解説

探しもの見つけます ～情報化社会に役立つ情報検索の技術動向～

岸田 和明（駿河台大学／国立情報学研究所）
kishida@surugadai.ac.jp

賀沢 秀人（NTT コミュニケーション科学基礎研究所）
kazawa@cslab.kecl.ntt.co.jp

◎情報検索への要求の多様化と検索技術の拡大

1970 年代から *80* 年代にかけての情報検索は，図書や雑誌論文などに関する文献情報を収録したデータベースを中心としていたが，インターネットをはじめとする最近の電子的な情報資源の広がりに伴って，その状況は一変した．情報検索に対する要求は多様化し，それまでのデータベース検索サービスが提供していた学術情報やビジネス情報だけでなく，我々の日常的な生活の中で，より一般的かつ幅広い情報を探し出す必要性が高まっている．

これに伴って，従来的な，文書（*document*）を単位とした検索以外の，やや異なった形態の情報検索も重要となっている．たとえば，本来，情報検索の第 *1* の目的は，ユーザが知りたいこと（検索要求）に適合した「情報」の提供にあり，その情報は必ずしも「文書」でなくともよい．大量に蓄積された文書やテキストの中から，ユーザが求めている「探しもの」さえ見つかればよいわけである．あるサッカー選手の次の試合の日程を知りたいと思って *WWW* を検索したとき，そのサッカー選手のホームページが特定されるよりも，直接，その試合日程や放送予定が出力されたほうが便利であろう．この要求を満たすには，検索された文書から特定の情報を抽出する質問応答の技術が必要になる．

さらには，*WWW* のように大量で多様な内容を持つ情報資源が利用可能となった現在においては，単に情報を探してくるだけでは不十分であり，その情報をユーザが活用しやすいように加工・提示することもまた求められている．このためには，検索された文書の内容を短くまとめるためのテキスト自動要約の技術や，文書クラスタリング，検索結果の可視化（*visualization*）が必要になるだろう．そのほか，画像や音声に対するマルチメディア検索や，大量に届けられる新規の情報の中からユーザの求めるものを抽出する情報フィルタリングのような形態の検索も重要な役割を担うようになっている．

同時にまた，従来的な文書検索に関しても，最近の状況に対応して，新しい要求が生まれつつある．たとえば，電子文書の普及によって言語に関するボーダーレス化が進んだのに伴い，言語横断検索（後述）に焦点が当てられるようになったことはその一例である．もちろん，文書検索技術は，広義の情報検索における中核的な存在として，さらに洗練されていく必要もある．たとえば，現在の *WWW* の検索はまだ満足できるものとは言いがたく，より正確かつ高度な文書検索が求められている．

以上のような情報検索に対する要求の多様化・高度化に対応して，実際に，*TREC*[☆1] や *NTCIR*[☆2]，*CLEF*[☆3] といった，情報検索に関する国際的な大規模実験プロジェクト[1] においても，さまざまな研究・開発が進められている．たとえば，本年 *2* 月に新たに開始された *NTCIR* の第 *4* 回目のワークショップにおける研究課題（タスク）には，(*1*) 言語横断検索タスク，(*2*) テキスト要約タスク，(*3*) 質問応答タスク，(*4*) 特許検索タスク，(*5*) *Web* タスクの *5* つが設定されている．

☆1 http://trec.nist.gov/
☆2 http://research.nii.ac.jp/ntcir/
☆3 http://www.clef-campaign.org/

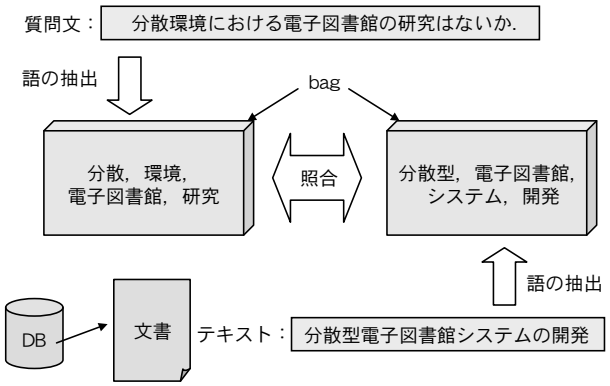

図 -1 “bag-of-words”の間の照合

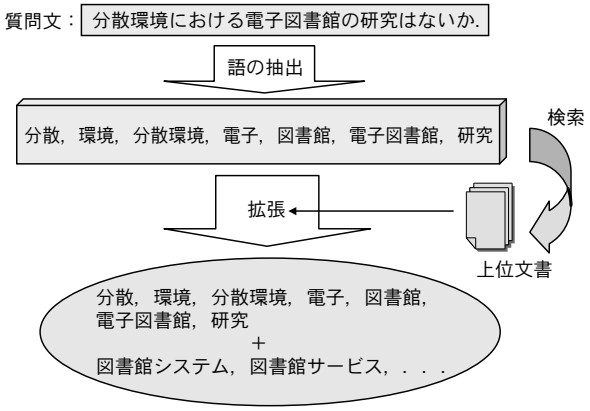

図 -2 擬似適合フィードバック

本稿では，以上の多様な検索技術のうち，特に，従来的な文書検索技術と，質問応答の技術とを中心に，その研究課題や最近の動向を取り上げて解説する．なお，テキスト自動要約については，すでに本誌の特集として詳細な解説がある[2]．

◎擬似適合フィードバック

“bag-of-words”方式

現在の文書検索システムの大部分では，昔ながらの，“*bag-of-words*”（語の袋）方式が使われている．つまり，これは，**図 -1** に示したように，文書のテキストを断片化して索引語の集合へと変換し，それと検索語の集合とを照合する方式である．その結果一致した語に関する何らかの統計量に基づいて各文書の得点を計算し，その得点順に文書を並べ換えれば，ユーザに検索結果を提示できる．ユーザが検索要求を質問文として表した場合も同様であり，文書のテキストを断片化した方法を使って，質問文を語の集合へと変換する．

文書テキストや質問文が語の集合に還元されるということは，それらの語間に存在する関係が捨て去られてしまうことを意味している．それでもこの方式が使われる理由の *1* つは，大量のテキストを迅速に処理するには，それを索引語へと分解し，*B* 木などの探索アルゴリズムを適用することが実用に適っているためである．さらには，複雑な自然言語処理技法を適用して構文関係や意味関係を活用しても，それに見合うほどの顕著な性能向上がもたらされないという経験的な知見もある．“*bag-of-words*”方式の是非は，長年にわたって時折思い出したように繰り返されてきた，情報検索分野における古くて新しい議論であり，おそらく，将来的にも再び取り上げられる研究課題であろう．

質問拡張

図 *-1* に示したように，現在の“*bag-of-words*”方式では，ユーザは，自らの検索要求を検索語の集合（または質問文）で表現しなければならない．しかし，自分自身にとって不明確なことがらを検索しようとするのであるから，ユーザが的確な検索語をうまく思いつくことができるとは限らない．実際に，インターネットの検索エンジンに投入される語の数は平均して *1* ～ *2* 語といわれており，このような少数の検索語から十分な結果を得ることは難しいだろう．

この問題を解決する方法の *1* つは，ユーザが入力した検索語に対して，システムが自動的または半自動的に，語を追加することである．これを質問拡張（*query expansion*）と呼ぶ．この結果，検索集合の意味的な表現がより豊かになり，検索性能の向上がもたらされる可能性がある．

擬似適合フィードバックのしくみ

実際に検索語を追加するには，何らかの既存の機械可読型のシソーラスや，文書集合中の語の共起関係に基づいて自動的に作成された統計的なシソーラスを用いることができる．しかし，これまでの研究結果によれば，これらはそれほど顕著な性能向上はもたらさない．

それに対して，*TREC* や *NTCIR* などの最近の検索実験で頻繁に活用されているのが，擬似適合フィードバック（*pseudo relevance feedback*）である．この場合，検索は，

①初期検索：ユーザによる検索語を使って検索を実行
②質問拡張：初期検索で得られた上位何件かの文書を「適合」であると仮定して，その中から有用な検索語を選別して，元の検索語に追加
③ *2* 次検索：拡張された検索語の集合を使って，検索を再度実行

といった手順で実行される（**図 -2** 参照）．

擬似適合フィードバック研究の課題

擬似適合フィードバックは最近の検索実験ではすっかり「おなじみ」であり，一定の検索性能の向上をもたらすことがほぼ認められている．技術的には，*1990* 年代半ばから後半にかけての *TREC* において研究され，その基本的な方法がおおよそ完成された．*NTCIR* においては，その後，日本語テキストへの適用可能性が検証され，いくつかの修正が加えられている．最近の要素技術の中では，目立ったトピックの *1* つであるといってよいだろう．

検索システムの基本が“*bag-of-words*”である以上，検索性能の向上には，検索語の拡張は避けては通れない．そのための方法として，今のところ，確かに擬似適合フィードバックは有力な候補である．しかし，その実用化に向けては，いくつかのハードルがある．まず，擬似適合フィードバックが検索性能の向上をもたらすといっても，それはあくまでマクロ的な傾向にすぎない．検索実験では，通常，複数の検索課題が用意され，それらに対する平均的な性能をシステム間で比較するが，その「平均」における性能向上を達成しているのにすぎない．つまり，検索課題の中には，擬似適合フィードバックによって，検索結果が逆に悪化する場合もあり，この点の技術改良は必須である．また，この方法は検索を繰り返し実行するため，応答時間やリソースの点では問題を抱えている．このあたりの解決も重要である．

◎言語モデルを応用した言語横断検索

言語横断検索とは

言語横断検索（*cross-lingual retrieval*）とは，検索語と文書とが異なる言語で書かれている場合の検索を指す．たとえば，ユーザが日本語で検索語を投入して，英語の論文データベースを検索するような場合である．各国語で書かれた電子文書が増加し，他国語のテキストへのアクセスが容易になるにつれて，言語横断検索への重要度は高まっている．現在でも，インターネットの検索エンジンのサービスとして，検索語の翻訳機能が提供されていることがあるが，これはその *1* つの現れであろう．

言語横断検索の方法としてはさまざまなものが考案されているが，対訳辞書や機械翻訳システムなどを使って検索語を翻訳し，文書集合の言語に合わせるのが一般的である（ただし文書集合を翻訳する場合もある）．検索語の翻訳さえしてしまえば，あとは従来のベクトル空間モデルや確率的モデル[☆4]を使って，文書得点を計算できる．

ただし，このとき，翻訳の曖昧性（*ambiguity*）が問題になる．たとえば，「情報」という日本語に対して和英辞典を調べてみると，

information, intelligence, a report, news

などが掲載されている．このような場合に，どの翻訳を採用するかが，言語横断検索の性能を左右する．

ベクトル空間モデルや確率的モデルでは，この問題をモデルの外側で解決しておく必要がある．たとえば，辞書に最初に掲載されている語のみを検索語として採用する，あるいは，対訳コーパスを使って翻訳確率を計算し（後述），閾値を超える訳語をすべて投入する，などの方法がある．

言語モデルを応用した検索技術

それに対して，言語モデル（*language model*）を使うと，翻訳確率を無理なくモデルに組み込むことができ，合理的である．検索性能もかなり高く出るので，*TREC* や *CLEF* などの検索実験において，採用するチームが増えている．従来的な文書検索技術としては，ここ *2*，*3* 年における最もホットなトピックであるといってよいだろう．

言語モデルとは，ある言語に関して語（または語の列）が生起する確率を与えるモデルを指す．このモデルは，自然言語処理や計算言語学の分野において発展してきたものであり，コーパスに基づくデータ主導型のアプローチにおいて重要な役割を果たしている[3)]．もともと，文書検索は，自然言語で書かれたテキストに対する処理を基盤としており，これに言語モデルを応用しようとする試みはむしろ自然な流れといえる．

言語モデルを使って文書得点を計算するには，「ある *1* 件の文書が与えられたときに，その文書のテキストから検索語の集合が標本として抽出される確率（あるいは，検索語の集合がその文書から生成される確率）」を考えて，その確率を文書得点とすればよい．このとき，文書中に含まれる語をそれぞれ別個に独立したものとして扱うことが多い．

その結果，統計的自然言語処理におけるゼロ頻度問題（*zero frequency problem*）を考慮すれば，上記の生成確率は，各検索語の抽出確率を掛け合わせて，

$$P(Q \mid d) = \prod_{t \in Q} aP(t) + (1\text{-}a)P(t \mid d), \qquad (1)$$

となる．ここで，

Q：検索語の集合，

☆4 検索のためには，図 -1 に示された 2 つの“bag”を照合して，検索要求に対する各文書の得点を計算する必要があるが，その計算方法としては，線形代数の理論に基づいたベクトル空間モデルや，確率論に基づいた確率的モデルがある．

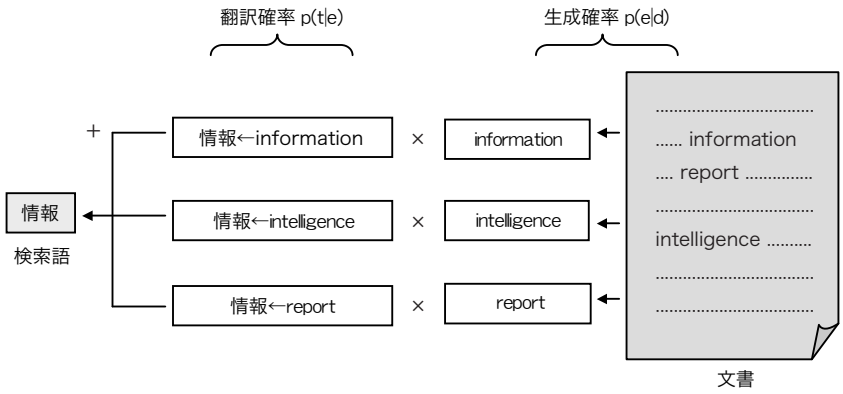

図 -3　翻訳確率を利用した言語横断検索

d：1 件の特定の文書を示す記号，
$P(t)$：語 t が文書集合全体から無作為抽出される確率，
$P(t|d)$：語 t が文書 d から無作為抽出される確率，
a：パラメータ（$0 \le a \le 1$），
である．

直感的には，語がその文書から無作為抽出される確率 $P(t|d)$ のみを掛け合わせればよさそうであるが，実際の文書自体が，その文書についての言語モデルから得られた標本にすぎないと考え，(1) 式では，文書に出現しない検索語の確率が 0 にならないよう $P(t)$ を入れている．これはちょうど，ゼロ頻度問題に対処するための補正（すなわち，ディスカウンティング）に対応している．

(1) 式に含まれる確率の推定量としては，$P(t|d)$ については「その文書に含まれる索引語の延べ語数に対する語 t の出現回数の割合」，一方，$P(t)$ については 1 つの近似として，「全文書数に対する，語 t が出現した文書の割合」が使われることが多い．その結果として，言語モデルを応用した (1) 式は，ベクトル空間モデルや確率的モデルと同様に，*tf-idf* の原理[☆5] に沿ったものとなることが知られている [4]．

翻訳確率の組込み

ある言語における語が他の言語の語に翻訳される確率を翻訳確率（*translation probability*）と呼ぶ．訳語の曖昧性が生じたときに，これを確率的に解決するための方策の 1 つである．翻訳確率を言語モデルによる (1) 式に組み込むには次のようにすればよい．まず，次の記号を導入する．

T_t：検索語 t に対するすべての訳語の集合

$P(t|e)$：文書集合中の語 e が検索語 t に翻訳される確率（翻訳確率）

これらを使えば，(1) 式中の $P(t|d)$ は，

$$P(t|d) = \sum_{e \in T_t} P(t|e)P(e|d)$$

となる（**図 -3** 参照）．これはちょうど，

文書→文書中の索引語→検索語

の遷移確率を考えて，中間の状態に関して総和をとるしくみになっている．そこで結局，(1) 式は，これをそのまま代入して，

$$P(Q|d) = \prod_{t \in Q} \left[aP(t) + (1\text{-}a) \sum_{e \in T_t} P(t|e)P(e|d) \right]$$

となる（なお，翻訳確率の組込みには他の方法もある）．このように翻訳部分が文書得点の計算式中に明示的に入ってくるので，言語横断検索のためのモデルとしては，翻訳をその外側で処理しなければならない他のモデルに比べて，よりエレガントである．

あとは翻訳確率さえ手に入れば，文書得点を計算できる．この確率は基本的には，対訳語のリストや対訳コーパスを使って求められる．後者の対訳コーパスとは，異なる言語で書かれた同一内容のテキストが並列したコーパスを意味する．そこから文や段落の並置（*alignment*）を構成し，その共起情報から，統計的に翻訳確率を求める．この推計モデルには，*IBM* で開発されたもの [5] などがあり，実際に *TREC* や *CLEF* などで使用されている．特に，昨年 11 月の *TREC2002* の会議では，英語とアラビア語との間の横断検索に関して，*IBM* の翻訳確率の推計モデルを使った研究成果がいくつか発表され，多くの聴衆の関心を集めていた．

◎「文書」の検索から「情報」を単位とした検索へ

以上説明した文書検索技術は，基本的には，ユーザの検索要求に適合した「文書」を的確に特定することを目

☆5 tf は検索語が当該文書中で出現した回数を意味する．この回数が多いほど，文書の得点を高くすればよい．しかし，tf をそのまま使うと，主題に関係なく数多くの文書に出現するような非専門語の重みが不当に高くなる可能性があるので，出現文書数の逆数（idf）によって重みを下げる必要がある．これが tf-idf による重み付けである．

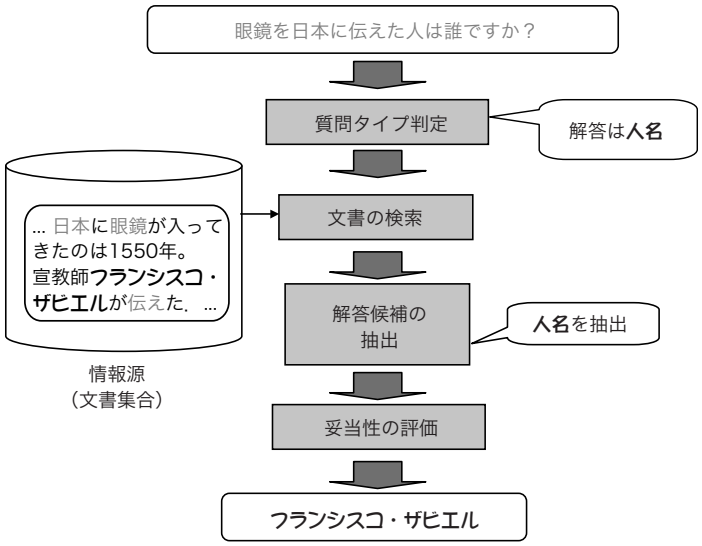

図 -4　典型的な質問応答システムの構成

的としている．しかし，本稿の冒頭で述べたように，ユーザの検索要求を満たすための単位は「文書」である必要はない．文書自体を出力せずとも，ユーザが求める情報をシステムが提供することは可能であるし，むしろそのほうが望ましいことも少なくない．

一口に「ユーザの求める情報」といっても，それが情報源の中に現れる形式はさまざまである．たとえば,「日本の最初の総理大臣の名前」について知りたい場合は，単に「伊藤博文」という名前が分かれば十分だが，「日本の最初の総理大臣」について知りたい場合には，伊藤博文の生涯やその時の時代背景などについてまとめたモノを提示することが必要となる☆6．

このように,「情報」を単位とした検索を行うためには，文書やホームページのようにあらかじめ決まった単位で，情報を探すだけでは不十分であり，検索要求に応じて動的に検索の単位を変更することが必要となる．また，検索の単位が，文書やホームページといった従来の情報検索で対象とした単位とは異なるため，検索結果のスコアリングについても従来とは異なる手法が必要となる．

ここでは，検索対象を動的に決定する必要のある情報検索タスクの例として，*TREC* の質問応答タスク（以下，*TREC-QA* と呼ぶ）について紹介する☆7．

TREC-QA では，端的に答えられるような事実関係に関する英語の質問文が与えられたときに，その解答を，約 *100* 万記事の新聞データベースから探し出すことが求められている．また, 検索対象となる「解答」に関しては,「解答以外の余分な情報を含まないこと」が正解の条件となっている．そのため, 質問応答システムにおいては, 質問に応じて検索単位を適切に決定する技術と，記事中からそれを切り出し，妥当性を評価する技術が非常に重要とされている．なお，典型的な質問応答システムの構成を**図 -4** に示しておく．

検索単位の決定

通常，質問応答システムにおける検索単位の決定は，質問が「何」を尋ねているのかを判定することから始められる（以下, 質問タイプの判定と呼ぶ）．たとえば,「日本の最初の総理大臣は誰ですか」という質問は人名を尋ねており,「伊藤博文の出身地はどこですか」という質問は地名を尋ねている，ということを判定する．このような判定を行うことで，質問応答システムは文章中の人名あるいは地名を検索対象とすればよいことを認識できる．逆にいえば，ここを間違えてしまうと，適切な検索結果を得ることは非常に困難になるため，質問タイプ判定は質問応答システムの要ともいえる．

TREC-QA に参加した多くのシステムでは質問タイプ判定に相当する処理が行われており，用いられる質問タイプの分類も「人名，地名，組織名」といった比較的粗いものから,「国家元首名，大学教授名，都市名，州名」といった細かいものまで多岐にわたっている．また，人名や地名といった固有名称以外にも，花の名前といった事物のクラス名まで含んだ幅広い分類を用いているシス

☆6 ここでの「モノ」は必ずしも文書である必要はなく，場合によっては映像や音声を含んだり，インタラクティブな要素を持つコンピュータプログラムである可能性もある．

☆7 TREC のタスク設定は毎年変更されている．本稿では 2002 年に行われた最新の TREC におけるタスク設定をもとに述べる．

テムも多い.

しかし，このように質問タイプを限定してシステムを作ることは，想定した質問にはよく答えられるが，それ以外の質問にはまったく答えられないという結果になりがちであり，このアプローチに対する批判や不満も多い．そのため *TREC-QA* では，質問タイプの判定がそれほど大きな意味を持たないと考えられる人物や事物の説明を求める質問（たとえば，「伊藤博文とは誰ですか」「*DNA* とは何ですか」）や，出来事の原因や推移を尋ねる質問（「伊藤博文が暗殺された原因は何ですか」）をテストセットに含めることが検討されている．

「情報」の抽出

TREC-QA においては解答となる「情報」は，文書やホームページといった与えられた単位では存在せず，システム自身がテキスト中から適当な範囲を抽出してくる必要がある．そのため，前節で述べたような質問タイプの判定を行うシステムでは，テキスト中から質問タイプに該当する部分を抽出する必要がある．

このように，あらかじめ決められたタイプに該当する文字列をテキスト中から見つけ出す技術は，情報抽出（*Information Extraction*）や固有表現抽出（*Named Entity Extraction*）と呼ばれる分野で研究されてきたものである．

情報抽出とは，会社の合弁といったイベントに関して，合弁した会社の名称や社長名，合弁が実施される日時などの，あらかじめ決められた情報を文書中から抜き出すタスクである．また，固有表現抽出とは，イベントとは無関係に人名や会社名を文書中から抜き出すタスクである☆8．

従来，情報抽出や固有表現抽出は文書検索と独立して研究が進められることが多かった．しかし，本来，これらの技術はいずれも情報検索技術の一部と考えるべきものであり，*TREC-QA* に代表される質問応答は，現在は個別に研究されているこれらの技術を，本来のあるべき姿に統合する *1* つの試みととらえられる．

一方で，物事の説明を求める質問などに対して，固有表現のような決まった解答単位を想定することは難しい．しかし，今後，より現実的な質問応答技術を構築する上で，単に情報を「抜いてくる」のではなく，質問に応じて情報を「構築する」技術が必須となると考えられる．実際，昨年の *TREC-QA* 参加者の間では，単に解答を抜き出すだけではすまない“*How...*”や“*Why...*”といった質問をテストセットに含めることに肯定的な意見が多く聞かれた．

妥当性の評価

文書検索においては，ある文書の内容が丸々検索要求に適合するということは稀である．そのため，文書の適合性は「適合するか否か」という二者択一で決められるものではなく，「完全に適合する」文書から「完全に適合しない」文書まで，さまざまな文書があると考えるのが適切である．また，検索要求に対してもさまざまな適合の仕方があるため，適合している可能性のある文書に関しては，万遍なく探し出すことも重要である．

しかし，*TREC-QA* に代表される質問応答では，多くの場合解答は正しいか否かのどちらかであり，また，複数の正解がある場合でも，どれか *1* つだけ解答すればよいということが多い．つまり，質問応答における「情報」の妥当性評価では，再現率はそれほど高くなくてもよいが，精度は非常に高くなるような手法が望ましいといえる☆9．

この傾向は，*WWW* などの冗長性が高い情報源を対象にする場合には，さらに顕著になると考えられる．極端な話，ある質問に対する正解が *1* 万回出現する情報源があったとすると，たとえ再現率が *1%* であっても，十分正解に辿りつく可能性がある．

上記のような状況は極端であり，ある意味，*TREC-QA* 流の質問応答に特有の状況であるともいえる．しかし，文書という単位を離れた「情報」の検索を行う際には，従来の文書検索とは違った適合度の評価方法が必要という点で，今後の情報検索技術の展開に重要な示唆を与えていると思われる．

◎冗長性の削減～情報の Novelty への挑戦

インターネットの検索エンジンを用いると，とても目を通すことのできない分量の検索結果が出力されることがある．もちろん，検索結果のすべてが必ずしも検索要求に適合しているわけではないが，それを差し引いたとしても，同じような情報が何度も繰り返し現れるばかりでイライラする，ということはしばしば起こる．

このような場合，たとえ有用な情報が検索結果の中に存在しても，冗長な情報に埋もれてしまい，ユーザが到達できないということが起こりがちである．情報を活用

☆8 Message Understanding Conference Homepage, http://www.itl.nist.gov/iaui/894.02/related¥_projects/muc/

☆9 実際，TREC-QA では解答候補を含む記事を絞りこむ際に，「質問文中のキーワードをすべて含む」といった「きつい」条件を用いる方が，tf-idf 等の「緩い」条件で検索を行うよりも解答の正確さが上がるという報告がなされている．

するという立場からすると，このような状況は検索に失敗しているのと実質上同じである．したがって，あるべき情報検索の姿からいえば，単に情報を探してくるだけではなく，重複する情報はまとめてコンパクトなかたちでユーザに提示することが望まれる．

ここでは，情報検索により見つかった情報から，冗長性を排除して真に「新しい」情報のみ残すことを目的としたタスクとして，*TREC* の *Novelty* タスク（以下 *TREC-Novelty* と呼ぶ）について紹介する☆10．

TREC-Novelty では，検索要求と新聞記事の集合が与えられたときに，(1) まず，検索要求に適合する文をすべて見つけ出したのち，(2) 適合文の中から，それ以前に出現した適合文と重複した内容を持つものを削除する，という 2 段階の処理を行うことが要求される．このタスク設定は，冗長性の削減を検索技術の評価項目として積極的に掲げている点で，注目に値する．

しかしながら，現時点では *TREC-Novelty* は完全に成功しているとは言いがたい．というのも，事件の速報記事など一部の場合を除き，文全体で伝える情報が完全に重複することは稀であり，「冗長な文」を多く含むようなデータを準備することは容易でないためである．実際，*2002* 年の *TREC-Novelty* のテストデータには，完全に内容が重複する適合文はほとんど存在せず，「冗長性の削除」を積極的に行ったシステムほど成績が悪くなるという傾向がみられた．

このことは，情報の冗長度削減を目的とした研究をする際には，(1) 情報の単位を何にするか，(2) いかに冗長な情報源を用意するか，という点に注意を要することを端的に表している．これまで，情報の冗長性については，インターネットの検索エンジンでの出力結果などを念頭に，漠然とした認識で語られることが多かった．しかし，現在の情報検索研究で用いられている文書集合の量は，インターネットに比べてはるかに小さく，検索要求を注意して選ばない限り，冗長性を持つ適合文書集合が得られない可能性は高い．

インターネットがこれだけ普及した現在においては，情報の冗長度削減技術は必須の技術といっても過言ではない．今後の技術発展を促すためにも，上で述べたような問題をクリアしていくことは重要な課題である．

◎今後の情報検索技術の発展のために

図 *-1* に示されているように，文書検索技術は依然として，“*bag-of-words*”に依拠しており，それに基づく検索の性能向上を目指して研究が進められている．たとえば，ユーザによって入力される検索語の表現の不十分さを補うための技術が擬似適合フィードバックであり，文書テキストと異なる言語で検索語が投入された場合に対応するための手段が言語横断検索である．それに対して，本稿の後半で説明したように，「文書」の単位ではなく，さらに動的な「情報」の単位を設定し，そのためのテキスト処理技術を模索する動きも着実に進んでいる．その試みは，*Novelty* の検出など，より高度な方向へと向かっており，今後，広義の情報検索技術の研究成果を統合し，その実用化を目指すことこそが，文書やテキストの大規模な集合の中からユーザが本当に求めるものを探すための技術の実現へとつながっていくと考えられる．

参考文献

1) 神門典子編：特集 情報検索の力くらべ，情報処理，*Vol.41, No.8, pp.897-924*（*Aug. 2000*）．
2) 奥村　学，久光　徹，増山　繁編：特集 テキスト自動要約，情報処理，*Vol.43, No.12, pp.1286-1316*（*Dec. 2002*）．
3) 北　研二：確率的言語モデル，東京大学出版会（*1999*）．
4) *Hiemstra, D.: A Linguistically Motivated Probabilistic Model of Information Retrieval, Research and Advanced Technology for Digital Libraries, Springer, 2000*（*LNCS 1513*）*, pp.569-584* .
5) *Brown, P. F. et al.: The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, Vol.19, No.2, pp.263-311*（*1993*）．

（平成 *15* 年 *4* 月 *28* 日受付）

☆10 ここで紹介するタスク設定は 2002 年の TREC のものである．