Library and Information Science No. 23 1985

ファジィ集合理論に基づく重み付き文献検索システム

Weighted Document Retrieval Systems Based on Fuzzy Set Theory Approaches

細野 公男 *Kimio Hosono*
高柳 敏子 *Toshiko Takayanagi*
後藤 智範 *Tomonori Gotoh*
原田 隆史 *Takashi Harada*

Résumé

There exist unavoidable drawbacks in the conventional document retrieval systems based on Boolean logic. Every keyword assigned to the documents in the collection is given equal weight irrespective of its actual relatedness to each document. This aspect is also applicable to the search formula where keywords are treated equally, and there is no chance for a searcher to specify the extent of the keyword's significance. In addition, AND operators are excessively influential when narrowing the range of the documents to be retrieved.

In order to get rid of the defects, a variety of techniques have been introduced into retrieval systems such as models using conditional probability theory, vector space, and so forth. Fuzzy set theory approaches are among these and are quite promising for this purpose since human behavior is essentially so fuzzy that it is difficult to be represented by, for example, probability terms.

This paper describes following items concerning fuzzy retrieval systems from this viewpoint.

1) General characteristics of fuzzy retrieval
2) Roles and functions of the weight
3) Calculation formula for the 'Retrieval Status Value'
4) Manipulating methods of the metamorphic functions of Boolean logical operators

Finally, this paper mentions some problems of determining legitimate RSV and e_i formula, especially for handling NOT function, and of selecting adequate Membership Functions, when developing operational fuzzy retrieval systems.

細野 公男：慶應義塾大学文学部図書館・情報学科教授，東京都港区三田 2-15-45
Kimio Hosono: Professor, School of Library and Information Science, Keio University, Mita, Minato-ku, Tokyo.
高柳 敏子：独協大学経済学部助教授，埼玉県草加市栄町 600
Toshiko Takayanagi: Associate Professor, Faculty of Economics, Dokkyo University, Soka-shi, Saitama.
後藤 智範：慶應義塾大学大学院文学研究科図書館・情報学専攻博士課程，東京都港区三田 2-15-45
Tomonori Gotoh: Graduate School of Library and Information Science, Keio University, Mita, Minato-ku, Tokyo.
原田 隆史：慶應義塾大学大学院文学研究科図書館・情報学専攻修士課程，東京都港区三田 2-15-45
Takashi Harada: Graduate School of Library and Information Science, Keio University, Mita, Minato-ku, Tokyo.

ファジィ集合理論に基づく重み付き文献検索システム

I. 現在の情報検索理論の問題点

ブール代数と通常の集合理論に基づいて構築されている現在の情報検索理論では，キーワードと文献や検索質問の概念との関係は，個々のキーワードがこのような概念を表わしているか否かの二値的なとらえ方で規定されている。そのため，例えばあるキーワードが文献Aではその主題を良く表わしているが文献Bではほとんど表現していないというようなとらえ方，つまりキーワードの主題表現力を多値的にとらえることはできない。これは検索式においても同様であり，検索質問を表現するにあたって検索式中の各キーワードの重要度には差がなく，全てのキーワードの重要度は同じである。

しかし，本来キーワードとそれで表わされる各文献の主題概念との関係は，あるかないかの二値的ではなく，主題との関連度の強さに応じて変化すると考えるのが自然である。また，検索質問を構成するキーワード群の中には，他のキーワードに比べて検索者がきわめて重要と考えるものもあれば，付随的なものもあろう。つまり検索式中の各キーワード間の重要さには差異があると考える方が一般的である。

したがって，キーワードと文献との関連度を多値的にとらえ，0.3 あるいは 0.9 のように 0（全く関係なし）から 1（最も強い関係）までの値で表現する方が望ましい。一方，検索式では質問概念に対する各キーワードの重要さの程度に応じて 0 から 1 までの値を割り当てることが考えられる。この種の値は，重み（weight）とよばれる．キーワードと文献との関連度も重みの一種とみなすことができよう。

さらに現在の理論では検索式中での各キーワードの機能的関係は，AND, OR, NOT の論理演算子で規定されているが，この種の演算子が常に検索プロセスを的確に表現できるかどうかにも疑問がある。例えば，t_1 AND t_2 AND... AND t_n のように多くのキーワードが AND 演算子で結合されている場合には，たった 1 つのキーワードが含まれていないために適合文献が検索されないという事態が生じうる。これは AND 演算子が固有にもつ制限性のためであるが，その結果，情報利用者にとってそれが重要度が低いキーワードであったとしても，重要度の高いキーワードの場合と同様に検索もれが生じる。これは，利用者にとってきわて不合理な事態といえよう。

これらは現在の検索理論では回避できない欠点であるが，この種の問題点を解決するために，文献や検索質問中の概念とキーワードとの間の関係を多値的に表現しようとするアプローチがいろいろ考えられている。確率論やファジィ集合理論による重みの概念の導入がその一例である。

そこで本稿では，このような関係をファジィ理論に基づいて構築した考え方のうち，現在のファジィ検索理論に大きな影響を与えている基本的なモデルを取り上げ，その利点，特徴を分析する。

II. 重み付けアプローチの種類とその特徴

文献の主題を表現するために文献に付与されるキーワードおよび検索式中に含まれるキーワードに，重みを導入し検索を行う方法が，現在に至るまでいくつか試みら

れている。重みは，文献に付与されているキーワードおよび検索式中のキーワードの両方に与えられている場合と，どちらか片方のみに与えられている場合との3通りが考えられている。いずれの場合でも，検索結果は検索式に対する蓄積文献の適合度の降順に示されるのが，つまり，文献が適合度の高いものから順番に出力されるのが通例である。この種の出力を適合度順出力 (ranked output) という。

重みの定義，捉え方，考え方はさまざまであり，重みを導入した手法ごとに異なっているが，キーワード相互間での相対的な重要度を示す尺度とする場合と，各キーワードが満たす基準値と考える場合とに大別できよう。後述するように前者は relevance weight，後者は threshold value とよばれるものにそれぞれ該当する。

一方，通常よく使用されている種類の重みとは全く異なるものもある。Angione は，重み付き検索とブール検索とが等価であるとし，両者の相互置換可能性を論じている[1]。その例として，重み付き検索式 [$t_1:2$, $t_2:2$, $t_3:1$, $t_4:1$, 5] (2, 2, 1, 1 は各キーワードの重み，5 は各文献が満たすべき閾値を意味する) は，キーワード t_1, t_2, および t_3 と t_4 のどちらか一方あるいは両方が付与されている文献を検索するとしている。これはキーワードの重みの合計値が5以上となり閾値を満足するからである。したがって，上式はブール論理式 [$t_1 \times t_2 \times (t_3 + t_4)$] と等価であることになる。つまり，ブール検索式と重み付き検索式とは置換可能とする考え方である。しかし，この方法ではキーワードに付与された値は，検索メカニズムにはなんら反映されておらず，単にキーワード間のブール論理関係を数量的に表現したにすぎないため，一般的な重み付き検索の概念にはなじまないといえよう。検索に重みを導入する場合には，それと関連した検索メカニズムを提示すべきであるからである。

重みの概念を導入した方法として，条件付き確率の使用などいくつかあるが，顕著な例は SMART システムで使用されているベクトル空間モデルである[2]。

このシステムでは，文献をキーワードを直交座標軸とする空間上の点 (ベクトル) として表現し，検索式 ($QUERY_j$) と各文献 (DOC_i) との一致度を次式で示す COSINE 尺度で求めている。DOC_i は $TERM_{ik}$, $QUERY_j$ は $QTERM_{ik}$ をそれぞれ要素とするベクトルであり，$TERM_{ik}$ は文献 i におけるキーワード k の重みを，$QTERM_{jk}$ は検索式 j におけるキーワード k (k=1, 2, 3, ..., t) の重要度をそれぞれ表わす。この重みはいずれもキーワードの出現頻度に基づいて算出される。

$$\mathrm{COSINE}\ (DOC_i, QUERY_j) = \frac{\sum_k (TERM_{ik} \cdot QTERM_{jk})}{\sqrt{\sum_k (TERM_{ik})^2 \cdot \sum_k (QTERM_{jk})^2}}$$

なお，このシステムで採用されている重みの概念は，relevance weight とみなすことができよう。ベクトル空間モデルの大きな欠点は，各キーワードの独立性を前提としていることである。しかし，シソーラス，件名標目表などに見られるようにキーワード間には上位，下位，関連などの関係があり，独立ではない。

文献および検索式中の概念とキーワードとの関係を多値的に表現する他のアプローチとして，ファジィ集合理論に基づいて重みを導入する方法が，近年注目されている。この方法はキーワードの独立性を仮定しないこと，他の手法に比較して検索メカニズムの決定に柔軟性があることなどの利点がある。

III. ファジィ集合理論を用いた重み付け

文献全体の集合をDとし，キーワード全体の集合をTとする。D の中で，任意のキーワード $t \in T$ に関連のある文献の部分集合を考えると，任意の文献 $d \in D$ がこの部分集合に属しているか，属していないかは，特性関数 $f(d,t)$ を使用して次のように表現することができる。

$$f(d,t) = \begin{cases} 1 & (d \text{ が } t \text{ に関連のある部分集合に属するとき}) \\ 0 & (d \text{ が } t \text{ に関連のある部分集合に属さないとき}) \end{cases} \quad (1)$$

これは写像表現により次のようにも表わすことができる。

$$f : D \times T \to \{0, 1\} \quad (2)$$

したがって，キーワード t に関連のある文献の部分集合は次のように表現できる。

$$S_t = \{d \mid d \in D,\ f(d, t) = 1\} \quad (3)$$

通常，ブール検索システムではキーワード t により検索した結果として，部分集合 S_t が得られる。すなわち，キーワードが文献 d に付与されているときは $f(d,t)=1$，付与されていないときは $f(d,t)=0$ となる。言い換えるとブール検索システムとは，文献全体の集合から与えられた質問の表わす概念に完全に一致した文献の部分集合を選びだすことである。すなわち，文献の部分集合間の二項演算 (和集合，積集合) と単項演算 (補集合)，そして全集合 D と空集合 ϕ に関するブール代数である。

ここでブール代数の性質をあげておこう。A，B，C をD の任意の部分集合とすると，次の 9 個の性質が成り立つ。

i) 巾等律　$A \cup A = A$, $A \cap A = A$

ii) 交換律　$A \cup B = B \cup A$, $A \cap B = B \cap A$

iii) 結合律　$A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$

iv) 吸収律　$A \cup (A \cap B) = A$, $A \cap (A \cup B) = A$

v) 分配律　$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

vi) 補元律　$A \cup \overline{A} = D$, $A \cap \overline{A} = \phi$

vii) 対合律　$\overline{(\overline{A})} = A$

viii) ド・モルガンの法則　$\overline{A \cup B} = \overline{A} \cap \overline{B}$, $\overline{A \cap B} = \overline{A} \cup \overline{B}$

ix) D, ϕ に関する法則　$\overline{D} = \phi$, $A \cup D = D$, $A \cap D = A$, $A \cup \phi = A$, $A \cap \phi = \phi$

さらに，ブール検索システムはこれら文献の部分集合間の包含関係 $\supseteq$ を順序関係とする特色をもっているのでブール束である。

ここで，キーワード t の付与を二値選択すなわち {0, 1} ではなく区間 [0, 1] の値で与える，つまり，文献 d へのキーワード t の関連の強さ（重み）を区間 [0, 1] の値で表現することにしよう。その結果 (1) は次のように拡張される。

$$f(d, t) = k \quad (0 \leqq k \leqq 1) \qquad (4)$$

このとき，k は d の t との関連の強さを表わす。すなわち，k=1 は t への最も強い関連を示し，k=0 は t への関連が全くないことを示す。

ファジイ集合理論はあいまいな概念を数量化するための一方法であるが，情報検索システムにおいてもブール検索システムを一般化するために，種々の試みがなされている。ファジィ集合論を使用することによって，(2) を下記のように拡張することができる。

$$f : D \times T \to [0, 1] \qquad (5)$$

なお，f(d, t) を特性関数の拡張として，文献 d のキーワード t への関連の強さを示すメンバーシップ関数 (Membership Function: MF) とする。キーワード t に関連のある文献 d の部分集合は，その MF すなわち f(d, t) の大きさによって，部分集合への所属の強さのみが示されるだけであり，集合としての境界が明確でないことからファジィ部分集合と呼ばれる。キーワード t に関連のあるファジィ部分集合を S'_t で示し，次のように表わす。

$$S'_t = \{\langle d, f(d, t)\rangle \mid d \in D\} \qquad (6)$$

ファジィ部分集合の包含関係および演算はすべてその MF すなわち f(d, t) によって示される。S'_{t_1}, S'_{t_2} をキーワード t_1, t_2 と関連のあるファジィ部分集合とすると，すべての $d \in D$ に対して，以下のように定義される。

包含関係　$S'_{t_1} \supseteq S'_{t_2}$ ならば

$$f(d, t_1) \geqq f(d, t_2) \qquad (7)$$

和集合　$S'_{t_1} \cup S'_{t_2}$ の MF は

$$\max[f(d, t_1), f(d, t_2)] \qquad (8)$$

積集合　$S'_{t_1} \cap S'_{t_2}$ の MF は

$$\min[f(d, t_1), f(d, t_2)] \qquad (9)$$

補集合　$\overline{S'_{t_1}}$ の MF は

$$1 - f(d, t_1) \qquad (10)$$

また，MF の性質から f(d, t) について次が成立する。

$$f(d, t_1 \text{ OR } t_2) = \max[f(d, t_1), f(d, t_2)] \qquad (11)$$

$$f(d, t_1 \text{ AND } t_2) = \min[f(d, t_1), f(d, t_2)] \qquad (12)$$

$$f(d, \text{ NOT } t_1) = 1 - f(d, t_1) \qquad (13)$$

ファジィ集合理論の応用により，検索システムは文献全体の集合から質問の表わす概念に，より近いファジィ部分集合を選択するシステムに拡張される。つまり，ファジィ部分集合間の二項演算 (8)，(9) と単項演算 (10) に関する代数である。(8)，(9)，(10) および MF の性質 (11), (12), (13)から，ブール代数の性質のうち，補元律以外はすべて成立することが知られている。このため，(10) の定義は，擬似的補集合 (pseudo-complementation) とも呼ばれる[8]。（しかし，検索システムについては補元律の成立は基本的に重要ではない）したがって，ファジィ部分集合のシステムはブール代数の性質の 1 つである補元律が成り立たないので，ファジィ部分集合間の包含関係 (7) を順序関係とすると，分配束となっている。

例 1　$(t_1 \text{ AND } t_2) \text{ OR } t_3$,　$t_1, t_2, t_3 \in T$

この検索式で検索する場合には，各文献 d についてその MF を次式で求め，MF の値の大きい文献ほど，検索式

によって表現される概念に，より近い文献，すなわち適合文献と考える。

$$f(d, (t_1 \text{ AND } t_2) \text{ OR } t_3) = \max[\min\{f(d, t_1), f(d, t_2)\}, f(d, t_3)]$$

通常このfの値を Retrieval Status Value (RSV) という。検索システムは RSV の大きい文献から適当な数の文献を出力する。例えば閾値を定めておき，その閾値以上の RSV をもつ文献のみを出力する。つまり，RSV は検索式が表わす概念への文献の適合性を示す一尺度である。なお，$f(d, t)$ が (2) で示した通常のブール検索システムでは，検索された文献の $f(d, t)$ すなわち RSV はすべて 1 となる。

例1にみるように，通常のブール検索式においては，検索式中のキーワードの重要性をまったく区別していない。しかし，利用者が重要度（興味の強さ）によって検索式中のキーワードを区別できるとしたら，検索式はより融通性の高い表現となろう。このようなキーワードの重要性を検索式に反映した質問表現を重み付き検索式といい，それに基づくシステムを重み付き検索システムという。

例 2　$\{(t_1, a_1) \text{ AND } (t_2, a_2)\} \text{ OR } (t_3, a_3)$

この検索式において a_1, a_2, a_3 は，各キーワード t_1, t_2, t_3 に利用者が指定した重要度（重み）であり，通常，区間 [0, 1] の値で与える。重み付き検索式での RSV の計算は，各 $f(d, t_i)$ とそのキーワードに与えられた a_i を加味して行なわれる。

任意の文献 d に対する (t_i, a_i) の評価関数を $e(f(d, t_i), a_i)$ で表わす。言い換えると，$e(f(d, t_i), a_i)$ は a_i という重みをもったキーワード t によって表わされた概念に，より近い文献のファジィ部分集合として再定義された MF と考えることができる。したがって，(8), (9), (10) および MF の性質 (11), (12), (13) を応用すると，例2の場合には，任意の文献dに対する RSV の計算を次のように考えることができよう。

$$\begin{aligned} RSV &= [e_1(f(d, t_1), a_1) \text{ AND } e_2(f(d, t_2), a_2)] \\ &\qquad \text{OR } e_3(f(d, t_3), a_3) \\ &= \max[\min\{e_1(f(d, t_1), a_1), e_2(f(d, t_2), a_2)\}, \\ &\qquad e_3(f(d, t_3), a_3)] \end{aligned}$$

なお，$e(f(d, t), a)$ は，(5) と同様に次のように表わすことができる。

$$e : (D \times T) \times [0, 1] \to [0, 1] \qquad (14)$$

一般に，重みが増加すればeの値が増加すると考えるのが妥当であろう。したがって，e_1 の計算において，重みを $f(d, t_1)$ の係数として用いるアプローチが一般的である。

このように $e(f(d, t), a)$ を，aという重みをもったキーワードtによって表わされる概念に，より近い文献のファジィ部分集合として再定義された MF と考え，(6) と同じようにこれらファジイ部分集合を S_t'' で示し，次のように表わす。

$$S_t'' = \{\langle d, e(f(d, t), a)\rangle | d \in D\} \qquad (15)$$

次に，MF として $f(d, t)$ をもつファジィ部分集合 S_t' の集まり $\{S_t'\}$ と，MF として $e(f(d, t), a)$ によって再定義されたファジィ部分集合 S_t'' の集まり $\{S_t''\}$ との間の対応関係 E を考えてみよう。

$$E : \{S_t'\} \to \{S_t''\} \qquad (16)$$

前にも述べたように，ファジィ部分集合の集まり $\{S_t'\}$ は順序関係を包含関係とした分配束である。このことから，E が束の準同型写像となっていれば，演算 $\cup$ と $\cap$ を共に保持する写像となる。すなわち，任意の S_{t1}', S_{t2}' に対して次が成り立つ。

$$E(S_{t1}' \cup S_{t2}') = E(S_{t1}') \cup E(S_{t2}') \qquad (17)$$

$$E(S_{t1}' \cap S_{t2}') = E(S_{t1}') \cap E(S_{t2}') \qquad (18)$$

同様に順序関係つまり包含関係も保持され，次のことがいえる。

$$S_{t1}' \supseteq S_{t2}' \text{ ならば，} E(S_{t1}') \supseteq E(S_{t2}') \qquad (19)$$

言い換えると $\{S_t''\}$ すなわち，重み付き検索システムも分配束となる。Buell はこの順序関係から $f(d, t_i)$ が単調増加ならば，$e(f(d, t_i), a_i)$ も a_i を固定したとき単調増加となり，このときに限ってEは束の準同型写像となるとしている[4]。

重み付き検索システムの場合，一般に，論理的には等しいが異なった表現で示される検索式間では文献に対する適合度 RSV の同等性を保証することが困難とされている[5,6]。これを保証するための前提の一つとして E が束の準同型写像であることが望ましい。

このような写像 E のもとで，$\{S_t''\}$ の任意の S_{t1}'', S_{t2}'' に対して (7), (8), (9) はそのまま定義できよう。また，補集合の定義として (10) をそのまま使用するならば，その補集合 $\overline{S_t''}$ の MF は，$1-e(f(d, t), a)$ のように定義される。一方，任意の S_t' の補集合の写像Eによる像 $E(\overline{S_t'})$ と考えるならば，その MF は，$e(1-f(d, t), a)$

のように定義される。一般に，写像 E については補集合に関する演算は保持されないので次のようになる。

$$\overline{S_t''}=\overline{E(S_t')}\neq E(\overline{S_t'}) \tag{20}$$

しかし，Buell も指摘しているように，$\overline{E(S_t')}=E(\overline{S_t'})$ と定義することも可能であるが，数学的には制限が厳しすぎ，情報検索システムにおいてはこのような制限はむしろ必要としない[4]。上記のいずれの定義を選ぶにせよ，補集合の定義は演算の順序（補集合の演算と重み a の処理のどちらを先に行うか）の問題を引き起こすだけでなく，現実の情報検索システムではどちらがより妥当であるかの問題をも含んでいる。

IV. 代表的な重み付き検索システム

前章で重み付き検索システムの基本的な概念，および数学的基礎について述べた。文献とキーワードとの関係に対する考え方は研究者間で共通している。一方，検索式中のキーワードに付与される重みの考え方は研究者によって異なっている[7]。特に，RSV の計算方法，およびその基礎となる検索式中の各キーワードに付与される重みの考え方，すなわち重みの解釈と機能が研究者によって異なっており，様々なモデルが提案されている。

ここでは現在までに提案されている代表的なモデルについて，RSV の計算方法および重みの解釈と機能の概要を説明する。

A. Radecki のモデル[8,9]

Radecki のモデルは II. で説明された適合度順出力システムの発展型と考えてよい。

〈RSV の計算方法〉

各文献の RSV は次のように定義される。

$$RSV=\begin{cases} f(d,t_1) & (f(d,t_1)\geqq a \text{ のとき}) \\ 0 & (f(d,t_1)<a \text{ のとき}) \end{cases} \tag{21}$$

したがって，$f(d,t_1)\geqq a$ である文献だけが検索式に対応する対象文献集合となる。最終的な適合文献は上式を満足する文献集合に対し，(11), (12), (13) 式に基づいて AND, OR, NOT を計算して求められる。

〈重みの解釈と機能〉

重みは検索式中の各用語に対して与えられる同一の閾値である。Radecki のモデルでは，重みは閾値と解釈できるが，検索式中の各キーワードに独立に異なった重みを与えることはできない。

このモデルでは重みは検索対象とする文献集合を限定するものである。検索対象は a=0 のときは重みを用いない場合に得られる文献集合と同一である。a が 1 に近づくほど限定力は高くなるので，a=1 の時には，a=0 の時と比較して適合文献の数は著しく減少するか，もしくは適合文献が存在しないこともある。

B. Waller と Kraft のモデル[10]

Waller と Kraft は，検索式を構成するキーワードにつけられた重みをキーワード間の相対的重要性とするモデルを次のようにまとめている。

〈RSV の計算方法〉

a) 1 つのキーワード（単一語）t_i による検索の場合

$$RSV=e_i=e\{f(d,t_i),a_i\}=a_i\cdot f(d,t_i) \tag{22}$$

b) OR 演算子によって結合されたキーワードによる検索の場合

検索式 (t_1, a_1) OR (t_2, a_2) に対し，a) に基づいて 2 通りのモデルが考えられている。

i) hard OR : $RSV=\max(e_1, e_2)$ (23)

ii) softer OR : $RSV=e_1+e_2-e_1\cdot e_2$ (24)

ブール検索における OR 演算子が持つ役割をファジイ検索で表現する場合には，max を用いるのが一般的である。しかし，max 演算では RSV に対する e_1, e_2 のうち小さい方の値が無視されるので，(24) のように e_1, e_2 の両者を RSV の評価に組み入れることも考えられる。

c) AND 演算子によって結合されたキーワードによる検索の場合

検索式 (t_1, a_1) AND (t_2, a_2) に対し，a) に基づいて 3 通りのモデルが考えられている。

i) hard AND: $RSV=\min(e_1\cdot e_2)$ (25)

ii) softer(firm) AND: $RSV=e_1\cdot e_2$ (26)

iii) soft AND : $RSV=(e_1+e_2)/(a_1+a_2)$ (27)

AND 演算子の場合では，(25) のように min を使用するのが一般的である。しかし min 演算では e_1, e_2 のうち大きい方の値が無視されるので，(26), (27) のように e_1, e_2 の両者を RSV の計算に組み込む方法も考えられる。RSV の計算においては，(26) よりも (27) の方が値の小さい e_i の影響をより強くうける。

d) NOT の場合

検索式 NOT (t_1, a_1) に対しては，3 通りのモデルが考えられている。

i) $RSV=a_1\cdot(1-f(d,t_1))=a_1-a_1\cdot f(d,t_1)$ (28)

ii) $RSV=-a_1\cdot f(d,t_1)$ (29)

iii) $RSV=(1-f(d,t_1))^{a_1}$ (30)

なお，(30) の場合には，$e_1=f(d,t_1)^{a_1}$ である。(29) は (28) の1つの変形として提案されている。しかし，(28) では a_1 の増加に比例して RSV も増加するのに対し，(29) では逆に減少する。(30) は (28) と同様に a が増加するにつれて RSV が増加する。

また彼らは AND, OR, NOT だけではなく，独自のアプローチとして AND と OR を結びつけた新しい演算子 ANDOR を提唱している。

e) ANDOR の場合

$$RSV=z\cdot\min(e_1,e_2)+(1-z)\cdot\max(e_1,e_2) \quad (31)$$

(31) で z は限定係数 (coefficient of restriction) とよばれ，その値の大きさによって以下の2つの機能を実現している。

- z が1に近づく：AND 演算子の機能がより強くなる，すなわち適合文献の数を減少させる。
- z が0に近づく：OR 演算子の機能がより強くなる，すなわち適合文献の数を増加させる。

このように，新しい演算子 ANDOR は係数 z の大きさを変化させることによって，AND 演算子と OR 演算子の両者が持つ相反する機能を1つの演算子で実現するものである。

〈重みの解釈と機能〉

ここでの重みは検索式中のキーワード間の相対的重要性を表わす尺度である。検索式中に含まれるすべてのキーワードの重みが1の場合には，検索式の RSV はブール代数に基づく検索システムと一致する。一方，検索式中のあるキーワードの重みが0の場合には問題を生じる。(25), (26) からわかるように，(t, 0) の時には，これが AND演算子として hard AND または softer AND で結合される場合には，もう一方の語の重みの値にかかわらず RSV は重み0のキーワードに完全に依存してしまうからである。

C. Bookstein のモデル[11,12]

Bookstein は，検索式を構成するキーワードに，他のキーワードとの相対的な重要性を反映させた重みを付与する検索モデルを提案している。彼のモデルでは，各キーワードについての評価関数 e の計算式が，キーワード間の関係を規定する論理演算子に依存するという特徴を持っている。

〈RSV の計算方法〉

a) 1つのキーワード (単一語) による検索の場合

重みが0または1に近づくときに，従来のブール検索システムと等価になるように評価関数が提案されている。検索式 (t_i, a_i) の RSV は (22) 式で定義される。

b) OR 演算子によって結合されたキーワードによる検索の場合

OR 演算子で結合されている個々のキーワードに関する評価関数 e の計算方式は，単一語の場合と同様であり，検索式 (t_1, a_1) OR (t_2, a_2) の RSV は，(23) 式で求められる。

一方の語の重みが0の場合，上記の計算によって他方の語だけを用いた式と結果は一致する。したがって，この場合，OR 演算における重み0は，その語が用いられないことを意味する。

c) AND 演算子によって結合されたキーワードによる検索の場合

OR 演算子によって結ばれる検索式の場合とは逆に，AND 演算子で結ばれている検索式においては，重みが0に近づくにつれて限定性を減少させるようなモデルを提案している。これを実現するために，AND で結合されている個々のキーワードの評価関数を次のように定義している。なお，検索式 (t_1, a_1) AND (t_2, a_2) の RSV は (25) 式で求められる。

$$e_i=e\{f(d,t_i),a_i\}=\begin{cases}1/a_i\cdot f(d,t_i) & (1/a_i\cdot f(d,t_i)\leqq 1 \text{ のとき})\\ 1 & (1/a_i\cdot f(d,t_i)>1 \text{ のとき})\end{cases} \quad (32)$$

d) NOT 演算子の場合

NOT 演算子は，次の2つの相反する機能を持ち，それぞれについて評価関数が定義されている。

1) 積極的機能

NOT 演算子を用いて検索される文献の数を増加させる機能。このような機能は，NOT 演算子が OR 演算子で結合される場合に用いられ，このとき評価関数は以下のように定義される。

$$e_i=e\{f(d,t_i),a_i\}=\begin{cases}1-1/a_i\cdot f(d,t_i) & (1/a_i\cdot f(d,t_i)\leqq 1 \text{ のとき})\\ 0 & (1/a_i\cdot f(d,t_i)>1 \text{ のとき})\end{cases} \quad (33)$$

2) 消極的機能

NOT 演算子を用いて検索される文献の数を減少させる機能。通常の NOT 演算子の機能がこれに相当する。

AND 演算子で結合される場合に用いられ，このとき評価関数は以下のように定義される。

$$e_i=e\{f(d, t_i), a_i\}=1-a_i\cdot f(d, t_i) \quad (34)$$

〈重みの解釈と機能〉

他の研究者と異なり，Bookstein のモデルでは，検索式中の個々のキーワードがいづれの演算子と結合されていても，重みを1に近づけることが，各演算子の機能の強化，0 に近づけることが演算子の機能の低下をそれぞれ意味する。

D. Kantor のモデル[5)]

Kantor は，文献検索には次に示すような2種類の目的があることを示して，文献検索を両者の側面を合わせ持つものと考えている。それに基づいて，他の研究者には見られない重み付き検索システムのモデルを提案している。

a）集中的探索 (focused browsing)

確固とした意図のもとで行なわれる探索で，通常の文献検索はこの側面が強い。

b）非集中的探索 (unfocused browsing)

目的なしに行なわれる漫然とした探索で，通常の browsing はこの側面が強い。この場合，検索式の文献識別力はないといえる。

〈RSV の計算方法〉

Kantor は RSV の計算に上述の考え方を導入している。そして，単一語および単一語の否定に対する検索式の RSV は，上記の両者を表わす各項の一次結合として，れぞれ (35), (36) のように定義されている。

$$RSV=a\cdot f(d, t)+(1-a)\cdot V \quad (35)$$

$$RSV=a\cdot\{1-f(d, t)\}+(1-a)\cdot V \quad (36)$$

なお，V は非集中的探索を表わす定数であり，文献の識別力を最小にするために通常 1/2 が与えられている。

a）OR 演算子によって結合されたキーワードによる検索の場合

検索式 (t_1, a_1) OR (t_2, a_2) の RSV は次式で求められる。

$$RSV=a_1\cdot a_2\cdot\max(f_1, f_2)+a_1\cdot(1-a_2)\cdot f_1+a_2\cdot(1-a_1)\cdot f_2+(1-a_1)\cdot(1-a_2)\cdot V \quad (37)$$

ただし，$f_i=f(d, t_i)$ である。

b）AND 演算子によって結合されたキーワードによる検索の場合

検索式 (t_1, a_1) AND (t_2, a_2) の RSV は次式で求められる。

$$RSV=a_1\cdot a_2\cdot\min(f_1, f_2)+a_1\cdot(1-a_2)f_1+a_2\cdot(1-a_1)\cdot f_2+(1-a_1)\cdot(1-a_2)\cdot V \quad (38)$$

ただし，$f_i=f(d, t_i)$ である。

〈重みの解釈と機能〉

(37), (38) で示した関数は文献とキーワードとの関連度を表わす帰属度関数と定数Vとの一次結合である。このとき，重み a と RSV の関係は以下のようになる。

$a=1$: $RSV=f(d, t)$

$a=0$: $RSV=1/2$ （検索式に対する文献の識別力がなくなる）

このことから上記のモデルにおいて，重みは検索式中のキーワードに対する利用者の信頼性の度合と考えてよいであろう。重みが1に近づくことは，そのキーワードに対する信頼性が増すことを意味し，また0に近づくことは信頼性が低くなることを意味している。

E. Buell と Kraft のモデル[13,14)]

Buell と Kraft は，検索式を構成する各キーワードに付与されている重みを閾値と解釈するモデルを提案している。

〈RSV の計算方法〉

a）1つのキーワード（単一語）による検索の場合

$$e_i=e\{f(d, t_i), a_i\}=\begin{cases}(1+a_i)/4+a_i/2\cdot\{f(d, t_i)-a_i\}/(1-a_i) & (f(d, t_i)\geqq a_i \text{ のとき})\\ (1+a_i)/4\cdot f(d, t_i)/a_i & (f(d, t_i)<a_i \text{ のとき})\end{cases} \quad (39)$$

なお，$(f(d, t_i)-a_i)/(1-a_i)$ は，$f(d, t_i)$ の最大値と a_i との差に占める現実の $f(d, t_i)$ と a_i との差の割合を示す。また，$f(d, t_i)/a_i$ は $f(d, t_i)$ が閾値 a_i をどの程度満足するかを示し，$(1+a_i)/4$ は $f(d, t_i)$ が大きい閾値を満足する場合と小さい閾値を満足する場合との識別を可能とするための項である。

b）OR 演算子によって結合されたキーワードによる検索の場合

検索式 (t_1, a_1) OR (t_2, a_2) の RSV は (23) で求める。

c）AND 演算子によって結合されたキーワードによる検索の場合

検索式 (t_1, a_1) AND (t_2, a_2) の RSV は (25) で求

める。

〈重みの解釈と機能〉

Buell と kraft は，閾値は検索式中の キーワードの f(d, t) が達すべき度合を示す値であるとしている。彼らは Radecki と異なり，閾値の概念を拡張して評価関数に組み込んでいる。(39) にはこのことが反映されている。

(39) の上式と下式は意味的に異なっている。$a_i=0.01$ のように閾値が小さく，$f(d, t_i)$ が閾値を容易に越え得る場合には，$f(d, t_i)$ が閾値をどの程度越えているかを比率で示すのが上式である。一方，$f(d, t_i)$ が閾値 a_i をどの程度満足しているかを表わすのが下式である。つまり $a_i=0.9$ のように閾値が大きい場合には，$f(d, t_i)$ が閾値を越えることが一般に困難であるため，評価関数 e_i は a_i に対する $f(d, t_i)$ の満足の比率を表わす形で示されている。

V. ファジィ検索モデルの特徴

A. Relevance Weight と Threshold Value

検索者の情報要求を表わす検索式中の各キーワードの重要性は一般に異なり，各キーワードは情報要求に対してそれぞれ異なった度合の重要性を持っていると考えられる。この重要性の度合を示すのが重みである。

検索式中の各キーワードに付与される重みに対する考え方は，IV. で示したように研究者によって様々であるが，概して Relevance Weight と Threshold Value の2つに分けられる。前者のアプローチをとるのは，Waller と Kraft[10]，Bookstein[11,12,15]，Kantor[5] であり，後者の考え方を採用するのが Radecki[8,9]，Buell と Kraft[13,14] である。

1. Relevance Weight (RW)

重みを RW とみなすアプローチは，各キーワードに通常 [0, 1] の値を付与することによって，情報要求に対する個々のキーワードの重要度を表現しようとするものである。この考え方をとる研究者によって提案されている単一語の評価関数は，(22) が基本となっている。これは情報要求に対するキーワードの重要度を表わすRW すなわち重み a_i に $f(d, t_i)$ を乗ずることによって，各文献の評価を行なうものである。

一方，RW と AND, OR, NOT の各演算子が持つ機能との関係は，RW の考え方をとる研究者間で異なっている。例えば，Waller と Kraft, Kantor は RW と各演算子の機能とは独立であると考え，キーワードを結合させている演算子の種類にかかわらず単一語の評価関数は同一であるとしている。つまり，単一語の評価関数 e_i は，常に Waller と Kraft では (22)，Kantor では (35) となる。これに対して Bookstein は，評価関数を演算子の種類に関連させて定義し，さらに RW の値の大きさが演算子の機能の強弱を変化させるという考え方をとっている。つまり，RW の値が小さいと演算子の機能は低くなり，逆に大きいと高くなる。例えば，AND 演算子は本来適合文献の範囲を限定する機能を持っているが，もし RW が小さい場合には，e_i が大きくなり限定力が弱まるように工夫している。(32) にはこの考え方が生かされている。そしてこのアプローチを RSV に反映させるために，演算子ごとに単一語の評価関数を定義している。

2. Threshold Value (TV)

RW 方式では，キーワードに付与される重みがキーワード間の重要度の相対的な差異を示すのに対し，TV 方式では，各キーワードに与えられる閾値を表わすので，RW の場合とは異なり，重みは互いに独立となる。したがって，$f(d, t_i)$ が閾値 a_i を越えるか否かに基づいて，2通りの評価関数が定義されている。これは Buell と Kraft のモデルから明らかなように，閾値を満足するか否かによって，検索環境が質的に異なることを意味する。ここに RW と TV との大きな差異がみられる。また，重みが演算子と独立である点も TV 方式の特色といえよう。

重みを閾値と考えるアプローチの最も単純なモデルは，次のものである。

$$e=\begin{cases}1 & (f(d, t)\geqq a \text{ のとき})\\ 0 & (f(d, t)<a \text{ のとき})\end{cases} \qquad (40)$$

これは閾値 a を基準としたブール検索システムと等価であると考えられる。これをファジィ検索に拡張したものとして Radecki のモデルがあり，そのモデルでは (21) から明らかなように閾値 a を越える $f(d, t)$ を持つ文献だけが検索対象となっている。$f(d, t)$ と閾値 a との差の程度は，両者ともに評価関数には反映されていない。一方，IV. の E. で既述した Buell と Kraft のモデルでは，この差が反映されている。

B. ファジィ検索システムの問題点

現実に重み付き検索システムを構築する観点からみた場合，ファジィ検索はキーワード相互の独立性を必要と

しないし，キーワードの単位が単一語でも複合語でもよいこと，さらにそれぞれの環境にあわせて f(d, t) や評価関数 e の形態を選択することが可能であることなどの利点がある。また，AND 演算子の使用によって生じるブール検索固有の限定性は，ファジィ検索では一応回避しうる。これは，ファジィ検索で AND 演算子の代りに使用される min 演算子では，重みの使用により演算子の効果を軽減することが可能だからである。

しかし，このアプローチに問題がないわけではない。特に，これまでの研究では，f(d, t)，e が抽象的なレベルでしかとらえられていないため，現実のシステムではこれらの関数の妥当性をいかに決定するかが大きな問題となる。f(d, t) に関しては，例えば文献集合中における各キーワードの出現頻度を利用することが考えられる。一方，e の妥当性に関しては準拠すべきものが現在のところ存在しないので，例えばeをブラックボックスとして，入力される検索式と検索結果の集合から望ましいeを決定せざるを得ない。また，複数の論理演算子で結ばれる複雑な検索式の RSV を求めるために提案されている種々の式の有効性を吟味する必要もある。いずれにせよ f(d, t), e, RSV の関数の決定には多くの実験を必要としよう。

さらにファジィ検索に限らず重み付き検索システム全体について，重み 0 の意味をどのように考えるか，また，ブール検索の NOT 演算子に対応する機能をどのような評価関数で表現するかという問題がある。現在のところ既存のモデルでは，重み 0 のとらえ方に十分検討がなされているとはいいがたい。また，NOT 演算子の機能の理由づけを含め，その評価関数を明確に表現しているものはない。例えば，(28), (30) と (29) とでは，重みの影響が逆に働いている。しかし実用システムの構築にあたっては，NOT 演算子を表現する有効な評価関数を決定することが必要になろう。

VI. おわりに

現在ファジィ集合理論を導入した情報検索システムが実質上存在していない理由の 1 つは，キーワードと文献との関連度を示すメンバーシップ関数および RSV の決定・選択が容易ではないからである。またファジィ理論の情報検索への応用は歴史が浅く，その分析・評価もあまり行われていないこともその理由であろう。その結果 Robertson のように，ファジィ集合理論ではなく確率論を使用すべきであるとの意見もみられる[16]。

しかし，人間の情報検索行動や文献の主題認識行動に伴うあいまい性は，自然現象にみられる不確実性を主として記述する手法として確立した確率論あるいは統計的方法では，表現しにくい面が多いことは明らかであろう[17]。したがって，本稿ではあつかっていないが，確率論よりも広範にあいまいな事象を表現できる可能性理論 (Possibility Theory) をも説明しうるファジィ理論の導入は，解決せねばならない点を含むにせよ，極めて有意義であると思われる。

1) Angione, P. V. "On the equivalence of Boolean and weighted searching based on the convertibility of query forms," Journal of the American Society for Information Science, Vol. 26, No. 2, p. 112-124 (1975)
2) Salton, G. and McGill, M. J. Introduction to modern information retrieval. McGraw-Hill, 1983. 448 p.
3) Kaufmann, A. Introduction to the theory of fuzzy subsets. Vol. 1. London, Academic Press, 1975. 416 p.
4) 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)
5) Kantor, P. B. "The logic of weighted queries," IEEE Transaction Systems; Man and Cybernetics, Vol. SMC-11, No. 12, p. 816-821 (1981)
6) Buell, D. A. "A general model of query processing in information retrieval systems," Information Processing & Management, Vol. 17, No. 5, p. 249-262 (1981)
7) Kraft, D. H. and Buell, D. A. "Fuzzy sets and generalized Boolean retrieval systems," International Journal of Man-machine Studies. Vol. 19, No. 1, p. 45-56 (1983)
8) Radecki, T. "Fuzzy set theoretical approach to document retrieval," Information Processing & Management, Vol. 15, No. 5, p. 247-259 (1979)
9) Radecki, T. "Generalized Boolean methods of information retrieval," International Journal of Man-machine Studies, Vol. 18, No. 5, p. 407-439 (1983)
10) Waller, W. G. and Kraft, D. H. "A mathematical model of a weighted Boolean retrieval systems," Information Processing & Management, Vol. 15, No. 5, p. 235-245 (1979)
11) 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)
12) 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)
13) Buell, D. A. and Kraft, D. H. "Threshold values and Boolean retrieval systems," Information Processing & Management, Vol. 17, No. 3, p. 127-136 (1981)
14) Buell, D. A. and 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 (1980)
15) Bookstein, A. "Brief communications: On the perils of merging Boolean and weighted retrieval systems," Journal of the American Society for Information Science, Vol. 29, No. 2, p. 156-158 (1978)
16) Robertson, S. E. "On the nature of fuzz: A diatribe," Journal of the American Society for Information Science, Vol. 29, No. 6, p. 304-307 (1978) しかし, Robertson が f(d, t) の選択, RSV の求め方, 重みの解釈などについて十分検討しているとはいい難く, 提示した例も情報検索の観点からは妥当とはいえない.
17) Cerny, B. A. "A reply to Robertson's diatribe on the nature of fuzz," Journal of the American Society for Information Science, Vol. 30, No. 6, p. 356-357 (1979)