言語処理は情報検索に役立つか？

○ 徳永 健伸 (東京工業大学 大学院情報理工学研究科)

1 はじめに

情報検索の研究には半世紀近い歴史があるが，その根幹には学術情報をどのように配布するか，あるいは収集するかという問題意識があった．したがって，情報検索の検索対象は，書籍や学術論文などのように均質で閉じた世界のものが中心あった．これに対して，1990 年代に爆発的な普及をとげたインターネットは情報検索の研究分野に大きなインパクトを与えた．インターネット上の情報は，変化の速度，絶対量，非永続性，非均質性，媒体の多様性，開放性などの点で従来の情報検索の研究が対象としていた情報とは異質である．このように質的に異なる検索対象を扱うためには，これまで情報検索で用いられてきた手法では必ずしも十分ではない．より知的で性能のよい情報検索システムが求められている．

2 情報検索：伝統的アプローチ

情報検索の基本的なモデルを図 1 に示す [2]．このモデルでは，現実世界の情報が文書で表現されていると仮定し，文書をコンピュータで扱えるような内部表現に変換する．一方，ユーザの情報要求も検索質問という形式で表現し，これをコンピュータで扱えるような内部表現に変換する．そして，これらの内部表現を比較することによって，ユーザの情報要求に適合した情報をみつける．

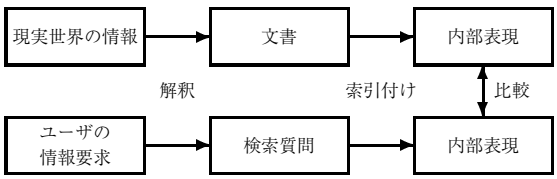

図 1: 情報検索の基本モデル

以下，情報検索の基本概念について簡単に述べる．

- **索引付けと検索モデル**
  文書をコンピュータの内部表現に変換する処理は「索引付け」と呼ばれ，情報検索の研究分野の中心的な研究テーマである．一般的には，文書中から抽出した索引語の集合で文書を表現する．索引語としては語や，語の活用部分を取り除いた語幹を用いる．また，各索引語にはその索引語の重要度を表わす重みを付与することもある．このような枠組では，文書は索引語の重みを要素とするベクトルで表現できる．検索質問も同様に表現することにより，各文書が検索質問にどれだけ適合しているかを 2 つのベクトルの間の類似性に帰着できる．これがいわゆるベクトル空間モデルである．

- **検索質問拡張**
  同じ概念を表現するのに文書中と検索質問中で異なる言語表現が用いられていると，文書と検索質問のベクトル表現を比較する際に問題となる．索引付けの段階で，使用する索引語の語彙を制限する方法もあるが，現在一般に用いられるのは検索質問拡張と呼ばれる手法である．検索質問拡張では，検索質問中に出現する索引語をそれと同義あるいは関連する索引語の集合に置換し，検索をおこなう．

- **適合性フィードバック**
  一度の検索でユーザが必要な文書集合を得ることはまれであり，通常は検索結果をユーザが吟味し，必要ならば検索質問を洗練することによって，よりよい検索結果を得る努力をする．このようなユーザからのフィードバックを一般に適合性フィードバックと呼ぶ．

3 言語処理：統計的アプローチ

1980 年代後半から，言語処理の研究分野ではコーパスに基づく言語処理と呼ばれる研究の流れが起った．これは，それまで人手で構築していた言語知識を大量の言語データから (半) 自動的に抽出しようとする研究アプローチである．もちろんこの背景にはハードウェアの性能の向上と低価格化，大量の電子化テキストの流通がある．特に後者に関しては Web の普及とも関連がある．

コーパスに基づく言語処理で用いられる手法は統計的なものが多く，やはり統計的手法を主として用いてきた情報検索の研究分野とは技術的に類似点が多い．これは，近年，多くの言語処理研究者が情報検索の研究に取り組むようになってきた背景のひとつである．

"Natural language processing and information retrieval"
Tokunaga Takenobu (Tokyo Institute of Technology)
日本音響学会招待講演 (2000.3, pp. 31–32.)

4　情報検索：言語処理の利用

情報検索の性能を改善するためのひとつの方向として言語処理を取り入れることが考えられる．インターネット上にはマルチメディア情報があふれているとはいえ，大部分の情報はテキストによって表現されているからである．言語処理研究の目的のひとつはテキストからその意味内容を取り出すことであるから，言語処理の研究成果を情報検索に利用しようとするのは自然な流れであるといえる．

過去にも情報検索の研究の中で言語処理を導入する試みは少なからずあったが，いずれも成功しているとはいえない．その理由としては言語処理の技術が十分に成熟していなかったことや必ずしも最先端の言語処理技術を使っていなかったことなどが考えられる．このような過去の試みがうまくいかなかったことが情報検索の研究者に言語処理に対する猜疑心をいだかせる原因となっている．

現在では言語処理の技術も進歩し，少なくとも形態素解析や統語解析のレベルでは最先端の技術が誰でも簡単に利用できるようなツールとして整っている．また，さまざまな言語資源も整備されてきている．言語処理技術を本格的に情報検索に利用する土壌はやっと整ったといえよう．

これまでに，言語処理技術を情報検索に応用した例として，索引語の洗練やシソーラスの自動構築などがある．文書や検索質問の索引付けは，英語であれば語の間の空白を手がかりに，日本語であれば文字種などを手がかりにして，いずれも表層的なテキスト処理によっておこなうことが多い．これに対して，形態素解析を導入すれば品詞情報が得られたり，特に日本語の場合は正確な語の境界を同定することが可能になる．さらに統語解析をおこなうことによって，複合名詞などの名詞句や係り受けの関係など，より多くの情報を含む索引語を抽出することができる．

また，前述した検索質問拡張では，ある語の類義語あるいは関連語としてどのようなものがあるかをあらかじめ定義しておく必要がある．このような知識はシソーラスと呼ばれ，言語処理の分野では広く利用されてきた知識である．特にコーパスに基づく言語処理では，シソーラスを言語データから自動的に構築する研究が盛んにおこなわれている．この技術を利用すれば，検索対象となる文書集合の分野に適したシソーラスを自動的に構築することができる．

このように，最近になって言語処理技術を情報検索に利用しようとする研究は盛んになっており，一部では成果を上げつつある．しかし，言語処理を導入したからといって，必ず従来の方法を凌ぐ性能が得られるわけではない．そのひとつの理由として評価尺度の問題が考えられる．情報検索システムの性能は，再現率と精度で測られることが多い．再現率は検索すべき文書をどの程度漏れなく検索できたかを表し，精度は検索すべき文書が実際の検索結果の中にどの程度含まれていたかを表わす．この両者はトレードオフの関係にあり，両方の尺度が高いほうがよいシステムであるといえる．

たとえば，日本語の文書において形態素解析をした結果得られる語を索引語として用いた場合と漢字のバイグラムを索引語として用いた場合を，再現率・精度の尺度で評価しても必ずしも顕著な性能の差はないかもしれない．しかし，これによって，形態素解析は不要だと結論するのは間違いであろう．適合性フィードバックをおこなうことを考えてみて欲しい．適合性フィードバックでは，システムが適合すると判断している文書を提示すると同時に，システムが重要だと判断している (重みの大きい) 索引語をユーザに提示しフィードバックを得ることもある．このときユーザはバイグラムのリストを見せられるのと，言語学的に意味のある語 (形態素) のリストを見せられるのでは，どちらを好むだろうか？このような種類の評価は伝統的なの情報検索の分野ではほとんど考慮されなかった．

5　音声認識技術の利用

音声情報を音のレベルで直接検索することは野心的だが現実的ではないだろう．しかし，音声入力の利点はある．情報検索において，ユーザが入力する索引語の数は一般に少ない．これはユーザの情報要求に関する情報が少ないことを意味する．たとえば，[1] では，音声による索引語の入力を許すことによって，ユーザが索引語を入力しやすくなり，精度 80 ～ 90%程度の音声認識器でも，キー入力に近い性能が得られた例が報告されている．これは対象言語が中国語だという事情もあろうが興味深い結果である．

6　おわりに

本稿では情報検索の概要と言語処理の貢献についてかけあしで述べた．本稿の標題に対する筆者の答えはイエスである．今後，両分野の研究者の交流によって両者のギャップが埋まることを望みたい．

参考文献

[1] L. Chen, H. Pu, M. Chen, H. Chen and M. Lee, “Natural language information retrieval with speech recognition techniques for Chinese network resources discovery”, in Proc. of the Workshop on Information with Oriental Languages, pp.135-142, 1996.

[2] 徳永健伸,「情報検索と言語処理」，東京大学出版会，1999.