論文

米国における全文データベースに関する論争

荒木啓介*

[著者抄録] 米国の情報・計算機関係の学会で10年以上にわたって繰り広げられてきた、データベースの全文サーチ方式の是非についての実験研究および論争を、その主要な六つの論文を中心に、時系列的にフォローし、論点、論争の推移、結論を探った。全文サーチを是とする立場の論拠は非常に小規模なデータベースでの実験に基づいているが、人手をかけたインデクシングも大同小異で、コスト的に機械的な方式が有利と説く。非とする立場は、かなりの大規模データに基づき、再現率平均20%で、全文サーチ方式は弾劾の対象であると注意を喚起している。息詰まる論争の経緯も興味あるが、過去の情報システムの評価実験についても、方法的な問題性が指摘され、新たな研究法の提案もなされている。

[著者付与キーワード] 情報検索, システム評価実験, 全文サーチ方式, 論争, 再現率, 適合率, 情報洩れ, STAIRS 研究, Medlars 実験, Aslib-Cranfield 研究, Smart システム, Tenopir 研究, Blair-Marion 研究

Article

Debates upon full-text database search in the United States

ARAKI Keisuke

[Author Abstract] During past more than 10 years, heated debates have been done on full-text search and been published on the journals in the United States. Author pursued the debates, their main points of view and the conclusions based on the 6 main articles. Approvers for full-text search claim the total advantages of mechanized search because of the cost-effectiveness compared with human indexed system showing even similar performance, though their conclusions were derived from their experiments on very small databases as pointed out by the opponents. Opponents to full-text search warn the unexpectedly low recall, average 20%, of full-text and denounce it based on their large scale, practical database experiment. Methodological defects, and accordingly their false results, included in the past famous experiments such as Medlars experiment, Aslib-Cranfield Study and Smart system were also pointed out. New methodology to enable objective comparison of IR system were proposed as one of the fruits of the hot discussions.

[Keywords by Author] information retrieval, system evaluation, full-text search, debates, recall ratio, precision, information loss, Stairs study, Medlars experiment, Aslib-Cranfield study, Smart system, Tenopir study, Blair-Marion study

* 科学技術振興事業団情報システム開発部 (〒102-0081 千代田区四番町5-3) Tel. 03(5214)8457
JST(5-3, Yonbancho, Chiyoda-ku, Tokyo, 102-0081)

1. 緒言

米国の主要な情報関連雑誌に、1980年代中頃に発表された全文データベース（フルテキスト）検索についての二つの相反する実験結果に端を発して、ここ10年以上にわたり賛否両論の様々な論争、反論、再反論が発表されている。

古くは1970年代に、統制語に対するタイトル・抄録からのフリータームの優劣についての研究・論争があり、一応の決着がついたかに見えた。その時点ではサーチできるのはせいぜい抄録までであり、抄録作成の必要性は依然動かぬものであって、統制語との併用が最適な検索結果を与え、抄録者もインデクサーもなお必要である、とする常識的な結論であった [Henzler, 1978]。

しかし80年代から90年代にかけての論争は違った様相も見せている。第1に文献全体がサーチの対象になり、検索目的にだけ限れば抄録を作成する意義は薄いという主張がなされている点、第2は全文サーチでは、多面的で特定性の良い結果が得られるなど、統制語では得難い利点があること、第3に出現語の統計分析や検索専用辞書の整備などで全文サーチの問題点に対する改善策が種々考案されていることなどである。もしそうであるなら、コストや時間を要する抄録作成やインデクシングはもはや不要となる、という議論に繋がってくる。

これに対する全文サーチ批判派の論点は、第1が言語自体の多義性によるノイズの増大は不可避であるという点、第2が全文における主題の表現方式が種々様々で予見困難なことにより再現率は低く、ノイズを減らすために検索条件をきつくと再現率はさらに平均20%程度にまで下がるという点である。

特にこの派の人々が最も強く指摘している点は、データベースの量が少ないうちは良いが、蓄積量が増大するほど検索される文献数が増える。そのままの結果では適合率が低いので数百件の検索文献から10-20%ほどの適合文献を眼で判定する労力をユーザは許容しないという主張である。

今までの全文サーチを擁護する研究は、ほとんどが小規模のシステムか実験規模システムについてのものであり、数万件の文献を対象にした実用システムにおける実地試験の結果からは、全文サーチは糾弾の対象である、とまで言っている。

日本においても最近では全文書を検索できるソフトが普及してきており、インターネット上の検索は若干のコンテンツ案内の上にこの全文サーチ機能に依拠する事例が多いほか、特に特許CD-ROMの検索ソフトのほとんどに装備されている。初心者にも手軽に扱えるという利点が強調されているが、特許の場合などには網羅的な検索が特に不可欠であるから、米国でのこの論争を踏まえた上で、今後の情報システムに対する方針や政策が定められるべきであると思われる。そのために両者の議論をしっかりと後付け、整理しておく必要がある。

2. 論争の要旨

論争の発端および論争している主要な論文は以下のものである。まずこれらを年代順に示し、全文サーチに対してどの立場の議論であるかを示しながら、その要点を記す。

(1) 1985 C. Tenopir. Full Text Database Retrieval Performance. Online Review, Vol.19, No.2. pp.149-164. (賛成派)

方法：

- ・BRSで提供される Harvard Business Review の1,000件以下の部分ファイルを対象にビジネス・労働問題関連の36質問についてテスト。
- ・試行錯誤を行わず、ただ1回のオンラインバッチ的な検索。
- ・全文、抄録、統制語およびタイトル・抄録・統制語の組み合わせの全4種の項目別のサーチにより検索文献、適合文献数を調べた。
- ・全4種の方法による適合文献の総和を分母とし、各方法による適合文献数を分子として、それぞれの方法による再現率（相対再現率）を求めた。

結果：

- ・全文サーチが平均検索文献数で17.8, 適合率18.0%, 再現率73.9%。
統制語ではそれぞれが, 3.1, 34.0%, 28.0%で, 全文サーチが勝っていた。

考察：

- ・全文サーチは副次的なテーマや観点, 言及・列挙されたのみの情報も検索することができて有益である。
- ・ただし, より網羅的なサーチには, どの方法も単独では不十分である。
なおこの文献が全文サーチを是とする根拠として, ほぼ唯一の, 最もよく引用されるものとのことである。

(2) 1985 D. C. Blair, M. E. Maron. An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. Commun of the ACM, Vol.28, No.3. pp.289-299. (批判派)

方法：

法律事務所です訟に関する判例やニュース等40,000件を独自に全文データベース化し, IBM, STAIRS 上で全文サーチのみを実施。裁判で実際に争点となる40テーマについて担当する弁護士が実務としてサーチした結果を集計した。

結果：

1 質問あたりの平均検索文献100-200件, 適合率75.5%, 再現率20.26%。

弁護士が内容的に知っているテーマについても, 具体的に用いられている用語を記憶して主題の表現法を予見することは, 極めて困難であった。

考察：

データベースの大きさが再現率悪化と相関している。

この文献が, 大規模な実験に基づいて全文サーチを批判し, 論争の口火を切ったものである。

(3) 1986 G. Salton. Another Look at Automatic Text-Retrieval Systems. Commun of the ACM, Vol.29, No.7. pp.648-656. (賛成派)

Blair, Maron 研究に対する批判と他の検索実験による反証, 全文サーチのための将来的技術開

発の提案。

論点：

- ・慣習法が基本の英米法では網羅的な判例等の検索が重視されるので, Blair, Maron の研究条件は厳しかった。
- ・Medlars, NASA, Aslib-Cranfield 研究では, 統制語検索もフリーターム検索も Blair, Maron の全文サーチと同じような効率であることを示しており, 一般の研究者, 大学等での利用にはこの程度で十分である。
- ・データベース量が大きくなるにつれての出力過重の問題はない。

(ただし Medlars 実験の紹介のところでは, Lancaster による, 「再現率を85-90%に上げるために, 500-600件の出力文献の通覧をユーザが好むであろうか?」を引用している)。

- ・共出現頻度分析等, 全文サーチを補う手法が考案されつつあり, 人手索引より優れている。未来は自動化システムのものである。
- ・検索用の語彙集, シソーラスは必要である。

(4) 1988 D.R.Swanson. Historical Note:Information Retrieval and the Future of an Illusion. Journal of the American Society for Information Science, Vol.39, No.2, pp.92-98. (賛成派→批判派に転向)

情報検索に関する論争のポイントを歴史的に展望したもの。

論点：

- ・適合文献の判定自体に, 立場により大きな違いが存在する。
- ・伝統的な方法よりも新しい方法に人々は注意を向けがちである。
- ・書籍目録の主題見出しを選ぶ実験での大きなバラツキの結果を引用。
- ・キーワード自動抽出法やテキストサーチの貧弱な実験結果について紹介。
- ・Blair と Maron のフルテキストの貧弱な実験結果を支持し, 過去の自分の小規模実験に基づく楽観的な見解を修正した。
- ・概念の問題である情報検索を機械にこれ以上期

待することは幻想である。

(5) 1990 D. C. Blair, M. E. Maron. Full-text Information Retrieval: Further Analysis and Clarification. Information Processing & Management, Vol.26, No.3. pp.437-447. (批判派)

(6) 1996 D. C. Blair. STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years after. Journal of the American Society for Information Science, Vol.47, No.1, pp.4-22. (批判派)

- ・自己の研究の再確認, Tenopir 研究の批判, G. Salton の批判への反論。
相互に比較可能な再現率を求める方法を提案。
- ・自分たちの研究結果は反論の余地のない事実である。
- ・小規模データベースでは出力過重の問題が生ぜず, 検索条件が緩やかにできるので, 一見再現率が良いように見えるにすぎない。
- ・Tenopir 研究では十分なサーチを行っておらず, しかもその3件, 5件という少数適合文献数のままで相対的な再現率を74%などと求めることは統計処理的にも許されない。
- ・Salton 自身の研究は極めて少数集合 (450件) のもので, 何の結論を引き出せない。
過去の他の研究も, 実施方法, 判定者, 結果の統計処理等で皆問題であり, 結論に疑問が残る。Cranfield 研究などは, 別途手作業で調べたら90%もの適合文献を見落としていた, という報告さえある。
- ・用語についての共出現統計等から得られる文献は, ユーザにとって有用という観点とは無関係で, 全文サーチを改善する単純な自動化法は存在しない。
- ・全文検索方式を強く弾劾する。決定的に重要な文書や高い再現率を要する文献については, 別の補強が必要である。重要な安全マニュアルが的確に検索できなかったために4機の原子炉を停止に追い込まれた電力会社もある。

なお Tenopir 自身は自分の研究が数多く引用され, また Blair らに批判されているにもかかわらず

ず, その後も何ら反論も継続的発表もしていない。

3. 各論文の内容

論争の要点は以上のとおりであるが, その内容についてこの順序に従ってやや詳しく見ていこう。

(1) Tenopir

主目的を同一文献についての, 異なる検索項目による性能比較としたため, 対象データベースとして, 全文, 抄録, 索引された統制語キーワードのいずれをも含む, BRS 上の Harvard Business Review Online (HBRO) を選んだ。これら全項目を含む分は1976年から実験時点までの (遅くとも1985年まで)1,000件未満の規模にならざるを得ず, この規模による繰り返し実験の弊害として, 少量である故に実験者が内容を記憶してしまい, 本当の客観的な実験にならないことが指摘されていることから, Tenopir は質問式作成・検索を各々について1回しか行っていない。この点は逆に, 会話型試行錯誤による検索が本領であるオンライン検索の利点を考慮に入れなかったことになる。

サーチテーマはビジネス関連, 特に労働問題にやや重点のある36テーマで, このうち31テーマについて文献が検索された。36テーマは紹介されているが, 個々に適合率, 再現率をまとめた表としては与えられていない。

研究者自身が事前に質問式を作成し, 自らがサーチを行った。結果の評価は, どの方法による検索かを知らされていない専門家パネルが担当し, 適合性の可否は多数決によって決定した。相対的再現率は, 四つの方法による全適合文献数で, 各々の方法による適合文献数を除して求めた (表1)。

表1からは, 全文検索について適合率が他の項目の半分程度と低い代わりに再現率が最も高いという結果を与えている。一見, 特定の用語でサーチするのであるから適合率は高く, 予期しない用語が漏れているであろうから, 再現率が低いように思われるが, 実際はこのように逆になっている。これは後ほど批判派が指摘する, 自然言語の曖昧

表1 Tenopir 実験の結果 [Tenopir, 1985]

サーチ項目	平均検索文献数	検索数幅	平均適合文献数	平均適合率	平均相対再現率
全文	17.8	0-65	3.5	18.0%	73.9%
抄録	2.4	0-11	1.0	35.6%	19.3%
統制語	3.1	0-24	1.2	34.0%	28.0%
書誌連合	5.3	0-29	2.0	37.0%	44.9%
合計(異なり)	21.2	1-69	4.5	21.2%	100.0%

表2 Blair, Maron 実験の再現率と適合率 (*印は準備用質問, Blair, Maron, 1985)

情報要求番号	再現率	適合率	情報要求番号	再現率	適合率
1	*	*	27	50.0 %	42.6 %
2	45.5 %	92.6 %	28	50.0	19.6
3	*	*	29	*	*
4	*	*	30	7.0	100.0
5	*	*	31	*	*
6	8.9	60.0	32	12.5	100.0
7	20.6	64.7	33	18.2	79.5
8	43.9	88.8	34	14.1	45.1
9	13.3	48.9	35	*	*
10	10.4	96.8	36	*	*
11	12.8	100.0	37	15.9	81.8
12	9.6	84.2	38	24.7	68.3
13	15.1	85.0	39	18.5	83.3
14	78.7	99.0	40	4.1	100.0
15	*	*	41	18.3	96.9
16	*	*	42	45.4	91.0
17	*	*	43	8.9	100.0
18	13.0	38.0	44	10.6	100.0
19	15.8	42.1	45	20.3	94.0
20	19.4	68.9	46	11.0	85.7
21	41.0	33.8	47	13.4	100.0
22	22.2	94.8	48	13.7	87.5
23	2.8	100.0	49	17.4	87.8
24	*	*	50	13.5	75.7
25	13.0	94.0	51	4.7	100.0
26	7.2	95.0			

さによるものである。また、この相対再現率自体が真の再現率とどう関係するのかが不明であるほか、この程度の少数例で一定の結論が引き出せるか否かの議論はなく、本文中で言及された言葉でもサーチできた点も含めて、全文サーチ方式の優位を結論づけている。ただし「どの方法も単独では完全ではなく、それぞれに独自の寄与がある」というのが彼女の本当の結論であって、この点を後の引用者は見落としているようである。

なお Tenopir は報文の後半で、各項目でのみ検索される文献をユニーク度として表現し、それぞれの項目の特徴を数値で表す試みを提案している。

(2) Blair, Maron 実験

実際に訴訟を代行している法律事務所で、担当弁護士たちが、判例、ニュース記事など40,000件、350,000ページの文書を自分たちで選定し、IBMのSTAIRSシステムに全文入力して検索実用に供し、併せて評価実験を行ったもの。ユーザである弁護士が51の質問（うち11質問は事前のチェック用で、40質問について結果の集計を行った）を行い、準弁護士が十分打ち合わせの上質問式を作成し、その結果を弁護士に送付し、満足の行くまで何度も修正した。訴訟においては、必要な文書の75%が検索されることが必須であるが、ほぼその値に達したと判断された時点でサーチを打ち切った。適合性の判定は「有用である」ことを基準に弁護士自身が行った。再現率については、いくつかの集合をAND演算して検索する場合に、演算項を一つ減らした集合から、先に検索した文献集合を差し引いた残りを再度、判定者に回付して（そのような措置をしたことは伏せて）判定を仰ぐ方法で、これらの中に含まれる適合文献の総和に最初の正規の検索結果の適合文献数を加えたものを分母として計算した。つまり判定者は知らずにAND条件を減らした検索結果を通覧させられたことになるが、40,000文献全部を通覧することが不可能である以上、次善のやむを得ない策であろう。Blairらは次報以下で、むしろこの方法を再現率計算の標準的な方法とするよう提案してい

る。結果の判定は個人差を避けるために複数で行い、またサーチも弁護士自身が行った場合とも比較し、数値の統計処理も行っている。

得られた結果は表2のとおりである。

適合率は最高100%、最低19.6%で非加重平均79.0%（標準偏差23.3）、加重平均75.5%であり、再現率は最高78.7%、最低2.8%、非加重平均20.0%（標準偏差15.9）、加重平均20.26%であった。弁護士別の値は、表3のとおりである。

表3 弁護士別の評価の違い

	適合率	再現率
弁護士A	76.0%	22.7%
弁護士B	81.4%	18.0%

この偏差は0.5のレベルで統計的に有意ではない。従って判定の個人差はない。

検索実務に慣れている準弁護士と、質問者自身である弁護士が直接サーチを行った場合も比較している（表4）。

表4 代行者（準弁護士）のサーチと質問者自身（弁護士）のサーチの違い

質問番号	再現率（準弁護士）	再現率（弁護士）
1	7.2 %	6.6 %
2	19.4 %	10.3 %
3	4.2 %	26.4 %
4	4.1 %	7.4 %
5	18.9 %	25.3 %
平均	10.7 % (s.d. = 7.65)	15.2 % (s.d. = 9.83)

すなわち結果には多少の違いはあるものの統計的に有意な差ではない。従ってサーチ業務に慣れているか否かとか質問者自身のサーチか依頼かでの違いはなく、システムそのものの特性として、この低い再現率しか得られなかった、と結論づけている。

その主たる原因はデータベースの内容が予見できないことで、データベースのサイズが大きくなるほど過度の出力数を減らさざるを得ず、この状況は悪化すると考察している。

検索から漏れた文書の例として、「交通事故」を扱ったもので、書き手の立場の違いから、必ずしも accident が用いられず、event, incident, situation, problem, difficulty などと表されていたり、what happened last week was…のような表現のものもあった。また、‘steel quantity’ (鋼量) が起訴できる事件に関連した重要なキーフレーズであったが、‘steel quantity’を全く報告せず、単に ‘girders’ (梁), ‘beams’ (桁), ‘frames’ (枠), ‘bracings’ (支柱) etc. としか言及していない未検索・適合文献の例を挙げている。

このような Blair らの研究に影響されて、WESTLAW システムのように人手による索引の追加を始めたところもある。

(3) Salton の反論

Blairs, Maron 研究の適合率平均 75% は良い値であり、貧弱な再現率の主張のほかは STAIRS の好印象を与えるものであり、人手索引との比較や大規模ファイルと小規模ファイルの両方での比較を行っていないのに Blair らの結論は感情的で

ある。判例を重んじる英米法の下では高い再現率による検索を必要とするが、一般の研究者や大学、学生向けにはこの程度で全く十分である。ファイル作成や索引にコストがかからないだけ優れており、高い適合率のため、不適合文献を除去する手間があまりかからず、全体として経済的である。従来の人手索引／自動化システムの比較実験でも自動化システムが幾分優れていることが示され、一方人手をかけたシステムである Medlars や NASA の International Aerospace Abstracts (IAA) による実験の結果からは、人手によるシステムにも問題が多いことを示している (表 5～7)。

さらに Salton は人手による索引方式の問題点として、過去の次のような報告を紹介している。

- ・二つの集団が同一の主題領域についてシソーラスを構築すると、索引語のわずか 60% しか両方で共通とならない。

- ・二人の経験あるインデクサーが同一の文献を同一のシソーラスを用いてインデクシングしても、わずか 30% の索引語しか両者に共通にならない。

- ・二人の仲介サーチャーが同一ホスト上の同一データベースで同一の質問についてサーチすると、出力のわずか 40% しか両者に共通にならない。

表 5 Medlars の効率 [Lancaster, 1968]

種別	再現率	適合率	検索文献数	適合文献数
高適合率サーチ	19%	80%	40-50	30-40
平均的サーチ	58%	50%	175	85
高再現率サーチ	89%	20%	500-600	135

表 6 NASA サーチシステムの比較評価 [Cleverdon, 1977] (44,000 文献, 40 質問)

索引付け方法	再現率	適合率
自然語 (タイトル・抄録のサーチ)	78%	63%
関連語で補足した自然語	73%	52%
人手統制語索引	56%	74%
自然語で補足した統制語	71%	45%

表7 Medlars サーチの典型的な失敗原因 [Lancaster, 1968]

失敗の原因	797 件の再現性の失敗	3,038 件の適合性の失敗
索引語 (適切な用語の欠如,誤組合せ)	10.2 %	36.0 %
検索式 (特定性過剰, 広すぎ)	35.0%	32.4%
文献索引付け (特定性過剰, 広すぎ)	37.4%	12.9%
ユーザ - システム (インタフェース不備)	25.0%	16.6%

い。

・二人の科学者ないし技術者が与えられた同じ質問に対する同一の文献集合について適合性を判定すると、判定が共通する部分は60%を越えない。

次に Salton は、出現用語の統計的な分析に基づく類似文献の自動検索方式について説明し、この種の方式が従来の人手による索引方式と同等かしばしばより優れている、と述べ、最後に Swanson の25年前の自動索引の実験から引用して、「…伝統的な検索法に対する機械的テキストサーチの相対的優位は、実験が重ねられテキストサーチの補助手段が改善されるにつれて一層大きくなるであろう。一方従来のインデクシングシステムを改善する確実な手法は証明されていないのである」として論文を締めくくっている。

(4) Swanson のレビュー

Taube のユニタム方式と ASTIA (Armed Services Technical Information Agency) の伝統的な主題見出し語方式の1951年に行われた比較実験において、98質問についての2,967の検索文献のうち1,577文献 (53%) もの文献が適合性について両方で不一致であり、方法論的に大きな問題があったにもかかわらず、両者とも勝利を主張して譲らず、世論はより新しい技術を支持する傾向を示したと、これが今日に至る論争の原型をなしているが如く指摘している。またコロンビア大学図

書館における340名の学生に6冊の図書について主題見出しを付けさせた実験では、1冊平均62もの見出しが付けられ、うち61は実際の見出しとは異なるものであったという事例を挙げて、いかに主題分析・表現の統一的な実施が困難であるかを論じている。一方 Luhn の KWIC 索引以来多くの自動索引が提案されているが、これは提案者があたかもクレイ射撃のように放り上げたアイデアを粉々に破壊し、他の者がその断片を拾い集めて有り難がっているが如きもの、と断じ、意味の問題を機械が扱う困難性を指摘している。小規模ではあるが自らの物理学の文献集合についての自動索引実験の結果が貧弱であったことを告白し、「科学者が自然語の自由さにまかせて相当数の表現法を駆使するので、表現の単純な一致に基づくシステムは打ち負かされてしまう」という Bar Hillel の言明を引用し、Blair, Maron の研究結果を支持している。

科学の歴史において、「無限運動の不可能性から熱力学が生まれ、無料の食料は存在しないの認識から経済学が発展し、全ての人に永久にウソをつけないことから政治理論が発達した」ように、不可能であることを明確にすることがその発展の基礎になるが、情報科学においても「9つの不可能な命題」を確認した上で、研究、開発に取り組むべきことを提案している。「9つの不可能な命題」(9PI, 9 Postulates of impotence) とは以下のも

のを言う。

PI 1: 情報ニーズを文脈的前提とは独立なサーチ要求としては完全には表現しきれない。検索は未知なものに対して行うものであるから、回答が見つかる前にそれに対する質問を完全には作成できない。

PI 2: 叙述した要求を適切なサーチタームのセットに翻訳することを機械に指令することはできない。

PI 3: ある文献は要求者が考慮するであろう他の文献とは独立に、情報要求に対し適切であるか否かを判定できない。適切性は相対的であり、より適切と思われる未知の文献が見つければ以前のものは否定されることもある。

PI 4: ある質問についての適合文献を全て見つけたか否かを確証することは不可能である。

PI 5: 機械は意味を認識できない。従って人間の判断が文献のインデクシングや分類作業で成し得ることを機械がまねることはできない。

PI 6: 単語の出現頻度統計は、意味を表現したりその代わりとなることはできない。しかしこの種のデータは人間が意味や適合性を見つける可能性のある箇所を示し、時には成功裏に利用することはできる。

PI 7: 反復のプロセスを支援する IR システムの能力は、人間が単純に行う反復的な適合性判定によっては評価できない。反復を繰り返すと、質問を創造的に修正するような別の基準が現れてくる。

PI 8: 緻密な適合性判定が極めて有効な機械の手順のいずれかを持つことはできるが、両者を同時に持つことはできない。

PI 9: 上記 8 つの命題は全体として、常に有効な全自動索引・検索システムは不可能であることを示唆する。

以上のいくつかは、丁度反対の表現ではあるが、Fugmann が定めた情報検索に関する 5 命題理論 [Fugmann, 1992] にも符合する。

さらに Swanson は科学者における情報検索、情報分析の重要性の認識が依然として浅いとし、

実はこの点に将来の情報システム発展の鍵があることを指摘している。

(5, 6) Blair, Maron の再反論

主として Salton の批判に答える形で反論(5)を発表し、また最初の(2)の論文発表後 10 年を経てなお引用され議論の的になっている点を踏まえて全文サーチの問題を改めて総括したのが(6)である。

Blair らは Salton に反論する理由として、以下の三つを挙げている。1: フルテキストサーチは情報検索の分野で急速に成長している活発で重要な技術であるが、一般には、むしろ工学的な問題である文書の「物理的なアクセス」に開発の力点が置かれ、より基本的な「知的なアクセス」が軽視されていること。2: 自分たちの STAIRS 研究は、その規模、研究計画、方法論および結果の統計処理の点で情報検索の実験的研究における画期的なもので、今後はそうしばしば行われまいであろうから、この意味するところを明確にしておきたいこと。3: 情報検索分野の先駆者として著名で影響力の大きい Salton の反論にきちんと答えることが必要と認めたことである。この上で、周到な方法による自らの実験結果の平均再現率 20% の数字は反論の余地のない事実である、と強調し、内容に精通している専門家が何回ものサーチを繰り返し、必要文書の 75% 以上は検索されたと信じてサーチを終了したにもかかわらずこの程度の値であったのは、適合文献中の単語やフレーズを事前に予見して的確な質問式を作成できなかったことが第 1 の原因である、と述べている。具体例として最初の報文にはない次の例を挙げている。

train accident (列車事故) についての文書が、書く人の立場などで列車の事故とは同義語ではない以下のような多様性で表現されている:

unexpected event (不測の事件)

unfortunate occurrence (不幸な出来事)

untimely incident (不時の事態)

tragic disaster (悲劇的な災害)

monumental blunder (記録的な大失策)

transportation system malfunction (運輸シス

テムの誤動作)

つまり、被害を被った人、事故を引き起こした当人、原因究明の責任者、報道人などで同じ列車事故でもこのように様々で、単に train と accident ではこれらが検索できない上に、trains personnel to avoid accident (事故を防ぐために要員を訓練する) や accidental uncoupling of trains (列車の偶発的な解き放ち (事故には至らない)) などがノイズとして大量に混入する。逆にこれらを検索しようとしてサーチタームに加えると、今度は別のノイズを大量にもたらししてしまう。ここでは言語に関する基本的な問題が指摘されている。一般に用いられる言語とは、そこで発話された言葉、そこにある書かれた文字のことであって、曖昧な言葉も文脈によって通常は明確に理解されるのである。しかし情報検索の言語は「いまだ知らない人が書いた、いまだ読んだこともない書き物の中の単語を問題にする」のであるから、同じ言語といっても似て非なるもので、根本的に異なる前提にある。フルテキストサーチはこの両者を混同し、「何もしないで何らかの成果を得ようとする試み」と見なされるから、種々の問題を抱えているのである。

単純なフルテキストサーチも数百件程度の規模なら我慢できるが、データベース規模が大きくなるにつれて上記のような言語に起因する問題を悪化させ、サーチャーは出力文献過剰の泥沼に陥ってしまう。これを妥当な大きさにまで減らす試みとして AND 条件を追加せざるを得ず、その結果今度は再現率が「劇的に低下する」ことになる。

以上のように一般的に反論し、次いで Salton の批判に誤解があることを指摘する。つまり STAIRS 研究は抄録でなく、元より抄録のない全文を対象とし、しかも Salton が言うような洗練された種々の補助機能は用いず単純な全文サーチを評価したもの、対象の文書は法律文書だけではなく、大会社の訴訟問題に関係する技術的・工学的文書、通信文、議事録などを含むものである。Salton が「出力過重の問題は無い」と言っている点については、Salton 自身が「不適合文献を多数

含んでいる可能性のある資料を利用者は喜ばない」という Lancaster のコメントを引用して実質的に「出力過重の問題」を認めている、と指摘しているほか、Zipf の法則を援用して、文献数が増えれば確実に各サーチタームの出現度数、ひいては出力文献数が増えることは経験的に証明されている事実であり、否定できない、として、理論的にも論駁している。結局、Salton が言う情報検索とは、必要な文献が適当に出てくる、大学等でそれなりに利用できれば十分なもの、ということになる。

ここで Blair らは問題を整理し、目標や価値観の相違が論争に色濃く反映されている点は切り離して、純粹に科学的に情報検索システムの評価方法を確率することを提案している。というのも、過去の評価実験が方法論的に不備であったからで、まず以下の4点を条件とすべきことを提案している。

1. 大規模なデータベースを対象とすべきこと。
2. 実験環境ではなく、実働環境で評価すべきこと。
3. 適合性判定は、質問者自身が有用性に基づき行うべきこと。
4. 得られた結果は統計処理すべきこと。

これに照らして情報検索システムについての過去の著名な評価実験を見ると、表8のように方法論的に不完全である。

次に Blair らは、大規模データベースで実験する場合の再現率算定の困難を軽減する方法として図1のような便法を実際に適用し、それをデータベースの違いを越えて結果を比較できる客観的な方法として提案している。つまり、A, B, C, D の四つの概念集合の積で表した元の質問式による集合1に存在する適合文献数を分子とする。条件を少しずつ緩め、しかも集合1の部分差し引いて得る、集合2から5、あるいは6から9、さらに場合により10から13までを求め、その中の全適合文献数と先の集合1の適合文献数の和を分母として再現率を求めるのである。最終ユーザである判定者に、1度見た文献集合を再度回覧することが無

表 8 評価研究法の特徴

実験 実験法	DB 規模	テスト環境 実働?	適合性評価者 質問者?	データの統計処理
1. MEDLARS (Lancaster,1968)	70 万件	Yes	No	No
2. Cranfield II (Cleverdon,1966)	1,400 件	No	No	No
3. NASA data base (Cleverdon,1977)	44,000 件	No	No	No
4. SMART-MEDLARS 比較 (Salton,1973)	450 件	No	No	Yes *
5. Tenopir 研究 (Tenopir,1985)	1,000 件以下	Yes	No	No
6. STAIRS 研究 (Blair,Maron,1985)	40,000 件	Yes	Yes	Yes

*注 SMART と MEDLARS の検索データを比較するために統計的手法を用いた。

しかし SMART システム単独のデータが有意か否かを見るための検定は行われていない。

1. $A \cap B \cap C \cap D$ (元の質問式)

2. $A \cap B \cap C \cap \neg D$

3. $A \cap B \cap \neg C \cap D$

4. $\neg A \cap B \cap C \cap D$

5. $A \cap B \cap \neg C \cap \neg D$

6. $A \cap B \cap \neg C \cap \neg D$

7. $A \cap \neg B \cap C \cap \neg D$

8. $\neg A \cap \neg B \cap C \cap D$

9. $\neg A \cap B \cap \neg C \cap D$

10. $A \cap \neg B \cap \neg C \cap \neg D$

11. $\neg A \cap \neg B \cap C \cap \neg D$

12. $\neg A \cap \neg B \cap \neg C \cap D$

13. $\neg A \cap B \cap \neg C \cap \neg D$

$\cap$ は AND, $\neg$ は NOT
条件を意味する。

図 1

いので、協力を求めるためにも有効であるという。

最後に Blair らは、文書中の出現単語の頻度統計などに基づくいくつかのランキングアルゴリズムを紹介し、いずれも「検索文書の適合性のランクを予測することはできず、全文検索システムを改善する問題での単純な自動化法は存在しない」とし、「単純な全文検索システムを強く弾劾する」

とまで言い切っている。「全文検索は無いよりはまし」という議論に対しては、「決定的に重要な任務や、高い再現率を必要とする用途においては、単純な全文検索のみでは目的を達成できないということであって、注意深く考えられた論理あるいは知性によって補強されなくてはならない」と結論で述べている。

3. ま と め

この一連の論争はいまだ終わっていないようであるが、両者の論点が明らかになっただけでなく、今まで曖昧であった事柄が明確になり、将来の一層の研究のための方向が示された点に、大きな意義があったものと思われる。これらを整理しよう。

第1に、事実の確認という意味で、全文検索は再現率が予想以上に悪いということが明らかにされた点である。

第2に、人手による索引にも大いに問題があることも、再度浮き彫りにされた。

第3は、今までの情報検索実験の方法自体に、従ってその結論にも問題があり、最終ユーザに適合性判定を委ねる大規模な実用環境での実験が必要条件であることが、指摘されたことである。

第4は、Saltonによる反論が無ければあまり明確には示されなかったかもしれない、Blairらの研究方法、特に簡便な近似的再現率の求め方が発表されたこと。ただしこの方法も、数百万規模のデータベースの場合にはもはや対応できそうもないものと思われる。

第5に、立場・見解の相違が明らかにされた点である。一方は情報検索に高い精度を要求し、他方は、トータルコストを考慮すれば全文サーチないし統計的方法で程々に十分役立つ、とする立場である。米国においては、この二つに情報関係者が2分されて、計算機関係者に後者の立場をとるものが多いという。ただここで心配なことは、全文検索の基本は1950年代に Taube が提唱しその後失敗であったことが判明したユニターム方式の計算機版であることや、情報検索における言語とは、いまだ未知なる対象概念を事前に予測して

サーチする手段なのであって、そこに書かれてある、あるいはその場で発言された表現を理解する一般の言語とは似て非なるものである、という認識がドキュメンテーション側から計算機専門家側へ正しく伝えられているか、が大きな問題である。

第6として、Swansonのようなこの分野における大家の一人が、この論争に大きく影響されて、以前の自分の結論を実質的に撤回し、情報検索における「9つの不可能の命題」を提示するに至ったことも、極めて興味ある出来事というべきであろう。ここで悟ってしまうと、その先の進歩は望めない。しかしこの部分の冒頭で述べている「無限運動の不可能性から熱力学が生まれ、無料の食料は存在しないの認識から経済学が発展し、全ての人に永久にウソをつけないことから政治理論が発達した」という指摘は、全て情報検索の状況について言っているともとれる。こういった意味で、情報検索の科学としての発展を期待すると共に、効率の良い情報システムの構築と運用に人手が不可欠であっても、人手の絶対数が不足する一方であるから、専門家の作業やユーザを支援する技術としてのエキスパートシステムの構築が今後の課題であろう。

論争を直接読むと、相手の非難する言葉を逆に投げ返し（互いにその結論を感情的である、などと）、生々しいやりとりで応酬しているが、Blairは、文頭に「学ばんとする願い切なところ、多くの議論、多くの記述、多くの意見あるべし：意見は良き人々にありては、造られつつある知識に他ならぬが故なり」という、イギリスの詩人、Miltonを引用して、論争が科学の発展に不可欠であることを強調している[Blair, 1990]。この点が我々の最も学ぶべきことかもしれない。

参 考 文 献

- 1) Blair, D. C.; Maron, M. E. An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. Commun. of the ACM. 28(3)289-299 (1985)
- 2) Blair, D. C; Maron, M. E. Full-text information retrieval: Further analysis and clarification. Inf. Process. & Manag. 26(3) 437-447 (1990)

- 3) Blair, D. C. STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years after. J. Am. Soc. Inf. Sci. 47(1)4-22 (1996)
- 4) Cleverdon, C. W. et al. Aslib-Cranfield Research Project. vol.2. Test Results. (1966)
- 5) Cleverdon, C. W. A Comparative Evaluation of Searching by Controlled Language in an Experimental NASA Data Base. Report ESA 1/432, European Space Agency, Frascati, Italy, 1977
- 6) Fugmann, R. Theoretische Grundlagen der Indexierungspraxis. 63p, Index Verlag, 1992. (日本語訳, 「情報システム・データベース構築の基礎理論」, 情報科学技術協会, 1994.)
- 7) Henzler, R. G. Free or Controlled Vocabularies. Intern. Classific. 5(1)21-32 (1978)
- 8) Lancaster, F. W. Evaluation of the Medlars Demand Search Service. National Library of Medicine. 1968.
- 9) Salton, G. Recent Studies in Automatic Text Analysis and Document Retrieval. J. of the ACM. 20(2)258-278 (1973)
- 10) Salton, G. Another Look at Automatic Text-Retrieval Systems. Commun. of the ACM. 29(7)648-656 (1986)
- 11) Swanson, D. R. Searching natural language text by computer. Science. 132 (3434) 1099-1104 (1960)
- 12) Swanson, D. R.. Information retrieval as a trial and error process. Library Quarterly. 47 (2)128-148 (1977)
- 13) Swanson, D. R. Historical Note: Information Retrieval and the Future of an Illusion. J. Am. Soc. Inf. Sci. 39(2)92-98 (1988)
- 14) Tenopir, C. Full Text Database Retrieval Performance. Online Review. 19(2) 149-164 (1985)