Library and Information Science No. 30 1992

80 年代における情報検索モデル研究の展開：文献レビュー

Progress of the research of information retrieval models in the 80's : a critical review

谷　口　祥　一
Shoichi Taniguchi

Résumé

In the research field of information retrieval(IR) models, there were many activities of research and development in the 80's. This paper reviews articles primarily concerned with IR models, especially which take quantitative approaches, for the purpose of tracing the progress in the research field critically. The number of those articles are too many to review comprehensively in this paper, so the author chose a subset of such articles written in Japanese or English.

This paper, first of all, describes some basic problems of conventional IR systems based on Boolean model. Secondly, it reviews several specific models in the following order: vector space model, fuzzy set model, probabilistic model, integrated model, and other retrieval models. Finally, it concludes that the progress of the field in the 80's can be summarized as continuation and refinement of the results of the research activities in the 60's and 70's, and that we cannot get any results which make a radical change in the framework of the research field. However, this paper also concludes that many useful, interesting results were produced in the same decade, and it suggests directions for further research which will bring fruitful results in the next decade.

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.
1992 年 5 月 10 日受付

I. はじめに

情報検索の領域においては，80年代においてもそれ以前に引き続き，さまざまな検索手法等の研究開発が行われてきた。それら多種多様な検索手法に対してはある範囲内で類型化が可能であり，まとめられた各類に属する検索手法を，ここではより抽象化して検索モデル (retrieval model) と呼ぶ。本稿の目的は，それら検索モデルの中でも特に数量的なアプローチを採用した検索モデルに限定して，80年代における研究の展開を文献のレビューを通して跡づけることにある。

本稿で取りあげる文献は，原則として 80 年代 (1980年～1990 年) に刊行されたものとするが，一部それ以後に刊行されたものも含めている。対象となりうる文献は相当な量にのぼることが容易に確認されるため，本稿ではそれらのうち，筆者が個人的に選択したもののみ取りあげるにとどまる (言語的には日本語および英語文献に限定される)[1]。また，取りあげた各論文の批判的評価を意図してはいるが，筆者の力量の限界もあり，単なる紹介にとどまるものも多い点，お断りしておかなければならない。

検索モデル研究を対象とし，かつ対象期間も本稿と一部重なる既存の文献レビューに，Belkin ら (Belkin *et al.*, 1987), Bookstein (Bookstein, 1985), および Kraft (Kraft, 1985) のものがある。特に前2者を組み合わせることで 1987 年当時までの広範囲にわたる論文をレビュー (少なくともリストアップ) することができる。Belkin らのものは検索モデルのある整理を示しており，その点でも有用なものである。また，限定された範囲の検索モデルを対象としたレビューも他にいくつか存在しており，それらについては適宜該当する箇所で紹介することとしたい。検索モデルを含めて情報検索理論に関しては 80 年代にいくつかの解説書が刊行されており (伊藤, 1986; Salton, 1989; Salton & McGill, 1983; 谷口, 1993), これらを上記レビューを補完するものとして活用することも可能であろう。

II. ブール型モデルの問題点

商用ベースで稼働しているオンライン情報検索システムはすべてブール型モデル (Boolean model) に従って構築されているものと捉えることができる。それはシステム内部の処理効率の点 (転置ファイルの活用等を含む), および2値論理 (または集合論) に依拠した処理過程の明快さの点で強力なシステムを実現している。しかしながら，ブール型モデルに対してはいくつかの基本的な問題点が繰り返し指摘されてきており，Bookstein (Bookstein, 1985) や Belkin ら (Belkin *et al.*, 1987) に依拠して最大公約数的にまとめると，次のものがその問題点として挙げられる。

1) 検索質問で指定された条件を部分的に満たさない文献は，たとえ有用であろうとも検索されない。検索条件を完全に満たす (完全に一致する) 文献のみ検索される結果となる。

2) 検索結果を検索質問に対する推定された適合度順に出力することができない。検索されるものとそれ以外のものとに分かれるにすぎない。

3) 検索質問あるいは文献の索引語表現において，それに含まれる索引語間に重要度の差異を設けることがで

きない。

4) 検索者の検索要求を，ブール式をなす検索質問に的確かつ十分なものとして変換できるとは限らない。

これらの問題点を解決する，もしくはその制約を緩めることを目的として各種の検索モデルが研究されている。

III. ベクトル空間型モデル

A. 類似測度の特性

ベクトル空間 (vector space) 型モデルとは，ベクトル表現された文献と，同じくベクトル表現された検索質問との似ている程度としての類似度 (similarity) を求め，その値を RSV (retrieval status value) とするものである。元来，同モデルは索引語（またはそれに対応する概念）ベクトルが基底をなすベクトル空間を想定し，その空間内に個々の文献および検索質問を位置づけ，その類似度を測定することを意味している。

最も基本的な類似測度は文献・質問両ベクトルの内積(スカラー積)を求めるものであり，さまざまな方法でこの内積を正規化した多数の類似測度が知られている。Noreault ら (Noreault *et al.*, 1981) は，24 の類似測度について同一環境下での評価実験を行い，その結果について報告している。取りあげられた 24 の測度すべてを狭義のベクトル型に属するものと見なすには困難があり，集合論型等，他の検索モデルに属させるべき測度も含まれている。しかしながら，それらはすべて論理演算子を用いず，文献および質問ともベクトル表現されているものに対する類似測度であり，広義に解釈したときのベクトル型に属するものと見なすことも可能であろう。同論文で報告されている実験結果のみに基づき個々の類似測度の優劣もしくは性能を即断するのは危険といえる。むしろ，報告された実験結果は1つの目安にすぎないと考えることが妥当であろう。この点は他の論文における実験結果にも等しく当てはまるものと個人的には考えている。上記論文は多数の測度を整理して示している点のみにおいても有用である。

このような実験結果に基づき類似測度の優劣もしくは性能を総合的に評価するよりも，むしろ個々の測度の特性（振る舞い）をいわば解析的に明らかにすることがさらに重要と思われる。この点で興味深い試みが Jonesら (Jones *et al.*, 1987) により示されている。それは，ある質問ベクトルを設定し，それに対して等しい類似度を与える文献ベクトル群の終点の軌跡を幾何的に示す試みである[2]。特に，1) 質問・文献両ベクトル間の角度の変化，2) 文献ベクトルの長さ（ノルム）の変化，3) 文献ベクトル内の個々の成分値の変化，のそれぞれに対応する類似度の変化，さらには 4) 極端に大きな値をとる単一成分の類似度への影響，および 5) 類似度の取りうる値の範囲，の5つに焦点を当て，個々の類似測度の特性および類似測度間の差異を示している。類似測度の特性を多面的かつ分析的に解明することの意義は，これまであまりそれに類する試みが行われてこなかっただけに，いくら強調してもしすぎることはないであろう。今後同種のアプローチが広く採用されることを個人的には期待している。

B. 索引語間の関連性の組み入れ

索引語間に存在する関連性を何らかの形で検索処理の過程に組み入れることは一般に検索効率をあげる上で有益と考えられており，ブール型検索システムにおけるシソーラスの活用等を含めて，多様な試みがこれまでに行われてきた。なお，索引語間関連性の組み入れは，検索質問の拡張処理と一致する場合もある点に留意されたい。

ベクトル空間型モデルに対する索引語間関連性組み入れの再検討は，前節で述べた類似測度を適用するベクトル空間の基底が直交基底であることへの自覚，すなわちこれまで無条件に直交基底を前提としてきたことへの自覚に始まった。索引語ベクトルを無条件に正規直交基底としてきたことへの指摘およびそれに対する再検討は Raghavan ら (Raghavan & Wong, 1986) によって最初に行われた。これを出発点として，その後いくつかの索引語間関連性を組み入れたモデルが提案されている。

その1つは一般化ベクトル空間 (generalized vector space) モデルと呼ばれるものである (Wong, Ziarko, Raghavan & Wong, 1987)。同モデルは索引語ベクトルを基底ベクトルとはせず，別に基本概念に当たるものを想定し，それらに対応し，かつ直交する基底ベクトルを新たに設定しようとするものである。そしてこれら新たに設定された基底ベクトルにより生成される空間上に，改めて個々の索引語ベクトル，ひいては質問および文献ベクトルを位置づけ，類似度を測定するものである。このような考え方において課題となるのは基本概念の設定法に関してであるが，同モデルにおいては検索対象となる文献集合自体における索引語の付与（出現）パターンそのものを基本概念に該当するものと見なしている。筆者にはその処理手順は了解できるが，それにより多様な

側面を有する索引語間関連性のいかなる部分が反映されたといえるのか，換言すれば残された部分はいかなるものであるのか，もしくは必要にして十分な索引語間関連性が組み入れられたといいうるのか等について疑問が残る。

2番目のものとして，singular value decomposition (SVD) モデルがある (Deerwester *et al.*, 1990)。文献への索引づけの結果を表す文献—索引語行列（換言すれば文献ベクトルを並べたもの）に対して SVD，すなわち固有値解析の一種を施し，当該行列を singular value (固有値の平方根に該当) からなる対角行列とその要素に対応した singular vector を並べた2つの直交行列との計3つの行列の積で効率的に近似するものである。singular value は，前記一般化ベクトル空間モデルの表現を借りれば，基本概念に該当するものと考えられ，ベクトル空間に即して考えればこれら基本概念による直交座標系が新たに設定されることになる。与えられた文献—索引語行列が有する階数を縮小する過程において，因子分析と同様の意味で索引語間の関連性が組み入れられることとなる。SVD を実行する計算量の点を除けば，当該モデルは今後の展開が期待できるものの1つといえよう。なお，SVD モデルと従来のベクトル空間モデルとを組み合わせることにより，さらに効率のよい検索結果がえられる点が実験結果として報告されている (Lochbaum *et al.*, 1989)。

上記2つのモデルが有する利点の1つは，新たに設定した基本概念に対応する1次独立なベクトルにより，検索処理を行う直交座標系が一意に決まり，そこにおいては索引語および質問・文献ベクトルの表現から冗長性を排除できる点にあろう。それに対してそれら冗長性をそのまま許容することを意図したモデルがある。それは従来のベクトル空間モデルにおいて基底ベクトルとみなした索引語ベクトルは無条件に直交系をなすとは考えず，索引語間の関連度に応じた角度でそれらが設定される斜交系に拡張して考えるものである（谷口，1990)。斜交座標系は，すべての索引語間に関連性がない場合に該当する直交系をその特殊例として含むことになる。これにより，与えられた索引語間関連度を反映した検索処理が可能となるが，これはあくまでも索引語の対（2つ組）に関してのみであり，3つ組以上の索引語間の関連度を直接反映させる方策・条件等は解明されていない。ただし，その代替案として2つの索引語間の関連度に，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次結合の関連度等，高次の関連度を加えたものとする方策は示されている。

C. その他

適合フィードバック (relevance feedback) 処理とは，当初の質問によりえられた検索結果を受けて，よりよい結果をえる質問にと修正を繰り返していくフィードバック処理を指す。そこでは，各段階で検索された文献集合に対して，個々の文献毎に検索者による適合・不適合の判定がなされるものとされている。このような適合フィードバックのベクトル空間型モデルへの適用は 60 年代後半には既に試みられており，各種の提案および実験結果がこれまでに報告されている。管見によれば，これに関わる 80 年代の成果として Salton ら (Salton & Buckley, 1990) によるものが挙げられる。同論文では主に確率型モデルとの性能比較を実験により検証しており，その結果によれば，ベクトル空間型のものは概ね確率型よりも良好な結果を示している。その限られた実験結果から両者の適否を軽々に論じることはできないが，当該論文が適合フィードバック処理の適切なまとめとなっている点をむしろ記しておきたい。

ベクトル型に密接に関連するものとして文献のクラスタリング，およびそのクラスタ化されたファイルに対する検索処理がある。これらの事柄に関する 80 年代の研究成果はその数が多く，また筆者の手に余る領域であるため，ここでは比較的新しい Willett (Willett, 1988) の包括的なレビューのみ挙げるにとどめる。

IV. ファジィ型モデル

A. ファジィ集合に基づくモデル

ファジィ集合 (fuzzy set) とはそれに属するか否かが明確でない対象物の集まりをも集合として扱うものであり，集合の各要素は当該集合に属する度合いを示すメンバーシップ関数値をもって表される。これにより，従来の2値表現が多値表現に拡張されたことになり，情報検索への適用を考えると，個々の文献への索引語の付与およびその重みづけ，質問中の検索語への重みづけは，いずれもファジィ集合として表現可能となる。さらに，ファジィ集合においては集合間の演算（和，積，補）がこれまでに多数定義されており，これらを検索質問中で指定された論理演算子（論理和，論理積，否定）の処理にそのまま適用することが可能となる。以上により，2値論

理に対応するブール型モデルをその特殊ケースとして含む，重みづけを用いた検索システムが実現されることになる。この点を丹念に論じたものに Radecki (Radecki, 1983b) のものがある。

ファジィ型モデルにおける1つの議論に，これら多数提案されている集合演算のうち，いずれを情報検索に最適なものとして選択すべきかとの問題がある。この議論の展開の中で，単一の演算子ではあるが論理和と論理積とを組み合わせた中間的な値をとる演算子，すなわち補償演算子 (compensatory operator) の導入が 70 年代末に提案されている。この補償演算子の有効性を Paice (Paice, 1984) は実験によって確認している。さらには，3つ以上の検索語が連結されているときに，ある条件下で結合律を擬似的に満足するような補償演算子の適用方法を併せて提案しているが，その提案の妥当性を十分に示しているとはいい難い。なお，補償演算子の中には情報検索への適用をこれまでに試みられていないものもある点に注意すべきである。また，演算結果が max 演算と min 演算の間の値となる平均演算子 (averaging operator; mean) も検索モデルへ適用が図られており (具体的な事例は統合型モデルの章で扱う)，両者の特性比較等も今後の課題となろう。

Koll ら (Koll *et al.*, 1990) は，個々の文献に対する検索者の適合性評価値とシステムの算出した RSV との相関を調べている。彼ら自身はファジィ型モデルと確率型モデルの比較として枠組みを設定しているが，実質的にはすべてファジィ型に属する max, min, 代数和，代数積の各演算間での比較と捉えることがより適切であろう。個々の索引語に対する適合性評価値がいかに組み合わされて文献に対する最終的な適合性評価値となるかに焦点を絞った実験であるが，その実験結果では論理積演算子については min 演算および代数積演算とも検索者の評価値よりも過小評価となる点が示されている。これはある意味では平均演算子もしくは補償演算子導入の必要性を示した実験結果と解釈できよう。ただし，このような実験のみに依拠して各種のファジィ演算の中から最適なものを選択しようとすることは適切さを欠くであろう。むしろ同論文は検索者の評価行動をモデル化する試みとみるべきであり，その点こそが重要と考える。ただし，モデル化という点では非常に単純なモデルを示したにすぎず，今後の展開を期待しなければならない。

ファジィ型モデルのもう1つの問題は，検索語への重みづけに関わるものである。大きく分ければ，検索要求に対する個々の索引語の相対的な重要度と捉えるものと個々の索引語に対するしきい値 (threshold value) と捉えるものとの2つがある。前者の立場からは検索語に付与される重みは relevance weight と呼ばれる。両者は基本的な部分で異なるものであり，一般には両者の接合は困難と考えられる。しかしながら，後に触れるように，ある特定のモデルについては両者の結果が類似することを示す報告もある。Buell (Buell, 1981) は，その時点までの重みの扱いに対する両者の考えを整理して示している。

この重みの扱いはファジィ型モデルが備えることが望ましいとされている条件の1つ，「分離要件 (separability)」とも密接に関係する。それは検索式中の個々の検索語に対するある文献の評価は，検索語間で互いに独立して行われねばならないとするものであり，結合律とも関わるものである。Buell (Buell, 1982) はこの問題を数学的な構造としての束において検討している。同時に，否定演算子に適用される補集合演算の問題についても言及しているが，数学的な特性よりも情報検索という現実の問題への適合性をより重視すべきとの結論を示している。

relevance weight の解釈にも多様なものがあり，従来は文献に付与された索引語の重みと検索語の重みを乗じることで個々の検索語に対する評価を行っていた。それに対して，当該重みは検索者の求める理想的な文献に付与された索引語の重みを表すと考え，従って検索者の提出した検索式は全体として理想的な文献を表現したものとの解釈をとるものがある。これより，指定された検索語の重みに近い値をもつ索引語を有する文献ほど高い評価値を与えることが考えられ，これをファジィ制約として定式化することが可能となる (Bordogna *et al.*, 1991)。このような考え方は以前のものに比べて異色であり，今後の展開が期待されるものの1つといえよう。

あるいは，relevance weight として検索語の重みを考える立場に立脚し，従来同様文献に付与された索引語の重みと乗じることを基本としていながら，検索の意図の明確さあるいは当該検索語への信頼度を表す係数を導入して，それらの一次結合からなるモデルを示したものがある (Kantor, 1981)。実際には，上記係数は明確な検索意図のない場合を表す値，すなわち文献の識別力を最小にする値 1/2 に設定される。これにより，検索語の重みがその最大値 1 に近づくにつれて通常の relevance weight と等しくなるが，逆に最小値 0 に近いほど上述

の係数に一致するような RSV 算出式をえる。その提出された式自体は基本において補償演算子と同じ構成をとるものであることがわかる。

同様に relevance weight としての解釈に属するものであるが，検索語間の関係を規定している論理演算子の種別に依存させて重みの適用を図る考え方がある。この立場からは前述の分離要件は満たせないことになる。これに該当するモデルに Bookstein (Bookstein, 1980) の提案がある。そこでは，その重みを最大値1に近づけるほど当該検索語を連結している演算子の機能を強化し，最小値0に近づけるほど演算子の効力を低下させる，すなわち論理演算子の指示する条件を弛緩させる結果となる。具体的には，論理和演算子のときには従来のファジィ型と同じものを用いるが，論理積演算子で連結されている検索語については重みが最大値1のときには従来のファジィ型と同じものとし，その重みが最小値0に近づくにつれて論理積が有する限定性を減少させ，重みが小さいときの論理和適用結果に近づくモデルとしている。当該モデルも前述の補償演算子を用いたモデルと似た側面を有しており，両者のアプローチの相互関係を整理する必要があるものと思われる。

一方，検索語への重みづけをしきい値と捉えるものにも，いくつかの考え方がある。そのうち最も直接的な適用法を示しているのが，α-レベル・ファジィ集合（λ-レベル・ファジィ集合ともいう）を用いたものであろう (Radecki, 1981; Radecki, 1983a)。そこでは単一検索式中の検索語にはすべて同一の値が与えられ，その値を越えた重みを有する索引語のみが評価の対象とされる。換言すれば，α-レベル・ファジィ集合の適用後は，文献に付与された索引語の重みは α 以上のものに限定され，それ以外はすべて重み0として扱われる。システムからの出力文献数はこの α の値で制御できることになる。Radecki の後者の論文では，さらに順編成ファイルや転置ファイル等，ファイル編成法への上記モデルの適応について考察を展開している。なお，当該モデルは，その構成が単純であるため，前述の分離要件を始めとしてブール型が有する殆どの特性を継承することは容易に推測されよう。また，同モデルにおいては，検索語への重みは単一検索式内ではすべて同じ値を設定するものとされているが，個々の検索語に対して別個の値を設定するよう変更しても問題はなかろう。この点への言及は Buell (Buell, 1981) にも見られる。同様に，α-レベル・ファジィ集合を用いた検索システムにおいて，その処理効率（計算量およびメモリ容量）の最適化を意図したアルゴリズムが示されている (Zenner *et al.*, 1985)。

前記 Radecki のモデルでは，しきい値を境界として個々の検索語の評価結果は当然のことながら不連続となる。それに対してしきい値の上下の値についても連続性を保つよう考慮したものも提案されている。Buell ら (Buell & Kraft, 1981a; Buell & Kraft, 1981b) は，文献に付与された索引語の重みがしきい値を越えないときには両者の比率に基づいた評価値を与えるようにし，しきい値を越えるときにはその越え幅の最大値に対する実際の越え幅の比率に基づき評価値を与えるモデルを複数示している。提示されたモデルにより，従来の relevance weight の立場では解決が困難な問題（例えば，検索語に付与された極めて小さな重みの意味など）を回避することが可能となり，かなり問題点も整理されたといえよう。なお，先に触れた relevance weight の立場に属する Kantor のモデルがとる振る舞いは，値 1/2 をしきい値とみなした場合の上記 Buell らのモデルと極めて似たものとなる点が別途報告されている (Buell, 1985)。

以上で見てきたような各種のファジィ型モデルを整理，評価する枠組みを示したものに「Waller-Kraft wish list」と呼ばれるものがある。これは前述の「分離要件」を始めとして，論理的には等しいが異なる表現をとる検索式間で最終的な RSV の同等性を保証すべきとする「自己一貫性 (self-consistency)」等，計5つの条件を含んでいる。しかしながら，示された条件すべてを同時に満たすことは不可能である点が Buell (Buell, 1982) により検証されており，その後 Cater ら (Cater *et al.*, 1989) は当初のリストに含まれていた検索語の重みに関わる曖昧さを指摘し，代わって8つの条件からなるリストを提示している。最終的に示された8つの条件についてその過不足を評価する能力は現在のところ筆者にはないが，それを基礎にして個々のファジィ型モデルを評価し，さらにはその特性を明確化することを可能とする具体的な基準にまとめあげる作業が今後必要になるものと思われる。その作業自体が同時に，これまで提案された各々のモデルに対する正確な評価の実施を意味し，かつ解決が必要な問題の明確化をも意味することになろう。

なお，ファジィ集合に基づくモデルについては，細野ら（細野ほか，1985）および Kraft ら (Kraft *et al.*, 1983) による解説ないしはレビューがある。特に前者は

その時点までに提案されたモデルを丁寧に跡づけ，適切な解説を示していて大変有用である。ファジィ集合型モデルの概要を把握するためのものとして推奨しておきたい。

B. ファジィ関係に基づくモデル

ファジィ関係 (fuzzy relation) とは，関係の有無を意味する2値にとどまらず，関係の多値表現を可能とするものである。ファジィ関係はその関係の度合いを表すメンバーシップ関数値により決定される。従って，前節では集合への所属度としてすべて表現されていた個々の文献への索引語の付与およびその重みづけ，質問中の検索語への重みづけは，それぞれ文献集合と索引語集合とのファジィ関係，検索質問集合と検索語（索引語）集合とのファジィ関係として捉え，かつ表現することが可能となる。

そこでは，検索処理とは上記の与えられたファジィ関係を合成することにより，文献集合と検索質問集合との最終的なファジィ関係を求めることに帰着する。このような枠組みで検索モデルを提唱したものに，宮本ら (Miyamoto & Nakayama, 1986; Miyamoto, 1989) および Kohout ら (Kohout *et al.*, 1984) によるものがある。前者は特に従来の転置ファイルの活用等を含めた効率的な検索アルゴリズムの開発をも意図したものである。また，ファジィ集合演算同様，ファジィ関係の合成についてもファジィ理論の領域でこれまで多様なものが提案されてきており，そのいずれを情報検索に用いるべきかが問題の1つとなる。Kohout ら (Kohout *et al.*, 1984) は論理演算子の1つである含意 (implication) をも含めて多様な関係の合成を列記しているが，その検討は十分にはなされておらず，単に可能性を示したものと受けとるべきであろう。

また，索引語間の関連性が関連度行列等，多値で与えられているならば，それも同様にファジィ関係として捉え直すことができ，上述のファジィ関係の合成処理による検索モデルにそのまま組み入れることが可能となる。宮本ら (Miyamoto & Nakayama, 1986; Miyamoto, 1989) はこの点を詳細に展開しており，また索引語の類似関係，包含関係等の関連度の算出法，すなわち擬似シソーラスの構成法についてもその効率的なアルゴリズムを併せて示している (Miyamoto, Miyake & Nakayama, 1983)。これに関連する研究として，部分的に与えられた索引語間関連度に対して推移的閉包 (transitive closure) を適用することにより，関連度行列を完成させる方法を示したものがある (Bezdek *et al.*, 1986)。

さらに宮本 (Miyamoto, 1989) は，従来ベクトル型において示されていた関連拡張フィードバック (association feedback) について，ファジィ関係の適用を図ることにより，より柔軟なモデル記述を試みている。関連拡張フィードバックとは，検索でえた文献に付与されていた索引語を自動的に質問表現に加える方法であり，質問拡張を行いながら自動的に検索を繰り返すものである。適合フィードバックとは異なり，その過程に検索者による適合性判定を介入させることはできず，あくまでも検索プロセスをモデル化したものにすぎない。上記論文は推移的閉包を導入し，その性質を巧みに利用しており，これによりベクトル型に付随した困難さの回避を可能としている。さらには同論文において，上記フィードバックはある種のクラスタリングに基づく検索と等しくなる点が示されている。なお，これら一連の宮本らの研究成果は成書としてまとめられており (Miyamoto, 1990)，80 年代における1つの成果を示すものといえよう。

以上のファジィ関係に基づくモデル化は，ある点ではベクトル型の拡張とも考えられ，ベクトル型を含めての統一的なモデル記述の可能性を含んでいる。他方，ファジィ関係に基づくものは，ファジィ集合に基づくモデル化が基本的に備えていた論理演算子への対応を困難とする。宮本 (Miyamoto, 1989) は論理演算子の適用について部分的に触れてはいるが，十分な解決法を示しているわけではない。

C. その他

ファジィ集合に基づくファジィ型モデルと確率型モデルとを，ともにブール型への重みづけを可能にするものとして Bookstein (Bookstein, 1981) は比較を行っている。その結果，前記2つのモデルはともに2値の範囲内ではブール型と一致する，すなわちブール型に収束する点を示してはいるが，それ以上の結論がえられているわけではなく，単にその研究の端緒を示したにすぎない。

ファジィ集合とは異なるが，新たな集合概念として示されたラフ集合 (rough set) を応用した検索モデルも提案されている (Srinivasan, 1989; Srinivasan, 1991)。ラフ集合とは，全体集合が同値関係によって同値類に分割されたところで定義されるものである。情報検索への応用は，仮に索引語（検索語）集合がいくつかの同値類に

分けられているとすれば，それらを用いた検索質問に対して別途定義されている近似の精度が最も高い文献から順に並べることができ，これが適合度順出力に相当することになる。また，同論文では，適切な同値関係を用いれば論理和結合される索引語は同一の同値類に，論理積結合される索引語は異なる同値類に割り当てることが可能としてブール型等，他のモデルとの整合性を検証している。全体としてアプローチ自体は明快であり興味深いものがあるが，研究自体は緒についたばかりとの印象を拭えない。ラフ集合自体は証拠理論の下界確率（確信測度），上界確率（願望測度）と対応することが知られており，その点での展開も今後検討する余地があるものと考えられる。

V. 確率型モデル

A. 索引語の独立性仮定に基づくモデル

確率型（probabilistic）モデルの特徴とは，情報検索において不可避であるところの不確実性の存在を前提とし，それを出発点に据えてモデル構築を図っている点にある。現在までに索引処理のモデルから，狭義の検索処理に関わるものまで多様なものが提案されており，それらの統合化を意図したものも見受けられる。本章では，それらのうち検索処理に関わるものを主として取りあげることにする。

確率型の検索モデルについては複数の導き方が可能であるが，いずれの場合にも最終的には，ある検索質問に関して適合の事象 w_1，不適合の事象 w_2 のそれぞれにおいて文献の索引語表現があるパターン x をとる確率 $P(x|w_1)$ および $P(x|w_2)$ を求めることに帰着する。そして，求められたそれら確率に基づき，個々の文献の相対的出力順序を決定するものである。ここで x は個々の索引語の付与の有無（または文献内出現の有無）を表す要素からなる2値ベクトル表現とする。これらの確率を文献毎に求めるに当たっては，その扱いを簡単にするため，各索引語は統計的に独立，すなわち個々の索引語の付与（もしくは文献における出現）は互いに独立しており，他の索引語の付与（または出現）には影響を及ぼさないとの仮定をおくことが多い。この独立性仮定を採用したモデルは，2値独立性（binary independence）モデルと呼ばれる。そこでは各索引語に関して，適合文献が当該索引語を含み，かつ不適合文献が当該索引語を含まない確率が高いほど大きな値となる関数が設定され，この関数値を質問中の対応する検索語の重みづけ，あるいは質問拡張に用いることになる。これらの値は relevance weight または term precision と呼ばれており，限られた情報（例えば検索者自身による適合判定結果等）に基づき，いかに正確に推定するかが重要な問題となる。

現実的には，試行検索による上位出力文献に対する適合判定結果をフィードバック情報として受け取り，判定された適合および不適合文献集合中での各索引語の出現確率から母集団での出現確率を推定する手順をとる。標本の大きさに関わる問題に対処するため，上記適合フィードバックからえられた値に補正値を導入することが行われている。この点については 70 年代から研究されてきたが，本稿が対象とする80年代においてはRobertson (Robertson, 1986) や Salton ら (Salton & Buckley, 1990) の報告があり，実験結果も併せて示されている。しかしながら，70 年代に提示されたものを含めて，これら多数のうち，いずれが最適であるのかを判断することはできまい。理論的な解決が困難であり，実験結果の積み重ねにしか依拠できない問題ではあるにしても，これらの実験結果からそれを判断することは殆ど不可能であろう。ここに大きなジレンマが存する。

また，これまで検索語の重みづけと質問拡張するために候補となる索引語を選び出すことは同一式で行われており，Smeaton ら (Smeaton *et al.*, 1983) は relevance weight を特に質問拡張に適用したときのいくつかの可能な選択肢について実験結果を報告している。しかしながら，他方では検索語の重みづけと質問拡張とは本来区別して考えることが必要との意見も提出されている (Robertson, 1990)。確かに情報検索において確率的な考え方はさまざまなところで必要とされるが，安易な適用を図ると確率型モデル自体が導き出された理論的な基盤から遊離してしまう，すなわち理論的に保証される範囲を逸脱してしまう点を戒めた意見として傾聴に値するものである。なお，Wu ら (Wu *et al.*, 1981) はこれら算出された relevance weight を，従来ベクトル型で用いられてきた形式での適合フィードバックに適用することを試み，その実験結果を報告している。

relevance weight の推定における他の課題として，標本の性質に関わるものがある。通常，試行検索結果の上位出力文献を推定に用いる標本とするため，必ずしも無作為標本とはいいきれない問題が残る。そこで Losee (Losee, 1987) は検索質問に最も合致する文献から抽出されるプロセスを反映するよう，上記2値独立性モデル

を修正している。示されたモデル自体の適否を簡単に判断することはできないが，これまでモデル化がなされてこなかった側面に正面から取り組んだものとして評価されよう。

また，relevance weight の値と，その構成要素である当該索引語をもつ文献数との関係も明らかにされている (Yu, Lam & Salton, 1982)。結論のみ記せば，当該索引語をもつ文献数が0から全適合文献数までの間は，その relevance weight は増加関数となり，それを越えると減少関数になる。このような，いわば解析的な解明によって初めて，relevance weight の性質も真の意味で明らかになるのであり，実験結果と両輪をなすことを強調しておきたい。

B. 索引語間の従属性を組み入れたモデル

現実的には，多くの場合に索引語間には何らかの従属関係が存在すると予想されるため，それら従属関係の検索処理への反映が重要な課題となる。索引語間に想定される従属関係の最も簡単な近似として，各索引語は最も緊密な関係にある他の索引語1つとのみ依存関係を有するものと仮定することができる。このときに導き出されるものが［1次］木従属 (tree dependence) モデルである。これは 70 年代に示されたものであり，当該モデルの適用に当たっては，与えられた制約の中で各索引語の1次従属関係を最もよく反映可能な測度の設定が課題となる。van Rijsbergen ら (van Rijsbergen *et al.*, 1981) はこの目的において使用可能な式を整理しており，その結果えられる従属木（最大生成木）を質問拡張に用いることを試みている。なお，2次以上の高次の木従属モデルを考えることも可能であるが，従属関係の導出がさらに複雑となる点は容易に予想されよう。また，上記論文につながるものとして前節で取りあげた Smeaton ら (Smeaton *et al.*, 1983) の実験報告があり，そこでは質問拡張に最大生成木に加えて，文献のクラスタリング結果である最近隣文献 (nearest neighbors) 等を用いて，複数の質問拡張の選択肢を実験している。

また，仮に検索者の設定した検索質問が検索語の列挙ではなく論理演算子（ただし和，積のみ）を用いて構成されているのであれば，この検索式を論理和標準形 (disjunctive normal form) に変換し，各々の節中の論理積で結合されている検索語間には従属関係が存在すると考えるものもある (Croft, 1986)。そしてこれはそのまま適合文献中における索引語間の従属関係を表すものと考え，前記2値独立性モデルに組み入れたものが提案されている。ただし，その採用した仮定の妥当性に関しては疑問が残る。

索引語間に存在するすべての従属関係を組み入れたモデルに，BLE (Bahadur-Lazarsfeld expansion) モデルがある (Salton, Buckley & Yu, 1983; Yu, Buckley, Lam & Salton, 1983)。ここにおいては一切の仮定が不要となるが，高次従属度を含めてすべての従属度を算出することは不可能であるため，現実的な適用に際しては適当な範囲内の関連度を選択して組み入れることになる。上記2つの論文は，BLE モデルの簡略化として，前述した1次従属木に新たな枝を追加し，部分的に閉路を含むグラフ構造に拡張する試みを示している。その実験結果によれば，採用した方法は必ずしも良好な結果を示していないが，BLE モデルの有効性を活かした近似法は他にも考えられよう。

索引語の従属関係をすべて組み入れた他のモデルとして，最大エントロピー原理 (maximum entropy principle) を用いたものも提案されている (Cooper *et al.*, 1982; Cooper, 1983)。これは既知の索引語間従属度を制約条件と考え，これら与えられた制約条件を満たす形で，分布に関わるエントロピーを最大にする，換言すれば制約条件が指示する情報量を越えたものとしての分布が与える情報量を最小にしようとするものである。これによりすべての索引語間の関係を反映することが可能になるが，必要とする計算量が極めて大きくなるとの問題を含んでいる。なお，Kantor (Kantor, 1984) は1つの効率的な計算法を示している。

C. 索引語の文献内出現頻度を用いたモデル

前節までにおいては文献の表現は個々の索引語の付与(出現)の有無を表す2値ベクトルからなっていた。それに対して，個々の索引語に関する表現を2値ではなく，当該文献における出現頻度を反映したものにすることも考えられている。例えば，Croft (Croft, 1981) は期待値の考えを適用して，索引語が文献に付与される（出現する）確率を導入している。そして，この確率を当該索引語の文献内出現頻度に対応したものとすることができる点を示している。同モデルを用いた実験結果も別途，報告されている (Croft, 1983)。

同様に，索引語の文献内出現頻度を用いたものに，非2値独立性 (nonbinary independence) モデルがある (Yu, Meng & Park, 1989)。これは従来の確率型モデル

における解釈の枠組みを変え，質問中の検索語は2値表現とし重みづけを行わず，逆に文献ベクトルは各索引語の出現頻度もしくはそれを反映した重みを与え，多値として扱うものである。なお，同モデルは確率型モデルが満たさなければならない重要な特性をそのまま継承していることが示されているが，同モデルの導出法およびその妥当性を説得力をもって示すにはさらにいくつかの点で説明もしくは証明を要するものと筆者には思われる。上記論文中の記述のままでは，その点十分とはいえないものと考えている。

次に，索引語の文献内出現頻度をポアソン分布に従うものと仮定した2-ポアソン独立性 (2-Poisson independence) モデルを取りあげる。当初，2-ポアソン・モデルは索引処理，すなわち索引語の選定およびその重みの決定を主眼にしたものとして提案されたが，それを検索処理に転用することを明確に企図したのは Robertson ら (Robertson, van Rijsbergen & Porter, 1981) であった。そこでは，ある索引語に関して蓄積された全文献は2つのクラス（当該索引語により主題・内容が表現されるものとそれ以外のもの）に分かれ，これら2つの文献集合内では各文献における当該索引語の出現頻度はそれぞれ異なるポアソン分布に従うとの仮定をとっている。同モデルにおいてもポアソン分布を規定する平均値の推定が大きな課題として残されており，当初に示された推定法（積率母関数を用いるもの）に関わる問題点を Raghavan ら (Raghavan, Shi & Yu, 1983) は指摘し，その改善を試みている。これらの研究の蓄積はあるにしても，筆者は，検索処理で取り扱う範囲内では索引語の出現頻度をポアソン分布を含めて理論的な確率分布に当てはめることはかなり困難と考えている。なお，索引処理の枠内においての議論であるが，ある索引語に関して3つ以上のポアソン分布を想定し，上記2-ポアソン・モデルを一般化する試みも見られる (Srinivasan, 1990)。ただし，実験結果によれば，一般化は必ずしも望ましい結果をもたらさず，逆に改悪につながっていることが示されている。

D. ベイズ推論に基づくモデル

ベイズ推論の方法に直接基づくものに逐次学習 (sequential learning) モデルと呼ばれるものがある (Bookstein, 1983a)。それは，その時点までに出力した文献に対する検索者の適合判定結果を逐次学習し，各々の索引語に関して，適合および不適合文献でのその出現に関する仮定された分布の母数を順次推定していくものと概括することができる。具体的には，検索開始時の事前知識を表す分布がそれまでの適合フィードバックの結果を受けて修正されたものがその時点での事前分布となり，標本情報をこの事前分布について期待値をとったものが予測分布となる。よって，残された適合および不適合文献において各々の索引語が出現する確率の予測分布が導き出され，それらに基づき残された文献の適合・不適合を識別する関数が求められる。以上の枠組みに依拠して，2値独立性モデルおよび2-ポアソン独立性モデルにそれぞれ対応する逐次学習モデルが提案されている。同モデルにおいては，検索開始時の分布を規定する4つのパラメータが新たに導入されたことにより，前記2つのモデルに比してさらに柔軟な検索処理をモデル化しえたといえよう。

同モデルにおいて必要となる4つのパラメータのうち，2つは適合文献に，他の2つは不適合文献にそれぞれ関連するものである。この不適合文献に関連するパラメータに関しては，蓄積されている全文献に基づき推定しても極端な不都合はない点が確認されている (Losee, Bookstein & Yu, 1986; Losee, 1988)。しかしながら，適合文献に関連する残り2つのパラメータについては，前記の不適合文献に関わるパラメータ値を参考とするいくつかの推定法が示されるにとどまり，確定的な解決法が見い出されるまでには至っていない (Losee, 1988)。以上の通り，逐次学習モデルにおいてもパラメータ推定の課題が完全に解決されたわけではなく，重要な問題として残されているのは事実としても，同モデル自体は学習とのプロセスをモデル化した試みとして大変重要な成果と考えられよう。今後の研究の進展を切に期待したいものの1つである。

なお，論理演算子を用いて表現された検索式を逐次学習モデルに適用することも試みられており，実験では概ねよい結果が報告されている (Losee & Bookstein, 1988)。そこでは，検索者の設定した検索式をすべて論理積標準形 (conjunctive normal form) に変換し，論理積で結合されている各々の節を1つの仮想的な索引語 (hyperterm; hyperfeature) とみなし，当該節に含まれる個々の索引語をまとめたものとして扱っている。例えば，出現頻度についてであれば，当該節中の個々の索引語の出現頻度を合計して用いることになる。そしてこの仮想的な語を単位にして逐次学習モデルを実行するものとしている。

ベイズ推論の枠組みを採用するもう1つのモデルに，主観確率 (subjective probability) を導入し，それに基づき推論を行おうとするものがある (Thompson, 1990a; Thompson, 1990b)。これは文献の索引者および検索者の主観確率をベイズ理論に従い組み合わせて，検索処理等を実施しようとするものであり，次節で述べる統一的な確率型モデル構築の可能性を探ったものの1つでもある。主観確率がベータ分布等に従うときの具体的なモデルが示されているが，筆者の皮相的な理解では，実際の検索処理への適用が大変困難なもののように思われる。同モデルおよびその基盤となる心理学分野でのこの点に関する展開を見守りたい。

E. その他

確率型モデルに関する研究の展開の中で注目すべきものの1つに，統合化されたモデルの提唱がある。本章の冒頭に述べたように，確率型モデルには索引作業を意図したものが一方に存在するため，これまで取りあげた検索処理のモデルとの統合が考えられている。検索処理はある検索質問と文献群との関係を確率的に示したものであり，確率型での索引処理はある文献と質問群との関係を確率的に示したものと捉えるならば，両者を統合して質問群と文献群との関係づけ，あるいは個々の質問と個々の文献との関係づけを扱う単一のモデルを想定することができよう。これを定式化したものが Robertson らによって示されている (Robertson, Maron & Cooper, 1982; Maron, 1984)。その定式化に準拠した，2値独立性モデルの枠組みでの具体的な統合モデルが Wong ら (Wong & Yao, 1990) により示されている。当該モデル化においては，従来のモデルにおける独立性仮定よりもさらに制約の厳しい独立性仮定が導入されており，その意味する事柄の再検討等，今後に残された部分が大きいものと考えられる。

各々の文献を索引語集合がつくる空間上の確率分布とみなすモデルがあり，確率分布 (probability distribution) モデルと名付けられている (Wong & Yao, 1989)。個々の索引語の文献内出現頻度をその文献を単位に正規化すれば，当該文献自体は索引語空間における確率分布を表すものと考えることができる。このような枠組みに依拠し，具体的には効用理論および情報理論に基づく2つの検索関数が提案されている。なお，効用理論に基づくものの1例が，ある仮定をおけば，索引語の文献内出現頻度を利用する Croft のモデル (Croft, 1981) に等しくなることが併せて示されている。確率分布モデルで示されたような新たな確率型モデルの構築はそれ自身において興味深いが，筆者はさらに他の確率型との適切な関係づけ，もしくは関係の把握が重要と考えている。同モデルについても，上記論文中で示された部分以上に他の確率型に属するモデルとの関係づけが可能であり，かつ必要でもあると考えている。

また，他の確率型モデルでは捨象もしくは極端に単純化して扱ってきた効用/損失/費用 (utility/risk/cost) との要素を全面的に導入してモデル化を図った試みに，Salton ら (Salton & Wu, 1981) のものがある。そこでは，従来の確率型に従い質問中の検索語に対する重みづけを行うに当たって，適合・不適合の事象と索引語の出現・非出現の事象との組み合わせに効用の要素を導入し，その上で最適な重みづけを考えたときのその重みの特性等を考察している。

これまでに取りあげた確率型モデルにおいては，適合フィードバック等における検索者による文献の適合判定をすべて2値，すなわち適合・不適合の2つに分け，適合と判定された文献はすべて同程度に適合するものと見なされていた。そこで，より柔軟な判定結果の表現を可能とするためには，適合判定自体の多値表現を認めることが考えられる。しかしながら，多値化された判定結果の検索処理への反映には多くの困難が予想される。Bookstein (Bookstein, 1983b) は問題の整理と解決の枠組みを示すことを試みているが，必ずしも成功しているわけではなく，問題はそのまま残されているといえよう。

以上見てきた通り，確率型にも多岐にわたる検索モデルが含まれている。それらの整理の仕方も一通りではなく，本章で採用したものが必ずしも最良のものとはいえないであろう。確率型モデルについてまとまった解説を行ったものに鈴木 (鈴木, 1991) によるものがある。また，文献レビューには Bookstein (Bookstein, 1985) のものがあり，それらを併せて参考にしていただきたい。

VI. 統合型モデル

これまで取りあげてきた各種検索モデルの統合化，さらには複数のモデルを組み合わせた性格を有する新たなモデルの構築は，情報検索モデル研究にとり大きなテーマをなす。この方向に沿った研究成果もこれまでにいくつか報告されている。

それら統合化の試みの中で最も完成度が高いものは，拡張ブール (extended Boolean) モデルと名付けられた

ものであろう (Salton, Fox & Wu, 1983)。その最大の特徴は，検索質問中の各論理演算子に付与された，いわば結合の強さを指示するパラメータの値を変化させることにより，単一のモデルがベクトル型として機能したり，ブール型もしくはファジィ型として機能したり，あるいはそれらの中間的な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，通常のユークリッド距離概念を一般化したミンコウスキー距離をその基礎に据えて構築したものであり，論理和・論理積演算子のそれぞれに適用する2つの式から構成される。論理和と論理積が組み合わされた複合質問についても，部分質問に対する処理を再帰的に適用していくことで対処可能となる。また，同モデルによる演算結果は max 演算と min 演算の間の値となり，ファジィ理論においては平均演算子と呼ばれるものに該当する。拡張ブール・モデルの提案は80年代における大きな成果であり，その特性は極めて有効なものと考えられる。上記 Salton らの論文においてもその数学的な性質が考察されているが，さらにこの点の検討が必要となろう。また，これまでに同モデルを応用したものとして，適合フィードバック処理に部分的に適用したもの (Salton, Buckley & Fox, 1983; Salton, Fox & Voorhees, 1985; Salton, 1988)，あるいは索引語間関連性の組み込みを意図したもの（谷口，1990）などが見受けられるが，今後の一層の展開を強く期待したい。拡張ブール・モデルは，それだけの価値を有するモデルといっても過言ではなかろう。

統合型モデルに属する他のものに，ベクトル型のところで取りあげた一般化ベクトル空間モデルを拡張したものがある (Wong, Ziarko, Raghavan & Wong, 1989)。そこにおいては当初，各文献はベクトル表現されていたが，それを索引語の論理和結合であると解釈し直し，定義された論理和演算を適用し，新たに設定された基本概念によるベクトル表現に変換する。同様に，論理演算子を用いた検索質問にも同じく定義された演算を適用し，基本概念によるベクトル表現に変換する。そして，最終的に質問と文献との類似度をえるためにはベクトル型に属する類似測度をそのまま適用するものとしている。以上により，ベクトル型モデルへの論理演算の導入，すなわちベクトル型とブール型との統合が実現されたことになる。しかしながら，筆者にとっては一般化ベクトル空間モデルについて述べた疑問がそのまま上記モデルにも当てはまることになり，従って上記モデル自体の適否についてはそれ以上判断することができない。

ベクトル型と確率型とを統合化しようとする試みもある (Bookstein, 1983c)。通常の確率型においては，式の形の上では relevance weight により重みづけされた質問ベクトルと各文献ベクトルとの内積をとったものが適合・不適合識別関数の基本となるため，質問・文献両ベクトルの内積を基本にして類似度とするベクトル型と同形のものと捉えることができる。この点は確率型が提唱され出した当時から自覚されていたが，より厳密にベクトル型，確率型両モデルの統合化を意図して検討を試みたのが上記 Bookstein の論文である。ただし，その作業は緒に着いたばかりといってよく，有益な進展がえられたわけではない。また，前章で触れた非2値独立性モデルのようなものが出てきた状況では，ベクトル型，確率型両モデルの境界がますます不明瞭となり，基本的なところから整理すべき必要があるものと思われる。

基本枠は従来のブール型システムのままとし，確率型の方法を一部導入して適合度順出力を行おうとする試みもある (Radecki, 1982; Radecki, 1988)。まず論理演算子で結合されたすべての検索質問は論理和標準形に変換する。次に論理和標準形を構成する各々の節毎に個々の検索語（否定演算子が付加されていないもののみ）について確率型に従い relevance weight を計算し，当該節内でのその合計値をもって，当該節が指示する条件を満たす文献の RSV とするものである。これはあくまでもブール型の枠組みを保持し，それに適用できるものを求めたものであり，その点で検索処理のモデル化として見たときには単純なものとなっている。

VII. その他のモデル

前章までに取りあげることのできなかったものを，以下にいくつか記す。1つは集合論に立脚して，索引語間の関連性の組み入れを行ったものである (Ito *et al.*, 1984)。検索質問と文献間の類似測度は集合論的に捉えれば，両者間の索引語の和集合に対する共通集合の比率と捉えることができる。そこでそれら集合の濃度を求めればよいことになる。その際，索引語間の関連性を集合的な重なり具合として捉え，集合論における包除原理（和積の原理）を適用して2つ組の索引語間の関連度から3つ組の索引語間関連度，さらにはより多くの n 個組の索引語間関連度までをすべて濃度の算出に組み入れることが考えられる。現実的には，それら2つ組から n 個組まですべての関連度が与えられることはないため，適切な近似法の検討が重要となる。上記論文では2つ組の

関連度のみ与えられている場合を仮定し，それを用いて3つ組以上の関連度を近似する方法が提案されている。同モデルは想定されるすべての索引語間関連度を組み入れることができる点で確率型における BLE モデルと類似する。

また，Belkin らによる類型化ではベクトル型，ファジィ型，確率型のいずれとも異なるネットワーク型に属するものとされている活性伝播 (spreading activation) モデルがある (Jones *et al.*, 1987)。これは索引語集合を基礎にして質問および文献をそれらとのネットワークで表現し，質問がネットワークの一部を活性化するとそれらの活性状態が次々と伝播され，最終的に活性化された節点（ノード）に対応する文献が検索されたものと考えるモデルである。示された最も単純なモデルにおいては，質問・文献両ベクトルの内積に対して，ある正規化処理を施したものとの形式を結果的にはとっている。しかし，多段階のネットワーク構成を考えれば，そこに索引語間関連度を組み入れることは容易であろうし，あるいは他の分野で盛んに研究されているニューラル・ネットワークにならい，活性を伝播する際にしきい値等の関数を設けることなどが考えられ，多様な展開が期待できよう。

独立した検索モデルとはいい難いものではあるが，論理演算子を用いた適切な検索式を自動的に構成させる方法も研究されている (Salton, Buckley & Fox, 1983; Salton, Voorhees & Fox, 1984; Salton, Fox & Voorhees, 1985; Salton, 1988)。検索者により提出されたベクトル表現の検索質問を，あるいは自然言語による文の形式で提出された検索質問から検索語を抽出しベクトル表現に落とした検索質問を，最終的にはそれらを論理和標準形の検索式に構成し直すものである。その過程においては，各索引語を有する文献数を基本にした重みを用いて，あるいは適合フィードバックを適用してえられる relevance weight を用いて，検索される予想文献数を基準にしながら最適な検索式に洗練していく手法をとる。筆者は，そのような課題設定の必要性，すなわちベクトル表現された検索質問（自然言語文から検索語を抽出した段階のものを含む）をあえて論理演算子を用いた検索式にシステム側で自動的に構成し直すことの意義について疑問を抱いている。

情報検索の隣接領域であるデータベース理論研究について，最後に一言触れておく。情報検索理論とデータベース理論との関係は，場合によっては文献検索とデータ検索との相違として論じられており，その関係づけはさまざまに捉えられている。両者の相違について，部分的にせよ言及したものに Blair (Blair, 1984)，細野 (細野, 1991)，宮本 (Miyamoto, 1990)，鈴木 (鈴木, 1991) のものがあり，特に Blair の論文はこの点を丁寧にまとめている。筆者は，端的にいえば，前者は文献の内容・主題等構造化されていないデータの検索，あるいは単純な照合にとどまらず抽象度の高い検索質問に答えることを主に想定したもの，換言すれば不確実性を常にはらんだ検索と考えている。一方，後者はその出発点において構造的に整理可能なデータを専ら検索対象に据えてきたものといえよう。しかしながら，80年代を通して情報検索モデル研究よりも一層の進展を見せた後者においては，最近特にオブジェクト指向データベース等において対象とする範囲を拡大させつつあり，その関係はもはや単純に割り切ることはできまい。今後の推移を見守りたい。ちなみに，データベース理論中のリレーショナル・データベース理論の枠組みを用いて情報検索の問題を扱うことを試みたものに Crawford (Crawford, 1981) や Blair (Blair, 1988) のものがある。これらデータベース理論もしくはシステムに準拠することにより，それらの成果・利点をそのまま流用できる点は，特にシステムの実現化段階では魅力的であろう。しかしながら，検索処理の観点のみからいえば，上記論文で扱っている範囲では，検索理論の点で新たな提案・進展が見られるわけではないことをここでは記しておきたい。

VIII. おわりに

本稿では 80 年代における情報検索モデル研究の成果についてかなり広範囲に取りあげることを意図したが，最初に述べた通り，当該領域に関わる成果報告は相当な量に上るため，取りあげることができなかったものも多い。それらについては紹介したレビュー等を適宜参照していただきたい。また，前章までに取りあげた個々のモデルないし論文の解説に付した見解等はあくまでも筆者の個人的なものにすぎず，客観的根拠を欠くものも多い。それを補う意味でも，他のレビュー等の参照を推奨したい。

最後に 80 年代の成果を総括しておこう。まず総体としては，それ以前の 60，70 年代からの継続とまとめられよう。すなわち，情報検索システム自体の枠組みあるいはモデル研究の枠組みを大きく変更するような成果はえられておらず，基本的にはそれ以前に提案されたモデ

ルの洗練化という性格が強い。むしろ，本稿では除外した自然言語処理あるいは知識処理に基づくアプローチが80年代を特徴づけるといえるのかもしれない。当然のことながら，本稿で取りあげた範囲内のものに関しても，前章までに記した通り，部分毎に見れば各種の取り組みや提案の蓄積が多数あり，有効なあるいは興味深い成果も同期間内に多数生み出されたものと評価すべきではある。このような評価は個人毎に異なるものと予想されるが，情報検索理論研究に対する評価のまとまった表明としては細野（細野，1989; 細野，1991）のものが挙げられよう。

先の総括を受けて，本稿で取りあげた各種検索モデルが内包する基本的問題点を，筆者の立場から簡単にまとめておこう。何よりもブール型を除くいずれのモデルも大規模な実用システムとして実現化されることはなかった点が多くの事柄を示唆しているといえよう。ただし，Salton らが中心となり Cornell 大学で 60 年代から開発および実験を重ね，多数の成果および論文を生み出してきた SMART システムが，90年4月から Individual 社により一般向けのサービスを開始したとのニュースを未確認ながら聞いている。公開されたシステム機能等の詳細は不明であるが，いずれにしても非ブール型システムの実用化を喜びたいし，今後の動向に注目したい。

まず，それら非ブール型検索モデルの基本的な問題点として，システム内部の計算量の飛躍的な増加が挙げられる。しかしながら，この点に関してはブール型で用いられる転置ファイルの部分的利用や，蓄積情報のクラスタリングに基づくクラスタ化ファイルの活用等，計算量自体を抑え込む試みや，さらにはマルチ・プロセッサによる並列処理の活用を含めて応答速度を上げようとの多様な試みが行われている。基本的にはハードウェアの性能向上により，これら計算量の問題はいずれ解消されるものと考えたい。

逆に，より重視すべき問題点として，これまでに提案されたいずれの検索モデルといえども，情報検索が内包する問題の複雑さに比べれば単純なモデル化にとどまる点が指摘できよう。単一の方式ですべての検索質問に答えようとすること自体に無理があり，本来は検索が行われる状況や対象領域等に応じた個別知識および検索手法の柔軟な適用を組み入れていくべきであろう。あるいは，文献の内容・主題の表現を単なる索引語の並び（またはそれらへの重みづけを含む）ではなく，より豊かなものとし，それに対応した検索方式を開発していく必要性があろう。問題を本稿で取りあげた検索処理のレベルに限定すれば，先に触れた自然言語処理や知識処理に基づくアプローチと本稿で取りあげたアプローチとの組み合わせも成果が期待できる課題であろう。いずれにしても，90年代には一層の研究成果が生み出されることを期待したい。

1) 情報検索モデル研究に属する論文は，一度国際会議等（特に ACM SIGIR Conference など）において発表された後に，フルペーパーに直され各種論文誌に掲載されるものも多い。従って，それらについては基本的に同内容のものが会議録と論文誌に掲載されることになる。本稿では引用の煩雑さを避ける目的で，そのような場合には論文誌の掲載論文のみ引用するにとどめる。
2) 同種の試みの嚆矢は，管見によれば Salton ら (Salton, Fox & Wu, 1983) に見られる。

Belkin, N. J.; Croft, W. B. "Retrieval techniques". Annual Review of Information Science and Technology, Vol. 22. Williams, M. E., ed. Amsterdam, Elsevier Science Publishers, 1987, p. 109-145.

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

Blair, D. C. The data-document distinction in information retrieval. Communications of the ACM. Vol. 27, No. 4, p. 369-374 (1984)

Blair, D. C. An extended relational document retrieval model. Information Processing & Management. Vol. 24, No. 3, p. 349-371 (1988)

Bookstein, A. Fuzzy requests: an approach to weighted Boolean searches. Journal of the American Society for Information Science. Vol. 31, No. 4, p. 240-247 (1980)

Bookstein, A. A comparison of two systems of weighted Boolean retrieval. Journal of the American Society for Information Science. Vol. 32, No. 4, p. 275-279 (1981)

Bookstein, A. Information retrieval: a sequential learning process. Journal of the American Society for Information Science. Vol. 34, No. 5, p. 331-342 (1983a)

Bookstein, A. Outline of a general probabilistic retrieval model. Journal of Documentation. Vol. 39, No. 2, p. 63-72 (1983b)

Bookstein, A. "Explanation and generalization of vector models in information retrieval". Research and Development in Information Retrieval. Salton, G. et al., eds. Berlin, Springer,

1983c, p. 118-132. (Lecture Notes in Computer Science, 146)

Bookstein, A. "Probability and fuzzy-set applications to information retrieval". Annual Review of Information Science and Technology, Vol. 20. Williams, M. E., ed. White Plains, N. Y., Knowledge Industry Publications, 1985, p. 117-151.

Bordogna, G.; Carrara, P.; Pasi, G. Query term weights as constraints in fuzzy information retrieval. Information Processing & Management. Vol. 27, No. 1, p. 15-26 (1991)

Buell, D. A. A general model of query processing in information retrieval systems. Information Processing & Management. Vol. 17, No. 5, p. 249-262 (1981)

Buell, D. A. An analysis of some fuzzy subset applications to information retrieval systems. Fuzzy Sets and Systems. Vol. 7, No. 1, p. 35-42 (1982)

Buell, D. A. A problem in information retrieval with fuzzy sets. Journal of the American Society for Information Science. Vol. 36, No. 6, p. 398-401 (1985)

Buell, D. A.; Kraft, D. H. Threshold values and Boolean retrieval systems. Information Processing & Management. Vol. 17, No. 3, 127-136 (1981a)

Buell, D. A.; Kraft, D. H. A model for a weighted retrieval system. Journal of the American Society for Information Science. Vol. 32, No. 3, p. 211-216 (1981b)

Cater, S. C.; Kraft, D. H. A generalization and clarification of the Waller-Kraft wish list. Information Processing & Management. Vol. 25, No. 1, p. 15-25 (1989)

Cooper, W. S. Exploiting the maximum entropy principle to increase retrieval effectiveness. Journal of the American Society for Information Science. Vol. 34, No. 1, p. 31-39 (1983)

Cooper, W. S.; Huizinga, P. The maximum entropy principle and its application to the design of probabilistic retrieval systems. Information Technology: Research and Development. Vol. 1, No. 2, p. 99-112 (1982)

Crawford, R. G. The relational model in information retrieval. Journal of the American Society for Information Science. Vol. 32, No. 1, p. 51-64 (1981)

Croft, W. B. Document representation in probabilistic models of information retrieval. Journal of the American Society for Information Science. Vol. 32, No. 6, p. 451-457 (1981)

Croft, W. B. Experiments with representation in a document retrieval system. Information Technology: Research and Development. Vol. 2, No. 1, p. 1-21 (1983)

Croft, W. B. Boolean queries and term dependencies in probabilistic retrieval models. Journal of the American Society for Information Science. Vol. 37, No. 2, p. 71-77 (1986)

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman, R. Indexing by latent semantic analysis. Journal of the American Society for Information Science. Vol. 41, No. 6, p. 391-407 (1990)

細野公男. システム志向から情報要求者志向へ. 情報管理. Vol. 32, No. 6, p. 489-500 (1989)

細野公男. 情報検索理論・技法の問題点とその解決の方向. 情報処理学会研究報告. 情報学基礎. No. 91-FI-24, 7p. (1991)

細野公男, 高柳敏子, 後藤智範, 原田隆史. ファジィ集合理論に基づく重み付き文献検索システム. Library and Information Science. No. 23, p. 137-147 (1985)

伊藤哲郎. 情報検索. 東京, 昭晃堂, 1986, 174 p. (ソフトウェア講座, 19)

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)

Jones, W. P.; Furnas, G. W. Pictures of relevance: a geometric analysis of similarity measures. Journal of the American Society for Information Science. Vol. 38, No. 6, p. 420-442 (1987)

Kantor, P. B. The logic of weighted queries. IEEE Transactions on Systems, Man, and Cybernetics. Vol. SMC-11, No. 12, p. 816-821 (1981)

Kantor, P. B. Maximum entropy and the optimal design of automated information retrieval systems. Information Technology: Research and Development. Vol. 3, No. 2, p. 88-94 (1984)

Kohout, L. J.; Keravnou, E.; Bandler, W. "Automatic documentary information retrieval by means of fuzzy relational products". Fuzzy Sets and Decision Analysis. Zimmermann, H.-J. et al., eds. Amsterdam, Elsevier Science Publishers, 1984, p. 383-404. (TIMS Studies in the Management Sciences, 20)

Koll, M.; Srinivasan, P. Fuzzy versus probabilistic models for user relevance judgments. Journal of the American Society for Information Science. Vol. 41, No. 4, p. 264-271 (1990)

Kraft, D. H. "Advances in information retrieval: where is that /#*&@$ record?". Advances in Computers, Vol. 24. Yovits, M. C., ed. Orlando, Fla., Academic Press, 1985, p. 277-318.

Kraft, D. H.; Buell, D. A. Fuzzy sets and generalized Boolean retrieval systems. International Journal of Man-Machine Studies. Vol. 19, No. 1, p. 45-56 (1983)

Lochbaum, K. E.; Streeter, L. A. Comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval. Information Processing & Management. Vol. 25, No. 6, p. 665-676 (1989)

Losee, R. M. The effect of database size on document retrieval: random and best-first retrieval models. Proceedings of the Tenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Yu, C. T. et al., eds. New Orleans, La., 1987-06. New York, ACM, 1987, p. 164-169.

Losee, R. M. Parameter estimation for probabilistic document-retrieval models. Journal of the American Society for Information Science. Vol. 39, No. 1, p. 8-16 (1988)

Losee, R. M.; Bookstein, A. Integrating Boolean queries in conjunctive normal form with probabilistic retrieval models. Information Processing & Management. Vol. 24, No. 3, p. 315-321 (1988)

Losee, R.; Bookstein, A.; Yu, C. Probabilistic models for document retrieval: a comparison of performance on experimental and synthetic databases. [Proceedings of the] 1986-ACM Conference on Research and Development in Information Retrieval. Rabitti, F., ed. Pisa, 1986-09. [s. n.], 1986, p. 258-264.

Maron, M. E. "Probabilistic retrieval models". Progress in Communication Sciences. Vol. 5. Dervin, B. et al., eds. Norwood, N. J., Ablex Pub. Corp., 1984, p. 145-176.

Miyamoto, S. Two approaches for information retrieval through fuzzy associations. IEEE Transactions on Systems, Man, and Cybernetics. Vol. SMC-19, No. 1, p. 123-130 (1989)

Miyamoto, S. Fuzzy Sets in Information Retrieval and Cluster Analysis. Dordrecht, Kluwer Academic Publishers, 1990, 259 p. (Theory and Decision Library. Series D, Vol. 4)

Miyamoto, S.; Miyake, T.; Nakayama, K. Generation of a pseudothesaurus for information retrieval based on cooccurrences and fuzzy set operations. IEEE Transactions on Systems, Man, and Cybernetics. Vol. SMC-13, No. 1, p. 62-70 (1983)

Miyamoto, S.; Nakayama, K. Fuzzy information retrieval based on a fuzzy pseudothesaurus. IEEE Transactions on Systems, Man, and Cybernetics. Vol. SMC-16, No. 2, p. 278-282 (1986)

Noreault, T.; McGill, M.; Koll, M. B. "A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment". Information Retrieval Research. Oddy, R. N. et al., eds. London, Butterworths, 1981, p. 57-76.

Paice, C. D. Soft evaluation of Boolean search queries in information retrieval systems. Information Technology: Research and Development. Vol. 3, No. 1, p. 33-41 (1984)

Radecki, T. Outline of a fuzzy logic approach to information retrieval. International Journal of Man-Machine Studies. Vol. 14, No. 2, p. 169-178 (1981)

Radecki, T. A probabilistic approach to information retrieval in systems with Boolean search request formulations. Journal of the American Society for Information Science. Vol. 33, No. 6, p. 365-370 (1982)

Radecki, T. Generalized Boolean methods of information retrieval. International Journal of Man-Machine Studies. Vol. 18, No. 5, p. 407-439 (1983a)

Radecki, T. A theoretical background for applying fuzzy set theory in information retrieval. Fuzzy Sets and Systems. Vol. 10, No. 2, p. 169-183 (1983b)

Radecki, T. Probabilistic methods for ranking output documents in conventional Boolean retrieval systems. Information Processing & Management. Vol. 24, No. 3, p. 281-302 (1988)

Raghavan, V. V.; Shi, H.; Yu, C. T. Evaluation of the 2-Poisson model as a basis for using term frequency data in searching. ACM SIGIR Forum. Vol. 17, No. 4, p. 88-100 (1983)

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)

Robertson, S. E. On relevance weight estimation and query expansion. Journal of Documentation. Vol. 42, No. 3, p. 182-188 (1986)

Robertson, S. E. On term selection for query expansion. Journal of Documentation. Vol. 46, No. 4, p. 359-364 (1990)

Robertson, S. E.; Maron, M. E.; Cooper, W. S. Probability of relevance: a unification of two competing models for document retrieval. Information Technology: Research and Development. Vol. 1, No. 1, p. 1-21 (1982)

Robertson, S. E.; van Rijsbergen, C. J.; Porter, M. F. "Probabilistic models of indexing and search-

ing". Information Retrieval Research. Oddy, R. N. et al., eds. London, Butterworths, 1981, p. 35-56.

Salton, G. A simple blueprint for automatic Boolean query processing. Information Processing & Management. Vol. 24, No. 3, p. 269-280 (1988)

Salton, G. Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer. Reading, Mass., Addison-Wesley, 1989, 530 p.

Salton, G.; Buckley, C. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science. Vol. 41, No. 4, p. 288-297 (1990)

Salton, G.; Buckley, C.; Fox, E. A. Automatic query formulations in information retrieval. Journal of the American Society for Information Science. Vol. 34, No. 4, p. 262-280 (1983)

Salton, G.; Buckley, C.; Yu, C. T. "An evaluation of term dependence models in information retrieval". Research and Development in Information Retrieval. Salton, G. et al., eds. Berlin, Springer, 1983, p. 151-173. (Lecture Notes in Computer Science, 146)

Salton, G.; Fox, E. A.; Voorhees, E. Advanced feedback methods in information retrieval. Journal of the American Society for Information Science. Vol. 36, No. 3, p. 200-210 (1985)

Salton, G.; Fox, E. A.; Wu, H. Extended Boolean information retrieval. Communications of the ACM. Vol. 26, No. 11, p. 1022-1036 (1983)

Salton, G.; McGill, M. J. Introduction to Modern Information Retrieval. New York, McGraw-Hill, 1983, 448 p. (McGraw-Hill Computer Science Series)

Salton, G.; Voorhees, E.; Fox, E. A. A comparison of two methods for Boolean query relevance feedback. Information Processing & Management. Vol. 20, No. 5/6, p. 637-651 (1984)

Salton, G.; Wu, H. "A term weighting model based on utility theory". Information Retrieval Research. Oddy, R. N. et al., eds. London, Butterworths, 1981, p. 9-22.

Smeaton, A. F.; van Rijsbergen, C. J. The retrieval effects of query expansion on a feedback document retrieval system. The Computer Journal. Vol. 26, No. 3, p. 239-246 (1983)

Srinivasan, P. Intelligent information retrieval using rough set approximations. Information Processing & Management. Vol. 25, No. 4, p. 347-361 (1989)

Srinivasan, P. On generalizing the two-Poisson model. Journal of the American Society for Information Science. Vol. 41, No. 1, p. 61-66 (1990)

Srinivasan, P. The importance of rough approximations for information retrieval. International Journal of Man-Machine Studies. Vol. 34, No. 5, p. 657-671 (1991)

鈴木志元. 確率的文献検索理論の展開. 社会教育学・図書館学研究. No. 15, p. 13-23 (1991)

谷口祥一. 索引語間の関連性を考慮した情報検索モデル. Library and Information Science. No. 28, p. 105-119 (1990)

谷口祥一. "情報検索モデル: その数量的アプローチ". 図書館情報学における数学的方法. 日本図書館学会研究委員会編. 東京, 日外アソシエーツ, 1993 刊行予定, (論集・図書館学研究の歩み, 第 12 集)

Thompson, P. A combination of expert opinion approach to probabilistic information retrieval. Part 1, the conceptual model. Information Processing & Management. Vol. 26, No. 3, p. 371-382 (1990a)

Thompson, P. A combination of expert opinion approach to probabilistic information retrieval. Part 2, mathematical treatment of CEO model 3. Information Processing & Management. Vol. 26, No. 3, p. 383-394 (1990b)

van Rijsbergen, C. J.; Harper, D. J.; Porter, M. F. The selection of good search terms. Information Processing & Management. Vol. 17, No. 2, p. 77-91 (1981)

Willet, P. Recent trends in hierarchic document clustering: a critical review. Information Processing & Management. Vol. 24, No. 5, p. 577-597 (1988)

Wong, S. K. M.; Yao, Y. Y. A probability distribution model for information retrieval. Information Processing & Management. Vol. 25, No. 1, p. 39-53 (1989)

Wong, S. K. M.; Yao, Y. Y. A generalized binary probabilistic independence model. Journal of the American Society for Information Science. Vol. 41, No. 5, p. 324-329 (1990)

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)

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)

Wu, H.; Salton, G. The estimation of term relevance weights using relevance feedback. Journal

of Documentation. Vol. 37, No. 4, p. 194-214 (1981)

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)

Yu, C. T.; Lam, K.; Salton, G. Term weighting in information retrieval using the term precision model. Journal of the Association for Computing Machinery. Vol. 29, No. 1, p. 152-170 (1982)

Yu, C. T.; Meng, W.; Park, S. A framework for effective retrieval. ACM Transactions on Database Systems. Vol. 14, No. 2, p. 147-167 (1989)

Zenner, R. B. R. C.; de Caluwe, R. M. M.; Kerre, E. E. A new approach to information retrieval systems using fuzzy expressions. Fuzzy Sets and Systems. Vol. 17, No. 1, p. 9-22 (1985)