情報検索における近年の理論的動向

後　藤　智　範

1．はじめに

1950年代のMoores,[1] Luhn[2]らの研究を端緒として始まった情報検索（*Information Retrieval*）の研究は，1970年代に入りオンライン情報検索システムの実用化に至り，図書館情報学において主要な研究分野となるに至った。情報検索における研究テーマについて過去30年間を概括すると以下のようになる。

50年代－70年代中頃

- 索引/抄録に関する問題
- 情報検索システムの開発，実験
- 自動索引/自動抄録

70年代中頃－現在

- 自動索引/自動抄録
- オンライン情報検索システムに関する問題
- 情報検索の理論
- オンライン情報検索システムのユーザーインターフェース

要約すると60年代の研究は，コンピューターでのオンライン検索システムの実働化に主眼がおかれており，70年代ではこれに関連する問題の解決に移っていったと言える。特に80年代に入り，システムの利用者が急増するに連れ，現在の情報検索システムがもつ多くの問題点が指摘されるようになった。問題点のいくつかは，情報検索システムが依拠している検索理論に起因していることが明らかにされるようになった。

本稿は，現在の情報検索システムが基礎としているブール代数に基づく検索理論の問題点を明らかにし，最近注目されつつある2つの検索理輪の概要および，その問題点を紹介する。

2．従来の情報検索理論の問題点

現在使用できる商用のオンライン情報検索システムは，全てその理論的基礎をブール代数（*Boolean Logic*）においている。ブール代数に基づく情報検索システム（ブール情報検索システム）は，以前から様々な問題点が指摘されており,[3] 以下のようにまとめることができる。

(1) キーワードは，文献に一度付与されると完全に固定したものとなり，どのような検索状況においても修正することはできない。

(2) 利用者が自分の情報要求をブール検索式で十分に表現することは非常に困難である。また，情報を検索式として正確に表現すると，検索式は必然的に複雑で長くなり，検索式の入力は非常に煩わしいものとなる。

(3) ブール検索システムは多くの利用者を驚かせ，惑わせる結果をもたらす。例えば，a OR b OR c...OR z という検索式を考えてみよう。ブール検索システムは，これらの用語のうち１つだけでも索引されている文献は，あたかもこれらの用語全てが索引されている文献と同じように適合文献とみなされる。同様な問題は，a AND b AND c...AND z についても当てはまる。ある文献が，a–z のうち索引語として１つでも欠けていれば，１だけしか付与されない文献と同様に，適合文献とみなされない。

(4) 索引者は，各文献にどのキーワードを付与するかしないかを決定するだけで，付与すべきキーワードの中で，どのキーワードがより重要であるかを指示することはできない。

(5) 利用者の立場からみると，検索の入力方式は過度に形式的であり，柔軟性に欠ける。利用者は a AND b という検索式をたてることはできるが，この検索式の中でキーワードａの方がキーワードｂよりも重要である，ということは表現することはできない。

(6) 同様な問題は適合文献の出力形式についてもあてはまる。出力すべき適合文献の中で，どの文献がより重要であるかは，ブール検索システムは指示することはできない。つまり，ブール検索システムは，検索質問に対して文献が適合しているか，していないかだけを判断する。

(1)および(2)の問題は，検索システムの機能の問題であり，とくに(6)の問題は，前節に挙げたユーザー・インターフェースの問題として独立した研究テーマとして取り上げられている。(3)の問題は，ブール代数の論理演算子の機能と，人間が日常使用している論理とのギャップにより生ずる問題であり，検索システムがブール代数に基づいている限りこの問題は不可避である。(4)および(5)の問題は，ブール検索システムが通常の集合論に基づいているため実質的に解決不能である。(6)の問題も，(4)および(5)の問題と同様の理由により本質的な解決は望めない。以上，(1)—(4)の問題は，検索システムが依拠している検索理論に深くかかわっている問題で，ブール検索システムでは解決不可能であると言ってよい。

確率検索（*Probabilistic Retrieval*）の概念は，60年代初頭，すでに Maron and Kuhns[4] によって提案されている。確率検索理論は，ブール検索システムとは異なった観点に立ち，確率・統計理論を基礎に，自動索引（*Automatic Indexing*）の研究と平行して研究されてきた。今日まで多くの実験システムが試作され，その有効性が確かめられている。

一方，ファジイ検索理論は，上記のブール検索システムの問題点が指摘され始めた70年代後期に Radecki[5] によって最初に提唱された。ファジイ検索理論はブール検索システムの問題点を克服しつつ，ブール論理関係をある程度維持できるという点で有望視されている。

3．確率検索理論

文献検索システムは，どんなシステムでも，どの文献を情報要求に対する適合文献（*Relevant Document*）とするかについて，文献から手に入る情報を操作することによって，情報要求に対する適合文献を出力する。しかしながら，完全な検索機構というものは考えられず，検索された文献のいくつかは不適合文献であり（*Irrelevant Document*），検索されなかった文献の中に適合文献が含まれている。

確率検索理論は，このような情報要求に対する適合文献の不確実性（*Uncertainity*）をはっきり認めることを前提に，理論を展開している。

3．1　確率検索理論の概要

確率検索理論は，文献の主題内容を表現する索引語に対して，不確実性の数量的尺度を表す重みを付与することから出発する。この重みは，索引語の出現確率から求められる。次に，各文献の索引語のもつ確率に基づいて検索質問を構成している索引語と各文献との類似度を計算する。類似度に域値（*Threshold Value*）を設定し，類似度が域値以上の文献を適合文献とし，類似度の高い文献から順に出力する。この場合，索引語は索引作業によって各文献に付与されるものである必要はなく，むしろ文献（実際には文献そのものではなく文献の標題および抄録）中の重要語を索引語として利用する。

上述のように，確率検索では文献中の語をキーワードとして用いるため，70年代は以下の研究に主眼が置かれた。

(1)文献集合中からのキーワードの描出方法

(2)重み付けアルゴリズム

Luhn[6] によって始められて以来，語の出現頻度に基づいてキーワードを描出する研究は自動索引として，60年代に多くの研究が行われた。この研究成果に基づいて，様々なキーワード描出のモデルが開発された。例えば，Bookstein と Swanson[7] は，キーワードの出現は適合文献集合および非適合文献集合で，それぞれポアッソン分布に従うという2-ポアッソン・モデルを提案した。類似のモデルは Robertson と Sparck Jones[8] によっても提案されている。

上記のような方法で描出されたキーワードに対して，重みを付与するアルゴリスムが，自動索引の研究者によって提案された。例えば，先の Bookstein と Swanson[7] は次のような重み付けを提案している。

$$n \log (r/r')$$

n：キーワードの出現頻度

r：キーワードの適合文献中における期待出現頻度

r'：キーワードの非適合文献中における期待出現頻度

一方，Robertson と Sparck Jones[8] は次のような重みを提案している。

$$\log (p/1-p) \;/\; (p'/1-p')$$

p：キーワードの適合文献中における出現確率

p'：キーワードの非適合文献中における出現確率

確率検索は，文献中におけるキーワードの出現確率を基礎としているが，ほとんどのモデルは，ある語の出現は他の語の出現とは独立しているという仮定に立っている。この仮説は，独立仮説（*Independent assumption*）と呼ばれるが，現実とは矛盾するものもある。したがって，このような仮説に基づいて推定された語の出現確率は，不正確にならざるを得ない。語の出現確率の推定には，問題が残されている。

3．2　確率検索理論の利点/欠点

確率検索理論について，現在までに指摘されている利点および欠点は，以下のようにまとめられる。

〈利点〉

(1) ブール代数の演算子を用いる煩わしさ，およびこれらを正しく使用するための困難さはない。

(2) 出力文献は適合度順に出力される。順位付は自然なもので，しかも重みに鋭敏に反応する。

(3) 自然語による検索質問が前提であるため，とくにユーザー・インターフェースを開発する必要はない。

〈欠点〉

(1) ブール代数の論理関係（AND, OR, NOT）は失われる。したがって，"a AND b", "a OR b" を意図している検索質問は識別されえない。

(2) 同様に，"～を除外する" というブール代数の NOT 演算を意図している検索質問は無視されるため，そのような検索質問に対する適合文献にはノイズが増える。

(3) キーワードの出現確率の正確な推定方法，および重み付けアルゴリスムに関して，研究者間で一致が見られない。

(4) 検索機構は複雑になり，したがって応答時間が長くなる。

4．ファジイ検索理論

伝統的なブール検索システムの枠内で，キーワード間の重要性の相違を検索式に反映させる（検索式中の各キーワードに重みを付与する）ことを目指す研究が70年代中頃 Angione[9] によって行われるようになった。通常，このような検索システムは，重み付け検索システム（*Weighted Retrieval System*）と呼ばれる。このモデルでは，キーワードに付与された重みは実際には検索機構には何ら反映されず，単にキーワード間の論理関係を数量的に表現したにすぎない。したがって，このモデルでは前節で挙げたブール検索システムの問題点を何ら解決していない。検索式中のキーワード，および文献に付与される各キーワードに重みを付与するこ

とを可能とし，しかも従来のブール検索システムが持っている代数的特性を保持するために，Zadeh 教授によって提案されたファジイ集合論[10]を情報検索に応用しようとする研究が行われるようになった。

4．1　ファジイ検索理論の概要

通常の集合論では，対象はある集合に含まれるか含まれないかであるが，ファジイ集合論では対象がある集合に含まれる度合，すなわち帰属の度合（*Degree of membership*）が認められる。帰属の度合は0から1の間の連続値で表される。通常の集合論で，集合に含まれる対象および含まれない対象はファジイ集合論では，それぞれ帰属度1，帰属度0をもつことになる。こうして全ての対象がある集合の帰属度をもつことになり，帰属度関数（*Membership Function: mf*）が定義される。ブール代数における和，積，差，の論理演算はそれぞれ，帰属度関数の max, min, 1−mf で定義される。

情報検索の文脈では，文献のファジイ集合は個々のキーワードに関係付けられる。あるキーワードに対して，個々の文献がそのキーワードに関連している度合が，帰属度関数として定義される。あるキーワードに対する文献のファジイ集合は，索引作業中に作られる。ファジイ検索システムにおいて，索引者は単に文献にキーワードを付与するだけでなく，それが文献にどの程度関連しているか（帰属度）を指示する。例えば，その語がある文献の中心主題を表していれば帰属度1を与え，別の文献では主題とほとんど関係無い場合には0.1（例えば）を与える。従来の索引作業において付与されない索引語は帰属度0を与える。このようにして全ての文献が索引されると，帰属度関数が個々のキーワードに対して実質的に定義される。情報検索において，あいまいさ（*Fuzziness*）を認める直接の価値は，索引者がやっかいと感じている絶対的な yes−no 決定（キーワードを付与すべきか，すべきでないか）を行う代わりに，キーワードが文献に適合している度合を指示することを，索引者に許容することである。

このように，索引時にキーワードおよびその文献との主題関連度を付与することによって，ブール検索システムはファジイ検索システムに拡張される。この結果，(2)の問題は解消し，(1)の論理演算子の限定性はより矛軟になる。Radecki の初期の検索モデルはこの線に沿ったものである[10]。しかしながら，このレベルの拡張では検索質問の入力方式は従来のブール検索式であるため，(3)および(4)の問題は以前として残されたままである。

索引語付与の際の重み付けだけでなく，検索式中の各キーワードの重み付けを可能とする検索モデルが Waller and Kraft,[11] Bookstein,[12] Kantor[13] らによって提案されるようになった。例えば，利用者が以下のような情報要求を持っていると仮定し，従来のブール検索システムと，上記のより拡張されたファジイ検索モデルにおける検索式の相違を比較してみよう。

〈情報要求〉　　SDI（戦略防衛構想）で配備される予定のレーザー兵器，そしてとくにビーム兵器の研究について知りたい。

〈検索式：ブール〉　　sdi AND weapon AND (raser OR beam)

〈検索式：重み付け〉　(sdi, 1) AND (weapon, 1) AND { (raser, 0, 4)OR(beam, 0, 8) }

上の検索式において，1，1，0.4，0.8は利用者が指定した各キーワードの重要度である。この例から明らかなように，上記の研究者が提案したモデルでは最後に残された(3)の問題も表面上解決する。

しかしながら，検索式中の各キーワードの重み付けを可能とする検索モデルには以下に挙げる重大な問題点を持つことが，指摘されるようになった。

(1) ブール検索システムが本来もっていた代数構造上の特徴が失われる。

(2) 検索式中の各キーワードに付与される重みに対して様々な解釈が考えられうる。

(1)は，ブール検索システムの代数構造的特徴の1つである分配律が成立しないことを示している。この問題は，本来情報検索の機能はどうあるべきか，という問題に発展してきている。

(2)は，検索式中の各キーワードの重みは，キーワード間の重要性の相違（*Relevance weight*）を示すものと解釈されていた。しかしBuell[14]は，この重みを文献が満足すべき値（*Threshold Value*）と解釈するモデルを提案した。検索式中の各キーワードの重みは，適合文献を求めるための評価関数の問題と関連し，研究者によってその解釈は様々である。

4．2　ファジイ検索理論の利点と欠点

ファジイ検索理論について，現在までに指摘されている利点および欠点は，以下のようにまとめられる。

〈利点〉

(1) ファジイ検索理論は，索引時および検索時において，キーワード間の重要性の相違を指示することができる。

(2) 上記の利点により，出力時に適合度順に文献を出力することができる。

(3) ブール代数のもつ代数構造がほとんど保持されているため，ブール検索システムに慣れ親しんだ利用者には使いやすい。

〈欠点〉

(1) 文献の適合度の計算はMax, Min演算を行っているため実際にはうまく機能しない。

(2) 検索式中の各キーワードに重みを付与することは，利用者を煩わしくさせる。したがって，自然語の検索質問を重み付け検索式に変換するユーザー・インターフェースが必要とされる。

(3) 検索式中の各キーワードに付与される重みに関して，研究者の間で解釈の一致が見られない。

5．終わりに

確率論，およびファジイ集合論の情報検索への応用によって，情報検索理論は最近10年間に

非常に発展した。2つのアプローチは，最初は対立する理論と考えられていたが，最近になって観点の相違が明らかにされ，両者を統合しようとする考え方もでるようになった。今後，両者の欠点を補完しあう，より一般的な統一理論に拡張されることが期待される。

引用文献

(1) Moores, C. E. "Datacoding and development in information retrieval." ASLIB proceeding. no.8, p.3–22 (1958).

(2) Luhn, H. P. "A statistical approach to mechanized encoding nd seaching of lieterary information." *IBM* Journal of Research and Development. vol.1, no.4, p.309–317 (1957).

(3) Bookstein, A. "Probability and fuzzy-set applications to intormation retrieval." Annual Review of Information Science and Technology. vol.20, p.117–151 (1985).

(4) Maron, M. E. and Kuhns, J. L. "On relevance, probabilistic indexing and inforamation retrieval." Journal of the Association for Computing Machinery. vol.7, p.216–243 (1960).

(5) Radecki, T. "Fuzzy-set theoretical approach to document retrieval." Information Processing and Mangement. vol.15, no.5, p.247–259 (1979).

(6) Luhn, H. P. "The automatic creation of lierature abstracts." IBM Journal of Research and Development. vol.2, no.2, p.159–165 (1958).

(7) Bookstein, A. and Swanson, D. "A decision theoreteic foundation for indexing." journal of the American Society for Information Science. vol.26, no.1, p.45–50 (1975).

(8) Robertson, S. E. and Sparck Jones, K. "Relevance weighting of search terms." journal of American Society for Information Science. vol.27, no.3, p.129–146 (1976).

(9) Angione, P. V. "On the equivalance of boolean and weighted searching based on the convertibility of query formas." journal of American Society for Information Science. vol.26, no.2, p.112–124 (1975).

(10) Zadeh, L. A. "Fuzzy Sets." Information and Control. vol.8, p.338–353 (1965).

(11) Waller, W. G. and Kraft, D. H. "A mathematical model of a weighted boolean retrival system." Information Processing and Management. vol.15, no.5, p.235–245 (1979).

(12) Bookstein, A. "Fuzzy request." journal of the American Sofiety for Information Science. vol.31, no.3, p.240–147 (1980).

(13) Kantor, P. B. "the logic of weighted queries." IEEE Transaction Systems; Man and Cybernetics. vol.SMC–11, no.12, p.816–821 (1981).

(14) Buell, D. A. "A geneal model of query processing in information retrieval systems." Information Processind and Management. vol.17, no.5, p.249–9 (1983).