は 1,432 組，0.1 以上 0.3 未満は 28,940 組となった。なお，これ以後の検索実験では実験を容易にする目的で，共出現値が 0.1 以上の組のみを索引語間関連度として用いることにした。

検索実験に使用する検索質問およびそれに対応する適合文献群は，5組準備した。検索質問に用いることができる索引語集合による制約，および実験結果の再現率 (recall ratio; 呼出率ともいう) 0.1～1.0 の範囲で適合率 (precision ratio; 精度ともいう) が広がりをもつよう，設定した検索質問に比較して多少とも拡大解釈した適合文献集合を割り当てている点など，必ずしも精密な実験が実施できたわけではないことをお断りしておく。

B. 実験結果および考察

今回は本稿で提案したモデルのうち，索引語間関連度を組み入れ拡張したコサイン関数モデル，および同様に拡張した拡張ブール型モデルを中心に実験をおこなった。えられた実験結果は第 1～5 表に，再現率の各段階における平均適合率の値を用いて示してある。ただし，前節末に記した制約および採用した補間法などに適合率の値は大きく依存するため，その値は絶対値として意味があるわけではなく，あくまでも今回の実験で取りあげたモデル相互の性能比較をする上で目安となる値と考えるべきものである。

最初にコサイン関数モデルによる検索 [(1) 式に該当]，および索引語間関連度を組み入れ拡張した同モデルによる検索 [(3) 式に該当] をおこなった。結果は第1表に示した通りである。質問ベクトルおよび文献ベクトルの両者とも重みづけを適用しないコサイン関数モデルによる検索結果が表中の a 欄である。それに対して質問・文献とも重みづけは適用しないが，(11) 式により算出された索引語間の共出現値を組み入れた結果が b 欄に示してある。同様に，(9) 式または (10) 式により各文献ベクトル内の索引語の重みを初めに計算し，それらの重みに基づく共出現値を (12) 式を用いて求め，質問・文献ベクトルとも2値のまま，その共出現値をコサイン関数モデルに組み入れたのが，c および d 欄である。また，第1

第1表　コサイン関数モデルおよびその拡張形による検索実験結果

重みづけの適用なし

再現率	適合率 a	b	c	d
0.1	0.7750	0.7946	0.7946	0.7946
0.2	0.6586	0.7341	0.7341	0.7675
0.3	0.6172	0.6554	0.7334	0.7924
0.4	0.3878	0.5137	0.5831	0.5636
0.5	0.3859	0.4878	0.4906	0.5636
0.6	0.2662	0.3312	0.3620	0.4683
0.7	0.1645	0.1958	0.2076	0.3579
0.8	0.1048	0.1628	0.1506	0.1998
0.9	0.0867	0.0899	0.1012	0.1051
1.0	0.0317	0.0562	0.0535	0.0561

重みづけの適用あり

再現率	適合率 e	f	g	h
0.1	1.0000	1.0000	1.0000	1.0000
0.2	0.9286	0.9583	0.8472	0.8348
0.3	0.7634	0.7974	0.7722	0.7685
0.4	0.6567	0.6418	0.7333	0.7374
0.5	0.6131	0.6045	0.6592	0.6863
0.6	0.3666	0.4531	0.5594	0.5646
0.7	0.2112	0.3075	0.3727	0.3147
0.8	0.1393	0.2461	0.2367	0.2090
0.9	0.0914	0.0931	0.1197	0.1234
1.0	0.0317	0.0562	0.0548	0.0831

a，e：コサイン関数モデルによる検索
b，f：(11)式による共出現値を適用
c：(9)式の重みづけに基づく(12)式による共出現値を適用
d：(10)式の重みづけに基づく(12)式による共出現値を適用
g：(9)式による重みづけおよび(12)式による共出現値を適用
h：(10)式による重みづけおよび(12)式による共出現値を適用

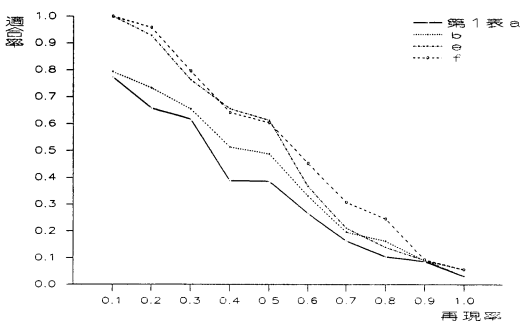
表中の下側は「重みづけの適用あり」としているが，ここでは文献ベクトルについてのみ (9)式または (10)式による重みづけをおこなっており，質問ベクトルについては重みづけを適用していない，すなわちすべての質問中の索引語の重みを1とした場合を指している。e欄は従来のコサイン関数モデルによる検索結果であり，f欄は(11)式を用いて算出された共出現値を組み入れた結果である。e，fともに文献ベクトルの重みづけに (9)式を用いたときと(10)式を用いたときでは等しい検索結果がえられており，そのため表中ではまとめた形で示してある。gおよびh欄については，表の下部に注記した通り，(9)式または (10)式による文献ベクトルの重みづけ，およびそれら重みに基づく索引語間の共出現値の両者を適用した結果である。

第1表からは以下の点が指摘できよう。

1)　質問・文献ベクトルとも重みづけを適用しないときに比較して，文献ベクトルについて重みづけを適用したときの方が，全般的に適合率は上昇している。この点は，表中の各欄について上側と下側とを比較することにより確認される。これはこれまでに報告されている多くの実験結果とも一致するものであり，併せて今回採用した重みづけ法の妥当性をも明らかにしているものと考えられる。今回の実験の範囲内では最大で 20% 近くの上昇がみられたところがある。

2)　本稿で提案した索引語間関連度の組み入れは，文献ベクトルに対する重みづけを適用しないときには，適合率を多少なりとも上昇させている。一方，重みづけを適用したときには適合率が上昇したとはいい難く，かえって減少しているところも見受けられる。これらの点は，a欄とb～d欄との比較，e欄とf～h欄との比較から確認される。最大で 10% 近くの適合率上昇がみられたところもあるが，必要とする計算量を勘案すると，再現率のすべての段階において期待されるほど検索効率が上昇しているとはいい難い結果である。ただし，今回の実験で用いた索引語間関連度は単純な共出現値であるため，より洗練された索引語間関連度が利用可能であれば，良好な検索結果がえられるものと期待される。また，文献ベクトルに対して重みづけを適用しないb～d欄の間で，あるいは文献ベクトルへの重みづけをおこなったf～h欄の間で，特に有意な差は見あたらないといってよかろう。a，b，e，f欄の結果について，グラフ化したものを第3図に示してある。

続いて，拡張ブール型モデルによる検索［(7)式に該当］，および同モデルに索引語間関連度を組み入れ拡張したものによる検索［(8)式に該当］をおこなった。拡



第3図　コサイン関数モデルおよびその拡張形による検索実験結果

第2表 拡張ブール型モデルによる検索実験結果

a：重みづけの適用なし

再現率	適合率 P=1～3	P=10
0.1	0.6833	0.6714
0.2	0.6117	0.6063
0.3	0.5193	0.5193
0.4	0.4218	0.4218
0.5	0.2618	0.2989
0.6	0.1954	0.2043
0.7	0.1563	0.1600
0.8	0.0977	0.0991
0.9	0.0808	0.0808
1.0	0.0317	0.0317

b：(9)式による重みづけの適用あり

再現率	適合率 P=1	P=2	P=3	P=10
0.1	0.9000	0.9375	0.9375	0.9375
0.2	0.8175	0.8889	0.8889	0.8472
0.3	0.7376	0.7320	0.7872	0.7287
0.4	0.6514	0.6148	0.6148	0.5935
0.5	0.5703	0.5960	0.5868	0.5577
0.6	0.3521	0.3243	0.3218	0.2900
0.7	0.2626	0.2585	0.2585	0.1533
0.8	0.1117	0.1333	0.1393	0.1294
0.9	0.0832	0.0839	0.0853	0.0814
1.0	0.0317	0.0317	0.0317	0.0317

c：(10)式による重みづけの適用あり

再現率	適合率 P=1	P=2	P=3	P=10
0.1	1.0000	1.0000	0.9375	0.9000
0.2	0.9015	0.9086	0.8281	0.7865
0.3	0.6709	0.8333	0.8229	0.7333
0.4	0.6063	0.5772	0.5723	0.6130
0.5	0.5356	0.5591	0.5591	0.5303
0.6	0.3252	0.3254	0.3217	0.2624
0.7	0.2512	0.2361	0.2335	0.2523
0.8	0.1117	0.1248	0.1333	0.1128
0.9	0.0832	0.0846	0.0867	0.0924
1.0	0.0317	0.0317	0.0317	0.0317

第3表 拡張ブール型モデルの拡張形による検索実験結果（重みづけの適用なし）

a：(11)式による共出現値を適用

再現率	適合率 P=1	P=2	P=3	P=10
0.1	0.5877	0.5717	0.5717	0.6121
0.2	0.3589	0.4792	0.5312	0.4253
0.3	0.3479	0.3820	0.3889	0.3097
0.4	0.3158	0.3282	0.3483	0.3363
0.5	0.2675	0.3378	0.3253	0.2244
0.6	0.2395	0.2159	0.2131	0.2091
0.7	0.1762	0.1704	0.1711	0.1493
0.8	0.1551	0.1486	0.1387	0.0992
0.9	0.0574	0.0819	0.0819	0.0594
1.0	0.0393	0.0393	0.0392	0.0395

b：(9)式の重みづけに基づく(12)式による共出現値を適用

再現率	適合率 P=1	P=2	P=3	P=10
0.1	0.3139	0.5341	0.5717	0.4883
0.2	0.2044	0.2515	0.2906	0.2908
0.3	0.1942	0.2273	0.3178	0.2821
0.4	0.1731	0.2113	0.2947	0.2872
0.5	0.1721	0.1779	0.1712	0.1865
0.6	0.1500	0.1468	0.1774	0.1627
0.7	0.1297	0.1324	0.1562	0.1533
0.8	0.1197	0.1314	0.1357	0.1310
0.9	0.0641	0.0715	0.0904	0.0589
1.0	0.0477	0.0598	0.0558	0.0439

c：(10)式の重みづけに基づく(12)式による共出現値を適用

再現率	適合率 P=1	P=2	P=3	P=10
0.1	0.2576	0.4770	0.6042	0.4542
0.2	0.2393	0.3366	0.3259	0.3279
0.3	0.1987	0.2929	0.3452	0.3031
0.4	0.1653	0.2490	0.3148	0.2904
0.5	0.1591	0.1809	0.1706	0.2316
0.6	0.1493	0.1580	0.1673	0.1674
0.7	0.1086	0.1424	0.1470	0.1541
0.8	0.0982	0.1112	0.1288	0.1267
0.9	0.0659	0.0920	0.1003	0.0762
1.0	0.0438	0.0486	0.0478	0.0439

張ブール型モデルによる検索結果を第2表に，さらに索引語間関連度を組み入れ拡張した同モデルによる検索結果を第3～5表に示してある。重みづけの適用の有無とはコサイン関数モデルの実験のときと同様，文献ベクトルに対する重みづけの有無を意味しており，質問ベクトル中の索引語に対する重みづけはおこなっていない。第3表には重みづけを適用しないで，索引語間関連度のみを適用した場合の結果を示している。第3表中のbおよびc表の意味は，コサイン関数モデルの実験において示したもの(第1表cおよびd)と同じである。また，第4表は拡張ブール型モデルの拡張式(8)式を用いたときの検索結果を示したものであり，第5表は(8)式中のδ_{ik}を各検索語q_kに対する値が最大となるもの$\delta_{ik}=\max_j x_{ij}y_{jk}$に置き換えたモデルを用いたときの検索結果であり，組み入れるべき索引語間関連度を限定した場合の一例とし

第4表　拡張ブール型モデルの拡張形による検索実験結果（重みづけの適用あり 1.）

a：(11)式による共出現値を適用

再現率	適合率			
	P=1	P=2	P=3	P=10
0.1	0.7188	0.8750	0.7917	0.9167
0.2	0.4251	0.5963	0.5872	0.5789
0.3	0.3734	0.4456	0.4456	0.3652
0.4	0.3654	0.3494	0.3660	0.3352
0.5	0.3203	0.3457	0.3639	0.2959
0.6	0.2820	0.2730	0.2971	0.2868
0.7	0.2167	0.1886	0.1982	0.1984
0.8	0.1661	0.1681	0.1614	0.1044
0.9	0.0566	0.0571	0.0587	0.0595
1.0	0.0393	0.0393	0.0393	0.0399

b：(9)式による重みづけおよび(12)式による共出現値を適用

再現率	適合率			
	P=1	P=2	P=3	P=10
0.1	0.4327	0.5917	0.7135	0.6436
0.2	0.3013	0.4966	0.6167	0.4454
0.3	0.2958	0.4330	0.5473	0.4294
0.4	0.2298	0.2907	0.2796	0.3315
0.5	0.2029	0.2807	0.2786	0.2248
0.6	0.1986	0.2447	0.2278	0.1922
0.7	0.1802	0.1878	0.1911	0.1803
0.8	0.1153	0.1137	0.1323	0.0955
0.9	0.0588	0.0641	0.0643	0.0572
1.0	0.0487	0.0498	0.0499	0.0418

c：(10)式による重みづけおよび(12)式による共出現値を適用

再現率	適合率			
	P=1	P=2	P=3	P=10
0.1	0.4833	0.7244	0.8077	0.7316
0.2	0.4077	0.5750	0.5940	0.4551
0.3	0.3962	0.4830	0.5877	0.4631
0.4	0.2242	0.3180	0.3566	0.3769
0.5	0.1920	0.2944	0.2404	0.2640
0.6	0.1888	0.2457	0.2227	0.1555
0.7	0.1586	0.2030	0.1855	0.1468
0.8	0.1203	0.1420	0.1527	0.1415
0.9	0.0678	0.0717	0.0641	0.0537
1.0	0.0443	0.0484	0.0478	0.0424

第5表　拡張ブール型モデルの拡張形による検索実験結果（重みづけの適用あり 2.）

a：(10)式による重みづけおよび(11)式による共出現値を適用

再現率	適合率			
	P=1	P=2	P=3	P=10
0.1	1.0000	1.0000	0.9357	0.9000
0.2	0.9083	0.9583	0.8281	0.6793
0.3	0.6612	0.8333	0.8229	0.6396
0.4	0.5737	0.5825	0.6024	0.5681
0.5	0.5462	0.5641	0.5659	0.5003
0.6	0.4080	0.4196	0.4131	0.3586
0.7	0.3088	0.2620	0.2572	0.2556
0.8	0.2288	0.2245	0.2251	0.1385
0.9	0.0856	0.0808	0.0819	0.0846
1.0	0.0407	0.0415	0.0413	0.0412

b：(9)式による重みづけおよび(12)式による共出現値を適用

再現率	適合率			
	P=1	P=2	P=3	P=10
0.1	0.8750	0.9375	0.9375	0.9375
0.2	0.8220	0.7778	0.7951	0.7639
0.3	0.6411	0.6896	0.6718	0.5325
0.4	0.6326	0.6134	0.6148	0.4524
0.5	0.5883	0.5989	0.5922	0.4321
0.6	0.4975	0.5445	0.5221	0.3936
0.7	0.3029	0.2695	0.2768	0.1552
0.8	0.2387	0.2491	0.2230	0.1347
0.9	0.1232	0.1374	0.1411	0.1224
1.0	0.0631	0.0679	0.0723	0.0513

c：(10)式による重みづけおよび(12)式による共出現値を適用

再現率	適合率			
	P=1	P=2	P=3	P=10
0.1	0.9375	1.0000	0.9375	0.8167
0.2	0.8869	0.9167	0.7976	0.7067
0.3	0.6344	0.8021	0.7917	0.4950
0.4	0.5806	0.5625	0.6307	0.4137
0.5	0.5668	0.5567	0.5723	0.3829
0.6	0.4586	0.4742	0.4766	0.2837
0.7	0.3071	0.3074	0.2760	0.2462
0.8	0.2812	0.2435	0.2298	0.1847
0.9	0.1125	0.1085	0.1095	0.0983
1.0	0.0711	0.0763	0.0791	0.0595

て示してある。拡張ブール型モデルにおいては論理演算子の強さを制御するパラメータ p の値に応じて検索結果が変化するため，実験ではそれぞれの場合にパラメータ p の値を 1, 2, 3, 10 に設定し検索をおこなった。

これらの実験結果は以下の点に要約されよう。

1)　拡張ブール型モデルによる検索においても，文献ベクトルへの重みづけは有効であり，パラメータ p の値に依存せず，適合率を上昇させている。この点は第2表中のａ表と下2つのｂ，ｃ表とを比較することによりわかる。また，(9) 式による重みづけと (10) 式による重みづけでは検索結果に有意な差がみられない点もコサイン関数モデルの実験結果と共通している。

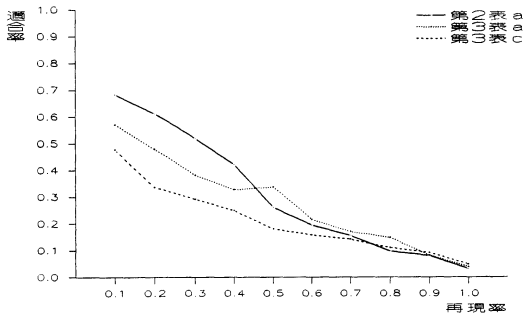
2)　拡張ブール型モデルにおける検索結果は，設定したパラメータ p の範囲内では大きな変化は現れていない

(第2表参照)。文献ベクトルへの重みづけを適用しないときには，$p=1\sim3$ の範囲では等しい結果がえられている。また，重みづけの適用の有無に関わらず，$p=10$ のときにも大きな適合率減少は見受けられない。全般的にいって，Saltonらが実験によって導き出した結論，すなわちパラメータ p は重みづけを適用していないときには $p=2\sim5$ の範囲内で，重みづけを適用しているときには $p=1\sim2$ の範囲内で最適な結果がえられるとした結論[21]とも大きく矛盾するものではないといえよう。

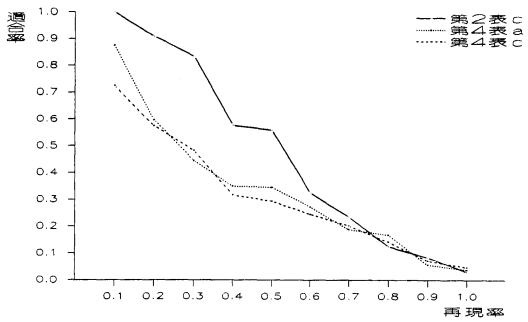
3) 拡張ブール型モデルによる検索結果をコサイン関数モデルのものと比較すると，重みづけを適用していない場合および (9)式または (10)式による重みづけを適用した場合の双方において，若干の適合率減少がみられるところがあるが，全体的には大きな差はないといってよかろう。特に $p=1$ のときの拡張ブール型モデルは論理演算子の効力がなくなり，コサイン関数モデルと正規化処理の点で異なるだけとなり，両者の検索結果の相違はこの正規化処理の相違に帰着することとなる (第1表a，e欄と第2表中の $p=1$ の場合を比較)。Salton らの実験において，コサイン関数モデルと比較した拡張ブール型モデルによる検索結果は，文献ベクトルへの重みづけをおこなわない場合には実験対象集合により適合率の上昇がみられたり，減少がみられたりして，一定していない。一方，同じ実験で文献ベクトルへの重みづけを実施した場合には，実験に用いた4つすべての実験対象集合において，$p=1\sim2$ の範囲内では拡張ブール型モデルによる適合率上昇が報告されているが[21]，今回の実験ではそれに合致する検索結果はえられなかった。

4) 拡張ブール型モデルに比較して本稿で提案したその拡張モデル [(8)式に該当] は，文献ベクトルへの重みづけの有無に関わりなく，今回の実験の範囲内では大きく適合率を減少させている。この点は第2表と第3，4表とを比べることで明らかである。特に再現率 0.1～0.4 の各段階に対応する適合率が大きく減少しているといえよう。また，文献ベクトルへの重みづけを適用したときにも，適用しないときにも，(12)式により算出された索引語間共出現値を組み入れた場合には極端に適合率が減少している。拡張ブール型モデルとその拡張モデルによる検索結果の比較を，重みづけを適用していない場合の比較を第4図に，重みづけを適用した場合の比較を第5図に，それぞれグラフ化して示してある。

5) 本稿で提案した形で拡張ブール型モデルに索引語間共出現値を組み入れた場合においても，文献ベクトルへの重みづけの適用は重みづけを適用しないときに比べてパラメータ p のすべての値に関して，適合率の上昇を示している。これは第3表と第4表とを比べることより理解される。この点より，重みづけ自体は拡張ブール型モデルの拡張型においても有効であることが確認される。



第4図 拡張ブール型モデルおよびその拡張形による検索実験結果 (重みづけの適用なし，p=2)



第5図 拡張ブール型モデルおよびその拡張形による検索実験結果 (重みづけの適用あり，p=2)

6) 拡張ブール型モデルに組み入れるべき索引語間共出現値を限定した場合の検索結果は，限定せずすべての利用可能な共出現値 (実際には実験の都合上，0.1 以上の共出現値をもつもの) を組み入れた場合に比較して，適合率の大きな改善がみられる。これは第5表と第4表との各欄を比較することで確認される。そして拡張ブール型モデルそのものによる検索結果と大差のない適合率をえていることも，第2表中のbおよびc表と比べることより確認できる。これより今回の実験の範囲内では，$\delta_{ik}=\max_j x_{ij}y_{jk}=x_{ik}$ となる場合が大半であったことが推測される。また，第5表から読み取れるように，組み入れるべき共出現値を限定した場合にも，(11)式により算出された共出現値を用いたときの方が(12)式による共

出現値を用いたときに比べて，概して適合率は良好である。

以上の諸点をまとめると，今回の実験ではコサイン関数モデルおよび拡張ブール型モデルの両者に索引語間共出現値を組み入れただけでは，有効な検索効率の上昇をえるまでには至らなかったといえよう。特に拡張ブール型モデルに共出現値を組み入れた場合には，大きく効率が低下してしまうことが観察された。その理由として，拡張ブール型モデルの拡張方法自体が適切でないことと，使用した索引語間関連度が単純な共出現値であることの2点が考えられるが，どちらがより基底的な原因として実験結果を左右しているかは今回の実験からは結論づけることが困難である。

V. おわりに

本稿では，索引語間の関連性が関連度として数値で与えられていることを前提にして，質問に対する各文献の類似度計測にその関連度を反映させることができるよう，既存の検索モデルの拡張を試みた。具体的には，コサイン関数モデルおよび拡張ブール型モデルに対して索引語間関連度を組み込んだ拡張モデルを提案した。これら拡張モデルを用いた検索実験ではその有効性を示しえなかったが，モデル化の点では当初の目的を十分果たしえたものと考えている。今後は，より精密な実験集合において，本稿で提案した拡張モデルが果たして今回の実験と同様な結果をえるかどうかを確認することが必要であるが，併せて本稿では取りあげなかった他の検索モデル（ファジィ型，集合論型，確率型など）との接合・統合をめざしていっそうのモデル化を進めることが課題となろう。

日頃より御指導をたまわり，本論文をまとめるに当たっても助言していただいた図書館情報大学　黒岩高明教授に深謝いたします。また，本研究の方向づけをしていただいた同大学　桜井宣隆教授，伊藤哲郎助教授に深謝いたします。

1) Salton, G. Mathematics and information retrieval. Journal of Documentation. Vol. 35, No. 1, p. 1-29 (1979)
2) van Rijsbergen, C. J. A theoretical basis for the use of cooccurrence data in information retrieval. Journal of Documentation. Vol. 33, No. 2, p. 106-119 (1977)
3) van Rijsbergen, C. J. "6. Probabilistic retrieval". Information Retrieval. 2nd ed. London, Butterworths, 1979, p. 111-143.
4) Salton, G.; Buckley, C.; Yu, C. T. "An evaluation of term dependence models in information retrieval". Research and Development in Information Retrieval. Salton, G.; Schneider, H. J. eds. New York, Springer, 1983, p. 151-173. (Lecture Notes in Computer Science, Vol. 146).
5) Yu, C. T.; Buckley, C.; Lam, K.; Salton, G. A generalized term dependence model in information retrieval. Information Technology: Research and Development. Vol. 2, No. 4, p. 129-154 (1983)
6) 谷口祥一．決定理論型情報検索モデルのファジィ確率による拡張．図書館情報大学研究報告．Vol. 8, No. 1, p. 37-47 (1989)
7) Ito, T.; Kodama, Y.; Toyoda, J. A similarity measure between patterns with nonindependent attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. PAMI-6, No. 1, p. 111-115 (1984)
8) Salton, G.; McGill, M. J. Introduction to Modern Information Retrieval. New York, McGraw-Hill, 1983, 448 p.
9) Wong, S. K. M.; Raghavan, V. V. "Vector space model of information retrieval: a reevaluation". Research and Development in Information Retrieval. van Rijsbergen, C. J. ed. Cambridge, Cambridge University Press, 1984, p. 167-185.
10) Raghavan, V. V.; Wong, S. K. M. A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science. Vol. 37, No. 5, p. 279-287 (1986)
11) Wong, S. K. M.; Ziarko, W.; Wong, P. C. N. Generalized vector space model in information retrieval. Proceedings of the 8th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Montreal, Canada, 1985-06. New York, ACM, 1985, p. 18-25.
12) Wong, S. K. M.; Ziarko, W.; Raghavan, V. V.; Wong, P. C. N. On extending the vector space model for Boolean query processing. [Proceedings of the] ACM Conference on Research and Development in Information Retrieval. Rabitti, F. ed. Pisa, Italy, 1986-09. [s.n.], 1986, p. 175-185.
13) Wong, S. K. M.; Ziarko, W.; Raghavan, V. V.; Wong, P. C. N. On modeling of information

retrieval concepts in vector spaces. ACM Transactions on Database Systems. Vol. 12, No. 2, p. 299-321 (1987)

14) Wong, S. K. M.; Ziarko, W.; Raghavan, V. V.; Wong, P. C. N. Extended Boolean query processing in the generalized vector space model. Information Systems. Vol. 14, No. 1, p. 47-63 (1989)

15) Deerwester, S. et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science. Vol. 41, No. 6, p. 391-407 (1990)

16) Giuliano, V. E.; Jones, P. E. "Linear associative information retrieval". Vistas in Information Handling. Howerton, P. W. ed. Washington, Spartan, 1963, p. 30-54.

17) Heaps, H. S. "12. Automatic question modification". Information Retrieval: Computational and Theoretical Aspects. New York, Academic Press, 1978, p. 293-308.

18) 水本雅晴. ファジィ理論とその応用. 東京, サイエンス社, 1988, 359 p.

19) Bezdek, J. C.; Biswas, G.; Li-Ya Huang. Transitive closures of fuzzy thesauri for information retrieval systems. International Journal of Man-Machine Studies. Vol. 25, No. 3, p. 343-356 (1986)

20) Tamura, S.; Higuchi, S.; Tanaka, K. Pattern classification based on fuzzy relations. IEEE Transactions on Systems, Man, and Cybernetics. Vol. SMC-1, No. 1, p. 61-66 (1971)

21) Salton, G.; Fox, E. A.; Wu, H. Extended Boolean information retrieval. Communications of the ACM. Vol. 26, No. 12, p. 1022-1036 (1983)

22) van Rijsbergen, C. J. Information Retrieval. 2nd ed. London, Butterworths, 1979, p. 18-19.

23) Salton, G.; Yang, C. S.; Yu, C. T. A theory of term importance in automatic text analysis. Journal of the American Society for Information Science. Vol. 26, No. 1, p. 33-44 (1975)

24) Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing & Management. Vol. 24, No. 5, p. 513-523 (1988)