1 情報検索技術とテストコレクション

岸田 和明　駿河台大学　kishida@surugadai.ac.jp

■テストコレクションとは■

情報検索（IR）の分野では，その黎明期から，さまざまなテストコレクションを用いた検索実験が積み重ねられ，文献（文書）の検索技術の発展に寄与してきた．初期のものとしては，1950年代から1960年代にかけて英国のクランフィールド（Cranfield）で行われた検索実験が著名である．この実験では，1,400件の文献からなるテストコレクション（CRANと呼ばれる）が作成され，ディスクリプタ（後述）を用いた検索の性能等が比較評価された．そして，この実験が1つの手本となり，検索実験の手順が確立され，以降，さまざまなテストコレクションが作成・利用されるようになった．よく知られたものとしては，コンピュータ科学の抄録を対象としたCACMや医学分野のMedlineデータベースから抽出されたいくつかのテストコレクションなどがある[1)]．

テストコレクションとは，より正確には，

(1) 文献データの集合（標題・抄録または全文）
(2) 検索質問の集合
(3) 各検索質問に対する各文献の適合／不適合の情報

の「3点セット」を指す．テストコレクションを利用することによって，複数の検索手法の比較評価が可能になる．すなわち，各質問に対する検索をそれぞれの手法で実行し，どれだけ効果的あるいは効率的に適合文献を見つけ出したかを比較すれば，各手法の性能に優劣をつけることができるわけである．

本稿では，このようなテストコレクションがどのようにIR技術の発展を支えてきたかを概観し，テストコレクションの必要性とその限界とをみることにしたい．

■概念からの検索：統制語彙 vs 自然言語■

現在では，画像や絵画，映像，音声にまで検索の問題が幅広く及んでいるが，以前は，情報検索といえば主に雑誌論文や図書に対するいわゆるテキスト検索を意味した．特に初期の頃は，雑誌論文をその標題や抄録を収録したデータベースから検索することが中心課題であり，その技術的問題に対する解決法をテストコレクションを使って比較評価しようとする試みがいくつかなされた．

その典型は，自然言語による検索と統制語彙による検索との性能比較である．前者は標題や抄録中に出現する語句をそのまま使った検索であり，後者は，人間の索引作成者がその文献の主題に相応しいと考えて付与した語句からの検索を指す．この場合，付与される語句はシソーラスと呼ばれる辞書に登録されたものに限られ，その語句は特にディスクリプタと呼ばれる（図-1参照）．

標題や抄録中に出現する語句のみを検索に用いる場合，用語の使用の多様性がその性能に悪影響を及ぼすことは容易に想像できる．たとえば，文献中で「計算機」，検索質問中で「コンピュータ」が使われているならば，文字列の単純な照合ではこれらは一致しない．一方，シソーラス中で“「計算機」ではなく「コンピュータ」を用いる”と規定されていれば，統制語彙による検索ではこの問題は生じない．

この問題は「概念」による検索の可能性にまで及ぶ．つまり，シソーラス中のディスクリプタが単なる語句ではなく「概念」を包括的に指示するものであり，そして，人間の索引作成者が各文献の概念を把握して，正確にディスクリプタを付与できると仮定すれば，文字列のレベルを超えた，概念からの検索が可能になる．

直感的には，概念からの検索の方が有用であるように思える．そのための方法としての統制語彙の開発と実証こそがIR研究の基本問題の1つであり，1960年代後半にテス

トコレクションを使った検索実験が試みられた．ところが，そこでの検索実験は，自然言語に対する統制語彙の優位性を示すことができなかった．逆にいえば，1960年代の自然言語からの検索技術でさえ，多大な人間の索引作成者の労力を要する統制語彙による検索と同程度の性能を実現できたわけである．これは，それまで実際にシソーラス付きのデータベースを開発・提供してきた人々にとっては意外な結果であったと想像できる．

現在の図書館などにおける情報検索サービスの現場では，その長短に応じて自然言語と統制語彙とを使い分けることが常識となっている[3]．つまり，1960年代のテストコレクションによる実験の結果は，その後の実際の場での経験から，ある意味で「追証」されることになったわけである．

■ベクトル空間型モデルとtf-idf■

IR研究のもう1つの方向は，自然言語からの検索の性能を向上させて，人間の索引作成者による統制語彙を用いた検索を超えたものにするということになる．すなわち，人間の索引作成者を介さない，完全に自動的な索引作成・検索システムの開発である．その典型例として，1960年代からG.サルトン（Salton）を中心に進められたSMARTシステムの研究が挙げられる[3]．

SMARTはさまざまな技術を組み合わせた総合的なシステムであるが，その中心は自動索引作成（automatic indexing）と照合（matching）のメカニズムである．この場合の自動索引作成とは，文献の標題や抄録あるいは全文中に含まれている語句から索引語（検索の手がかりとなる語）として有用なものを機械的に識別することを意味する．

基本的には，文献中で繰り返し言及されている語句はその文献にとって重要であると考えられる．したがって，その文献中の語の出現回数（term frequency: tf）が索引語選択の目安になる．しかし，いくら出現回数が多くても，たとえば「研究」や「方法」など，どの文献でも用いられるような語は索引語としての有用性は低い．この意味では「コンピュータ」のような専門用語も，コンピュータに関するデータベース中では索引語としての有用性は低い．そこで，数多くの文献に出現する語ほど索引語としての有用性が低いと仮定して，出現文献数の逆数（inverse document frequency: idf）を使ってこのような語を自動識別することが考えられる．具体的には，tfにidfを掛け合わせればよい．以上の手法はしばしば「tf-idf」と呼ばれる．

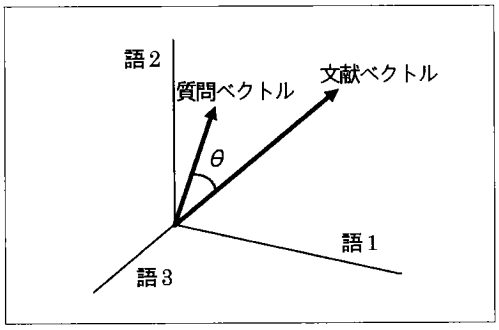
tf-idfを使って各語の重みを計算し，その羅列を実数のベクトルとみなせば，文献をベクトル空間中で表現できる．同様に検索質問の文を解析して，検索質問ベクトルとして表現すれば，次の問題はこれらのベクトル間の類似度をいかに計算するかということになる．これが「照合」の段階であり，SMARTではベクトル間の角度のコサインが使われる（図-2参照）．そして，類似度の降順に文献が出力される（いわゆる適合度順出力）．以上のような検索手法は一般にベクトル空間型モデル（vector space model）と呼ばれている．なお，現在のインターネットのサーチエンジンのいくつかはこのSMARTシステムに類似した方法を用いているといわれている．

SMARTシステムの開発・改良は長年に渡って進められているが，その性能の検証にはたびたびテストコレクション（すでに述べたCACMなど）が用いられてきた．それはあるときには，人手による統制語彙との比較であり，またあるときには，語の重みの計算方法間や照合の手法間の比較であった．この一連のSMARTの研究によって，IR技術の研究の1つのスタイルが確立されたといっても過言で

```
検索モデル USE 検索理論(HG.4)

検索もれ （HG.4.3.4） < 2> 94/04
    BT 情報検索
        ・検索理論
        ・・検索効率

検索料金 （HG.6.5） < 13> 94/04
    BT 情報検索
        ・データベースサービス機関

検索理論 （HG.4） < 20> 94/04
    UF 検索モデル
    BT 情報検索
    NT レレバンス
       アバウトネス
       検索効率
       ・再現率
       ・精度
       ・新奇性率
       ・検索もれ
       ・検索ノイズ

（以下略）
```

ゴチック体の見出し語がディスクリプタで，明朝体の見出し語が非ディスクリプタ．なお，BTは上位語，NTは下位語を意味する．

図-1 情報検索のためのシソーラスの例[2]



語の総数が3語のみと仮定．

図-2 ベクトル空間型モデル

はない．つまり，何らかの理論から始め，そこから具体的な計算方法を導き，最後にテストコレクションによる検証で閉じるという流れが，いつしかIR技術研究の標準とみなされるようになった．この流れは現在でも続いており，この点でも，テストコレクションおよびそれによる評価は，IR技術の発展にとって必要不可欠な存在となっている．

■適合度順出力の方法■

1970年代後半になると，確率型検索モデルと呼ばれる新たな検索手法の研究が盛んになった．これは，基本的には，各文献が検索質問に適合する確率を推計して，その降順に文献を出力するものである．この手法によって，ベクトル空間型モデルとは異なる枠組みでの適合度順出力が可能となるので，当然，ベクトル空間型と確率型とのどちらが優れているのかという疑問が出てくる．

この基本的な問題にある程度の解答を出したのが，米国における検索実験プロジェクトTREC（Text REtrieval Conference）である（その第1回目は1992年）[4]．このプロジェクトの最大の特徴は，それまでのものとは比較にならないほど大規模なテストコレクションを使用している点にある．それ以前のテストコレクションはいずれも規模が小さく，それがテストコレクションによる評価の妥当性を疑う理由の1つであった．たとえば，クランフィールド実験におけるCRANの文献数は1,400件，CACMでは3,200件である．規模が小さくなってしまう主な原因は，検索性能を正確に評価するには検索質問に適合した文献をデータベース中からすべて洗い出す必要があり，データベースが大きいとその作業が難しいためである．

TRECでは，数多くの研究チームを参加させて，それらが検索した文献を寄せ集めてプールして，それに対してのみ適合／不適合を調べることにより，この問題を解決した．この方法を使えば，データベース中の全文献を吟味する必要がなくなり，そのためテストコレクションの規模は飛躍的に大きくなった．その規模は，たとえばTREC-1では約70万件である．各研究チームがそれなりに適合文献を検索できていれば，この方法による評価の近似の信頼性は高くなる．

TRECによって，各手法の適合度順出力の性能に関する多くの知見が得られた．また，TRECでは各研究チーム間の競争形式をとっており，それによる効果もあって，各手法の検索性能自体が向上した点も見逃すことができない．

たとえば，TREC以前の確率型モデルはかなり理論的なものであり，実用的ではなかった（大規模な統計調査をするわけにもいかないIRの状況では，モデルに含まれる確率分布のパラメータを正確に推定することは難しい）．ところが，実際のテストコレクションに対してある程度の検索性能を示すには，モデルを調整して，そのあたりの問題を解決する必要が出てきた．その結果がtf-idfの導入である．

最も古典的な確率型検索モデルにはtfは含まれない．ただし，適合／不適合の情報が存在しないと仮定すれば，idfに似た項がモデル中に出てくることは知られていた．一方，その後提案されたポアソン分布に基づく別の確率型モデルには，tfは含まれるが，idfは含まれない．TRECでは，これらの2つのモデルが融合され，この結果，ベクトル空間型モデルと同様に，tf-idfが確率型モデルに組み込まれることになった．この手法は，ロンドンのCity Universityのグループによって提案され[5]，彼らが開発したシステムOkapiにちなんで"Okapi weighting"などと称されることがある．

この手法はベクトル空間型モデルと遜色ない，時にはそれ以上の検索性能を示すことができた．ただし共にtf-idfを基本とするために，両手法の間にそれほど大きな差は生じない．結局，現在のところ，tfとidfとdl（document length: 文献長）の3つの要因を組み込んだ手法はそれなりの検索性能を示すということが，適合度順出力に関するTRECでの1つの知見ということになっている[6]（dlはベクトル空間型モデルではコサインの計算式中のノルムのかたちで組み込まれ，確率型モデルでは，語の出現確率の計算の分母として出てくる）．

■自然言語処理技法の応用■

自然言語処理（NLP）技法のIRへの応用もまた古くから探究されてきた問題であり，その典型としては複合名詞の自動識別が挙げられる．たとえば，「情報の蓄積と検索」という表現を含む文献を「情報検索」という語から検索するには，「情報の蓄積と検索」という文字列から「情報検索」を正しく識別する必要がある．これを自動的に行うには，NLPの技法によって図-3に示すような語間の依存関係を示す木を生成すればよい．

しかし一般的には，検索性能へのNLPの効果に関しては，否定的な見方が強く，「computer」と「computers」のような語尾変化を統一するための語幹抽出（stemming）の技法等を除けば，NLPの手法はそれほど活用されてはいない．TRECにおいてもまた，NLPに関して肯定的な結果は今のところ出ていないようである[6]．

1つには，NLPのアルゴリズムが複雑であるにもかかわらず，単純な統計的手法に比べて，顕著な性能向上を実現できないという状況がある．しかし，特に日本語テキストに関しては，英語などの言語と比較して，NLPの技法を適用する余地が大きいように思われる．第1に，日本語の場合，語の間に切れ目がなく，語の識別が難しい．第2に，

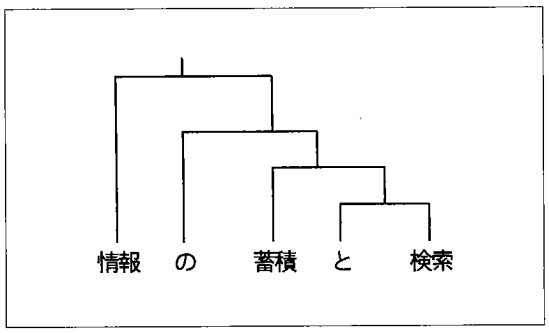


図-3 木構造による語間の従属関係の表現

送り仮名やカタカナの表記における「ゆれ」の問題がある．

言語横断検索（Cross Language Retrieval）もまた，NLPの応用が期待される領域である．一般に，検索質問とは異なる言語で書かれた文献の検索を言語横断検索と呼ぶが，このための方法としては機械翻訳を応用するものと，単純に語句を他の言語のそれに置換するものとがある．さらに後者には，既存の何らかの対訳辞書を使う方法や，いわゆる対訳コーパス（同一内容が書かれた複数の異なる言語でのテキスト）を使って統計的に語句を決める方法などのバリエーションがある．ここでもまた，「NLP対統計的手法」の構図が出てくるが，今回のNTCIRのテストコレクションの活用によって，これらに関する研究の進展が期待される．

■テストコレクションの限界と今後■

TRECでは，単純な適合度順出力や言語横断検索の比較評価以外にも，発話された文献（spoken document）の検索などの多くの研究課題が設定され，それらの問題を探究する契機を作り出している．各課題のためのテストコレクションが整備されれば，本稿でみてきたように，研究の進展に拍車がかかるわけであり，この点，現在のIR技術の研究はいわば「テストコレクション主導型」ともいえる状況になっている．この点，NTCIRにも大きな期待が寄せられることであろう．

しかし，同時に，テストコレクションの限界も認識しておく必要がある．テストコレクションによる評価の拠り所は各文献に対する「適合／不適合」の判定であり，これは通常，少数（たとえば1人か2人）の判定者によってなされ，その結果が客観的な評価基準として使われることになる．しかし，この種の判定は本来高度に主観的なものである．

主として図書館情報学の研究成果によって，同一質問に対する同一文献の適合／不適合の判断が人によって異なることが分かっている[3]．さらに，同じ人でも時間の経過に伴って適合／不適合の判定が変わることさえあるという．これらのことを考えれば，テストコレクションによる評価の妥当性に対する疑問が出てくるのは当然である．検索質問の数が多くなれば，適合判定の変動が平均の計算において相殺されるという期待もあるが，いずれにせよ，このあたりの問題はさらに探究される必要があろう．

検索手法を評価するには，数学や統計学に基づいて理論的に分析する方法と，テストコレクションによって実証的に分析する方法とが考えられる．伝統的なIRの分野では，良くも悪くも，後者の方法が重視されてきた．これは，文献の検索が現実の人間や社会に深く関連した問題であるがゆえのことであろう．しかし逆にいえば，それだからこそ，テストコレクションを使った研究室の中だけの分析からもう一歩外に踏み出すこともまた，今後必要になるかもしれない．

参考文献
1) 神門典子: 情報検索システムの評価プロジェクト, NTCIRワークショップ, 情報処理, Vol.41, No.6, pp.689-697 (June 2000).
2) 日本図書館情報学会文献目録委員会編: 図書館情報学シソーラス第3版, 日本図書館情報学会 (1998).
3) 岸田和明: 情報検索の理論と技術, 勁草書房 (1998).
4) http://trec.nist.gov/
5) Robertson, S. E. et al.: Okapi at TREC-4, NIST Special Publication 500-236 (1994). http://trec.nist.gov/
6) Voorheers, E. M.: Natural Language Processing and Information, Retrieval, Pazienza, M. T. (ed.), Information Extraction, Springer, pp.32-48 (1999).

(平成12年6月28日受付)