3.2 企業研究者から見たテストコレクションと評価ワークシップ

藤田 澄男　（株）ジャストシステム　Sumio_Fujita@justsystem.co.jp

■本音と立場■

情報検索システムを開発している企業の研究者としては，開発している商用システムの評価を行う必要があるのと同時に新しい方式の実験も行う必要がある．

評価にも，商用システムとしての実用性を考慮したさまざまな評価，たとえば，どのようなユーザモデルに対してどの程度有効であるといった視点からの評価と，情報検索の研究で伝統的に行われている精度／再現率による実験室評価があるが，評価ワークショップでの評価は，後者に属する．

ただ実験室評価といっても初期の頃とは比べものにならない規模の文書集合を対象にしており，企業でさえ，単独では，そう簡単に実施できないものである．

そこで評価ワークショップに参加することによって，製品化の際の評価のコストを安く上げることができるという実利的なメリットも考えられるが，これはちょっとリスキーでもある．

良い成績をとれば，研究者側としては，「これで売れなければ，後は営業の問題でしょ！」とごまかすこともできるが，逆の場合は立場がないどころか居場所もなくなりそうである☆1．

我々がNTCIR-1に対してとった立場は，既存の商用システムをベースに，日本語情報検索における句ベース索引言語の効果を評価するというものだった．

商用システムをベースに評価実験を行うと，システムが安定しているので大量の実験をこなすには好都合である反面，大規模な商用システムでは，改造も容易でなく，自由にさまざまな実験設定にチャレンジすることができない．新しい方式を試す実験の可能性が限られてしまうといった問題もある．

コンペティション自体は，1件の検索に数時間もかかるような，およそ非現実的な方法でさえ試すことが可能なほど自由である．しかしながら，これだけの実行時間がかかってしまっては，仮に原理としては優れていてもチューニングに十分な時間をかけられないという点で制約を受けてしまうだろう．

一方，実験の制約は，システムからだけでなく，テストコレクションからもくる．

NTCIR-1では，1）検索対象文書の文書長が短く，比較的均一，2）クエリー・文書ともに豊富なターミノロジーを含む，という性格を持っていて，我々が目的にした句の役割の評価には，非常に好都合だった．

NTCIR-2では，文書集合の性格も変わって，文書長が長くて均一でなくなれば，また別の要素が評価の中心になるだろう．

■意義と弊害■

テストコレクションの意義については，誰もが認めるところだろう．実際，1990年代になって，TRECコンファレンス[5]が始まってから，多くのことが分かってきた．

それ以前は，有用性が認められなかった情報検索の索引言語としての句の利用でも，TRECが始まってから，いつの間にか多くのサイトが採用するようになってしまった．

最近のRIAO2000[6]では，それをより一般化する見地から再検証する動きが見られているが，これもTRECやNTCIRのテストコレクションが充実してきたおかげである．数十年来の議論に，大規模テストコレクションを使った実験で新しい角度から光をあてることができる．

さて，ではこのようなコンペティションスタイルのワークショップにはどういう意味があるのだろうか？

まずは，テストコレクションを作成するためには，コンペティションを通して，さまざまなシステムの出力を集めてプーリングすることが必要になる．

プーリングができなくて，すべての文書について正解判定すれば，33万文書のNTCIR-1でさえ，たとえば50人が手分けして，1人あたり1トピックについて33万文書を調べていったら20年以上はかかってしまうだろう．このように今日の情報検索の実験室評価は，大学はもとより企業でさえ単独で実現するのは難しいような規模の問題になっている．

一方で，統合されたシステム間で比較しても，そこで使われている個別技術を比較したことにはならないという指

☆1 ただし評価ワークショップの結果をシステムの宣伝などに使うことは，IREX・NTCIRなどでは禁止されています．

摘もある．システムのパフォーマンスを精度／再現率で評価すると，統合されたシステムの差が，個々の要素技術の違いの効果を隠蔽してしまい，それが誤った結論を導く可能性があるという見方もある．

そもそも統合されたシステムのパフォーマンスでさえ，1件の検索に数秒のシステムと数時間のシステムを，対等のシステムとして評価することに無理があるかもしれない．

もし，商用システムを比較するならば，平均精度の数%の違いにはさしたる意味はないだろう．商業的な成功にとっては，実験室評価より他にもっと重要な要素があるかもしれない．

仮にフェアな競争の条件が整ったとしても，情報検索システムの力くらべのはずが，参加者の体力・持久力くらべになってしまう可能性もある：前回の評価結果提出の際も，チューニングのためにおそらく何日も徹夜された方たちもいたことだろう．

低コストで信頼度の高い評価を行うというテストコレクションの実利的な成果とは別に，どこかにシステム間の競争という分かりやすいタスク設定の効果を超えた目的を見出す必要があるかもしれない．

そこで，要素技術の評価という目的を設定したとして，どんな要素技術でも，特定のインスタンスを，特定の統合の仕方をした統合システムの中に組み込んで使わなければ，その効用は評価できないだろう．

では，どうしたら，たとえば情報検索における自然言語処理技術の役割のように，特定の要素技術の効用を，特定の文脈外で一般的に評価することができるのだろうか．

1つの可能性は，統合システムのうちある部分を共有したシステム間で比較することである．

現在主流の素性ベースの情報検索システムは，文書やクエリーに対する内部表現を作成する部分と，内部表現内で各タームに重み付けを行う部分に，分けられる．この両方を分けて，たとえばどちらかを共用したシステム間で評価できれば，かなり統合システムを透明化して，要素技術の評価ができるかもしれない．

TRECでは，クエリーの内部表現の提出などを行っているし，NTCIR-2では，トピック表現にタグ付けを行い，そのまま語句ベース検索の内部表現として利用できる形で提供するという．このような実験を何年かにわたって，さまざまな検索課題と文書セットに対して実施することによって，色々なことが分かってくるだろう．

実際，TRECでは，今までにNLPトラック，インタラクティブ検索トラック，クエリートラックのようにさまざまな要素技術に注目したトラックが企画されていた．

■地平線とベースライン■

まったく異なる技術を用いたシステム同士のパフォーマンスが非常に似ているとき，単なる偶然の一致の場合もあるだろう．しかし，そこに現状のさまざまな技術の地平線が存在するのかもしれない．どのような方法を使っても，それより先を見ることのできない機械情報検索の地平線があるのかもしれない．

一方で，ベースラインという用語がよく使われる．

ある特定の統合システムにある要素技術の特定のインスタンスなどを組み込む前と後を比較するための組み込み前のシステムのパフォーマンスである．

組み込み対象のシステムが違えば，当然別のベースラインをもとに比較をする．ある要素技術を組み込むことによって，評価結果がベースラインより，統計的に意義のある程度に改善されれば，その要素技術の組み込みは有効とみなされる．

ベースラインは，毎年変わって，去年のシステムは，今年のベースラインになるかもしれない．

もし，ベースラインが明らかに地平線に近づいているならば，それはタスクの内容を大幅に変える時期がきたということなのだろう．

■評価と評価の評価■

情報検索システムの評価において，適合性という主観的判断を評価に用いることが，議論の対象になってきた[7]．

ある検索課題に関してたとえば「適合する」・「適合しない」の2値で与えられる適合性判断が，文書集合に対して与えられたとする．これをたとえば50の検索課題について3人の異なる判定者の判断を用意する．そうすると3^{50}の異なる正解判定セットができる．それぞれの正解セットは，1つのユーザモデルを反映しているとみなすことができる．判定が，それぞれの課題に対するユーザの専門度・嗜好性・検索意図を反映しているからである．

実際，異なる判定者間で判断の一致をみることは，想像以上に難しいことのようだ．よって異なる正解セットを用いれば結果の平均精度は変わってくる．

ところが，TRECのデータを使って同一システムの複数の異なる実行結果をある正解セットで平均精度を計算して順位付けすることを，さまざまな正解セットを用いて行ってみると，異なる正解セットのために順位が逆転することは非常に少ないという[8]，☆2．

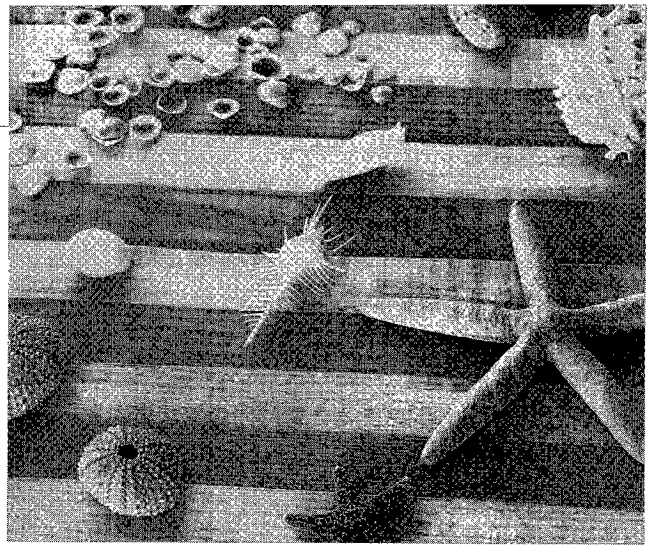
とはいえ，異なるシステム間の実行結果では，正解セットを変えることにより，順位逆転の可能性は上がるし，手

☆2 TREC以前にも，同様の問題意識による評価方法に関する実験が小規模テストセットを対象にして行われていた．文献7)，8) 参照．

動クエリー作成では，自動クエリー作成の場合よりも逆転の可能性は上がるという．

異なるユーザモデルが，適合性に基づく評価に与える影響は，ごく限られたものであるのかもしれない．

1990年代にTRECから始まった評価ワークショップを通して，システムについていろいろなことが分かってきたのと同時に，評価方法についてもいろいろなことが分かってきた．これらの成果は，商用システムの実験室的でない評価にも反映されるだろう．



3.3 NTCIRへの参加から学んだこと

Fredric C. Gey カリフォルニア大学バークレイ校 *gey@ucdata.berkeley.edu*
Aitao Chen カリフォルニア大学バークレイ校 *aitao@sims.berkeley.edu*
Hailing Jiang カリフォルニア大学バークレイ校 *hjiang1@sims.berkeley.edu*
（著・訳）
岸田 和明 駿河台大学 *kishida@surugadai.ac.jp*

■NTCIRへの参加■

カリフォルニア大学バークレイ校は，NTCIRワークショップに参加し，日本語の単言語検索と，日本語と英語との言語横断検索の2つのタスクに取り組んだ．前者においては，バイグラム（bi-gram）訳注1による方法と辞書を用いる方法とを比較検討し，後者においては，NTCIRのコレクション中の日本語と英語が対訳になっている部分から抽出した対訳辞書と，専門用語を含まない辞書を使った機械翻訳とを比較した．ワークショップに提出した我々の論文[10]では，日本語検索および日本語－英語の言語横断検索のための複数の方法を比較している．具体的には，語分割訳注2のための2つの方法と，検索質問を翻訳するための2つの方法をテストした．この研究は，Text REtrieval Conference（TREC）[5]への参加を通じて行った，全文に対する単言語検索および言語横断検索に関する諸研究[9), 11), 14)]の上に成り立っているものである．

日本人以外でNTCIRに参加したいと考えている人々は2つの障壁に直面している．それは，日本語への不慣れさと語の境界の検出である．それまでヨーロッパ言語で研究を行ってきたグループは，語分割のアルゴリズムと諸研究とを学ぶ必要がある．バークレイのグループでは，以前に中国語の語分割の研究[9)]があることと，日本人の研究者が1998年から1999年までバークレイに滞在していたことによって，この障壁は難しいものではなかった．

■文献のランキング■

NTCIRの検索では，我々はすべて，以前にTREC-2で用いた方式[11)]を使って，文献の順位を算出した．TRECのテストコレクションに対する随時検索（ad hoc retrieval）において，この方式は，長い検索質問と人手で修正された検索質問とに対して頑健であることが示されている．他の言語に適用した場合でも（この方式はTRECの英語のコレクション上での訓練結果である），TREC-4のスペイン語，TREC-5の中国語[13)]，TREC-6とTREC-7のヨーロッパ言語（フランス語，ドイツ語，イタリア語）[12), 14)]のように，この方式はよく機能する．すなわち，このアルゴリズムは，語の境界の検出（語分割）さえ可能ならば，どの言語であろうとも，頑健であることをこれまで示してきたわけである．この方法はロジスティック回帰に基づく確率型検索モデルであり，詳細については，我々のワークショップの論文[10)]を参照してもらいたい．

訳注1 たとえば，「情報検索...」ならば，情報，報検，検索，索...のように重複した2文字ずつに分ける．
訳注2 テキストを語の単位に切り分けること．