言語処理を利用した知的情報アクセス —検索，抽出，要約，分類，QA—

徳永　健伸

インターネットの普及により，世の中に大量の情報があふれかえるようになった今日，我々はコンピュータの助けなしにインターネット上の情報に効率よくアクセスすることができなくなった．インターネット上にはマルチメディア情報が多いとはいえ，情報の大部分は言語で記述されている．言語をコンピュータで理解することを究極の目的とする言語処理の研究は，ますます増大する情報に効率的かつ高度な方法でアクセスするのに役立つであろう．本稿では，言語処理を利用した情報アクセスの技術として，情報検索を中心に，その関連技術として情報抽出，文書要約，文書分類，QA システムについて紹介する．

キーワード：言語処理，情報検索，情報アクセス

1．はじめに

1990 年代から爆発的な普及をとげたインターネットによって世の中には電子化された文書があふれかえるようになった．特に，Blog に代表されるように，WWW（World Wide Web）は，個人による情報発信の敷居を下げ，インターネット上の文書の量を飛躍的に増大させた．その結果，現在ではインターネット上の文書に効率的にアクセスするためにはコンピュータの支援が必要不可欠となっている．

おりしも 1990 年代から，計算言語学・言語処理の研究分野ではコーパスに基づく言語処理という研究の流れが起った[2]．「コーパス」とは新聞記事や小説などのように，実際の言語使用の用例を集積した言語データのことである．この研究手法は，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ものである．そこで用いられる道具立てはさまざまな統計的手法や機械学習の手法であり，そのために大量のデータは必要不可欠となる．

これらの背景をふまえ，1990 年代以降，言語処理の分野では，その応用分野として大量の情報に効率的にアクセスするための諸技術が盛んに研究されてきた．情報検索はその代表例である．情報検索の研究には半世紀にわたる歴史があるが，その本来の動機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であった．したがって，その検索対象は，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，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対象を扱うためには，それまでの情報検索の手法では必ずしも十分ではない．言語処理などを援用した，より高度な情報アクセスの技術が求められるようになってきている．

本稿では，言語処理の応用分野としてこの 10 年間で盛んに研究されてきた情報アクセスの諸技術について解説する．特に情報検索を中心に紹介し，その関連技術として情報抽出，文書要約，文書分類，QA システムについて紹介する．

2．情報検索

情報検索は，広い意味ではユーザの情報要求を満足する情報を知識源から見つけ出すことである．ここで，情報要求とは，ユーザがある目的を達成するために現在自分がもっている知識では不十分であると感じている状態のことをいう．すなわち，情報検索はユーザの抱える問題を解決するための情報を見つける一種の問題解決であるともいえる．

図 1 は情報検索の概念図である．この図では，現実世界の情報が文書で表現されていると仮定し，「索引

とくなが　たけのぶ
東京工業大学 大学院情報理工学研究科
〒 152-8552 目黒区大岡山 2-12-1

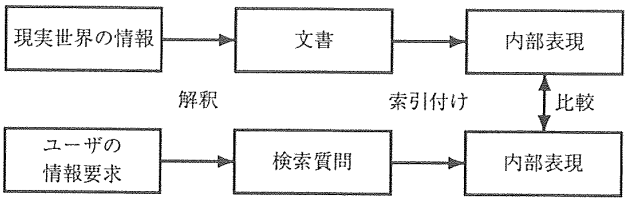

図1　情報検索の概念図

付け」と呼ばれる処理によって文書をコンピュータで扱える内部表現に変換する．同様に検索質問（query）の形式で表現されたユーザの情報要求も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見つける．図中の「解釈」の処理は人間によって行われることを仮定しており，Webの検索エンジンも含めていわゆる情報検索システムの守備範囲には入っていない．以下，情報検索の基本的な概念についてその概要を述べる．

2.1　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．多くの情報検索システムでは，文書中の語の順序関係や構文を無視し，文書の内部表現として索引語の集合を利用する．このため，このような表現方式はBoW（bag of words）と呼ばれる．索引語としては語や，語の連続，あるいは語と語の統語的・意味的関係などを用いる．

各索引語にはその索引語の重要度を表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．類似性の尺度としてベクトルの余弦などがよく用いられる．これがいわゆるベクトル空間モデルと呼ばれる検索モデルである．

2.2　検索質問拡張

ベクトル空間モデルに代表される多くの検索モデルは，文書中の索引語と検索質問中の索引語の表層的な一致に依拠している．しかし，「コンピュータ」と「計算機」など表層的には異なる文字列が同一の概念を表すこともある．このような表現上の違いは検索漏れを起こす原因となる．これを解決するために，検索質問拡張（query expansion）と呼ばれる処理を行う．検索質問拡張では，語や句などをあらかじめ意味の類似性によって分類したシソーラスあるいはオントロジーと呼ばれる知識を用い，検索質問中の索引語をその類義語に展開し，検索質問に追加する．上述の例では，「コンピュータ」と「計算機」がシソーラス中で類義語として定義されていれば，検索質問中の「コンピュータ」に加え，「計算機」も検索質問の索引語として自動的に追加する．これによって，「計算機」は含むが「コンピュータ」は含まない文書も「コンピュータ」という検索質問によって検索できる．

2.3　検索システムの評価

情報検索システムの性能評価には，効率性と有効性の2つの側面がある．効率性はユーザが許容できる時間的あるいは経済的範囲でシステムが答えを返すかどうかに関する指標である．一方，有効性はシステムを使うことによって，必要な情報を漏れなく（完全性），しかも，必要なものだけを（正確性）検索できるかどうかに関する指標である．これらの完全性と正確性はそれぞれ，再現率（recall）と精度（precision）という尺度で計測される．つまり，再現率は検索すべきものをどれだけ漏れなく検索できるか，精度は検索したものの中にほんとうに必要なものがどれだけあるかを表す尺度である．一般に再現率と精度はトレードオフの関係にあるので，これらをまとめた指標として両者の調和平均であるF尺度が用いられることもある．

システムを評価する際に，ユーザが必要な情報かどうかをどのように判断するかということが問題となる．仮にこのような判断ができたとしても，精度は検索結果に対するユーザの判断によって計算することができるが，再現率は，理論的にはユーザがシステム中のすべての情報を吟味した上で判断する必要があるため，事実上計算不可能である．通常は，あらかじめ検索質問とそれに適合する文書集合の正解を判定したテスト・コレクションと呼ばれる評価用のデータを用いて再現率や精度を計算する．

2.4　言語処理の利用

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，その大部分の情報は言語によって表現されている事実を考えると，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．初期の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．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．このような状況で，言語処理技術を情報検索に利用しない積極的な理由は見あたらない．

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，いずれも表層的な文書処理によって行うことが多い．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．さらに統語解析を行うことによって，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，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行われている．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．

3. 情報抽出

情報検索の目的がユーザの情報要求に適合する文書集合を見つけることであるのに対し，情報抽出はあらかじめ指定された情報を文書中から抽出することを目的とする．情報検索の検索質問に比べ，情報抽出では，どのような情報を抽出するかをより詳細に指定しておく必要がある．文書からいわゆる5W1H（誰が(who)，いつ(when)，どこで(where)，何を(what)/どれを(which)，どのように(how)）の情報を抽出するのは典型的な情報抽出である．例えば，図2の台風に関する記事を解析して図中の表を抽出する．この表のコロン(：)の左側はあらかじめ指定された情報であり，その値を記事から抽出することが情報抽出の役割である．

この例からもわかるように，情報抽出は通常の構造化されていない文書を構造化文書あるいはデータベースのレコードに変換する処理であると考えることもできる．通常の文書からこのような構造を抽出できれば，

「台風6号は6日正午現在，日本の南海上にあって，時速約25キロで西北西に進んでいる．中心の気圧は992ヘクトパスカル．中心付近の最大風速は25メートル，最大瞬間風速は40メートルの模様.」

⇓

台風ID	：6
時刻	：2007年8月6日 12:00
位置	：日本の南海上
速度	：約25キロ/時
方向	：西北西
中心気圧	：992ヘクトパスカル
最大風速	：25メートル/秒
最大瞬間風速	：40メートル/秒

図2　情報抽出の例

その結果をデータベースに格納して，データベースによる検索に用いたり，データマイニングの技術を使ってテキストマイニングをしたり，あるいは抽出された情報から元の文書の要約を生成するなど，その応用範囲は幅広い．最近では，ゲノム解析やWeb文書の解析のように大規模なデータを扱う応用分野への適用も始まっている．

情報抽出の手法は大きく規則ベースのアプローチと統計的手法によるアプローチの2つに分類できる．初期の研究では，対象領域に特化した抽出規則を人手で記述する規則ベースのアプローチをとる研究もあったが，対象領域ごとに抽出規則を人手で記述するのはコストがかかる．例えば，台風の情報を抽出するために作成した抽出規則をそのまま地震に関する情報を抽出するためには使えないことは自明であろう．そのため，抽出対象にあらかじめ印を付けたコーパスから抽出規則を自動的に学習する統計的な手法も研究されている．

情報抽出に関しては，対象領域への依存性はある程度避けられない問題であるが，できるだけ領域依存性を低減するために，対象領域に依存しない要素技術として，名前の同定（named entity recognition）や参照関係の同定を設定し，各要素技術について性能を改善する試みが続けられている．

名前の同定は，文書中の人，組織，場所などの名前や，時間表現，数量表現などを同定する技術である．これらの情報は，表のスロットを埋める情報となる可能性が高く，対象領域に依存しない方法で同定することが可能であることから要素技術として研究されている．

表のスロットに埋めるべき情報は場合によっては代名詞などの照応表現の形で文書中に現われる．このような場合は，その照応表現がどの実体を指示している

かを同定する照応解消の技術が必要となる．

情報検索に比べると，情報抽出における言語処理の役割はより本質的である．情報検索の分野でも，言語処理を利用して索引付けをより精密にし，性能を改善する試みが行われてきた．しかしながら，このような索引付けが一般的に情報検索の性能を大幅に改善するというコンセンサスは今のところ得られていない．これに対して情報抽出は，言語処理を導入して語と語，あるいは概念と概念の間の関係を同定しないと解けないタスクである．この意味で情報抽出は情報検索よりも言語処理の応用としては適しているといえる．

情報抽出の評価は情報検索の評価と同様に精度や再現率によって行うことが多い．情報抽出における精度は表中の埋まったスロットがどの程度正しいものか，再現率は埋めるべきスロットがどの程度埋められたかによって評価する．情報抽出では，抽出すべき情報が詳細に指定されているので，正解の判定が情報検索に比べて容易である．実際に抽出された情報が正しいかどうかは抽出の対象となった文書を見れば判定できることが多い．また，同じタスクを人間によって行うことができるので，人間の性能とシステムの性能を比較することが容易にできる．

4. 文書要約

2.1節で述べたように，情報検索では，文書の内容を索引語の集合という形式で表現するのが一般的である．このための索引付けは，文書の内容をよく表す特徴的な語を漏れなく抽出することを目的としている．この意味では，索引付けは，文書からコンピュータのための要約を生成する処理だと考えることもできる．これに対して，文書の自動要約は，文書から人間のための要約を生成することを目的とする．

文書の要約とはその文書で記述されている中心的な話題を簡潔にまとめたものであると定義できる．Paiceによれば要約は，その機能の観点から，以下の4つに分類できる[3]．

(1) 判断材料としての要約（indicative）
(2) 内容情報を提供する要約（informative）
(3) 評価を含む要約（critical）
(4) 比較を含む要約（comparative）

(1)の機能は，読者にその文書を読むかどうかを判断する情報を与える機能である．読者は内容を完全に把握できなくても，その文書が自分の現在の関心に関係があるかどうかを判断できればよい．したがって，索引付けの結果として得られる索引語の羅列でもある程度この機能をはたすことができる．(2)の機能は文書の内容を読者に伝える機能である．(3)の機能は文書の内容に加えてその内容の評価に関する情報も読者に伝える．さらに，(4)の機能は，その文書のみならず，関連する文書の内容も含めて，その話題に関する複数の文書の内容をまとめた情報を読者に提供する．(4)の機能をもつ要約はその話題に関する一種の概説に相当する．

通常，(1)，(2)の要約は著者によって作成されるが，(3)，(4)の要約は第3者によって作成される．また，これらの機能は包含関係にあり，(1)の機能は(2)の機能によって実現できるし，(4)の機能が実現できれば，すべての機能を実現できる．自動要約では(2)の機能をもつことが期待されている．

要約を作成するためには，文書の内容を理解し，中心的な話題を特定し，それを簡潔にまとめるという3つの作業が必要となる．したがって，必然的に言語処理の技術が必要となるが，残念ながら現在の言語処理の技術では，文書の内容を完全に理解したり，高品質の文書を生成することは難しい．現在，コンピュータによる自動要約と称して行われている研究のほとんどは，要約ではなく抄録を作成することを目的としている．要約が内容の理解と文書再生産をともなうのに対して，抄録は，重要な情報を伝えている文を文書から抜き出して並べたものである．

抄録を作成する手法は，主として重要な文をどのようにして同定するかによって分類することができる．具体的には，何らかの情報を利用して各文に得点を付け，もっとも得点の高いものから抜き出すことになる．抄録を作成する方法を文の得点を計算するのに利用する情報に基づいて分類すると，以下のようになる．

(1) 文書中の語の頻度を利用する．
(2) タイトル中の語を利用する．
(3) 文書中，段落中の文の位置を利用する．
(4) 手がかり語を利用する．
(5) 文と文の関係を利用する．

(1)の方法は情報検索の技術をそのまま利用できる．抄録の作成は文を索引語として索引付けを行うことだと考えることができるので，抽出した索引語（文）の集合が抄録に相当する．ただし，頻度に基づいて文の重要性を計算する場合，文の頻度を数えたのではほとんど頻度が1になってしまうため，語の頻度に基づいて語の重要度を計算し，文が含む語の重要度に基づいて文の重要度を計算する2段階の方式をとる．

(2)の手法の背景には，「タイトルに含まれる語は文書中の重要な概念を表している」という仮定がある．文書のタイトルの他にも章や節の見出しに含まれる語を重要語として利用することが考えられる．

(3)の手法では，文書あるいは段落中の文の位置によって文の重要度を判定する．例えば，技術論文などでは段落の最初をその段落の内容をよく表す文で始めるべきであるといわれる．もし，このようにして書かれた文書から抄録を作成する場合，各段落の第１文を抽出すれば抄録が作成できるだろう．

(4)の手法では，あらかじめ決められた手がかり語あるいは表現をリストとして用意しておき，このリストを利用して重要な文を探す．例えば，技術論文などにおいて「本論文の目的は，…」で始まる文は論文の目的を述べていると考えられるし，「…という結果を得た」のような表現で終わる文は結論を述べていると予想できる．目的や結論は抄録にとって重要な情報なのでこれらの表現を含む文は抽出されるべきだろう．

(5)の手法では，文と文の関係を解析し，その結果を利用して文の重要度を決定する．文間の関係としてどのようなものを利用するかによってさまざまな手法が考えられる．

自動要約システムの性能の評価は，人間が作成した抄録との比較によって行うことが多い．すなわち，人間が作成した抄録中の文の集合を正解とし，システムが生成した抄録の再現率と精度を計算することによってシステムの性能を評価する．

5. 文書分類

文書の自動分類とは，文書をあらかじめ決められたカテゴリに分類する，あるいは文書にカテゴリを付与することをいう．例えば，ある新聞記事をその内容にしたがって，「政治」，「経済」，「社会」などのカテゴリに分類することが考えられる．このようなカテゴリを検索質問とみなせば，文書の自動分類は情報検索と基本的に同じだと考えることができる．ただし，情報検索では検索対象となる文書集合は固定であり，入力される検索質問が開集合であるのに対して，文書の自動分類では，検索質問に相当するカテゴリ集合が固定で，入力となる文書が開集合となる．

文書の自動分類は情報検索と基本的に同じなので情報検索の基礎技術を利用することができる．自動分類の基本的な手続きは以下のようになる．

- 各カテゴリをあらかじめ内部表現に変換する．
- 入力文書を内部表現に変換する．
- 文書と各カテゴリの間の類似度を計算する．
- 文書にもっとも類似したカテゴリを付与する．

文書やカテゴリの表現形式や類似度の計算方式は情報検索と同様に用いるモデルによって異なる．例えば，ベクトル空間モデルでは，文書とカテゴリを索引語の重みベクトルで表現し，その間の類似度を両ベクトルの余弦などによって計算する．カテゴリのベクトルは，あらかじめカテゴリが付与された文書集合を訓練データとして用いて計算できる．

何らかのモデルを利用して対象文書に対する各カテゴリのスコアを計算し，スコアの高いカテゴリを文書に付与する．一般にひとつの文書には複数のカテゴリが付与できる可能性がある．例えば，景気対策のために減税を迫る諸外国に対する政府の対応を記述した新聞記事は「政治」のカテゴリに属すると同時に「経済」のカテゴリにも属する可能性がある．カテゴリを文書に付与する方法として以下のような方法が知られている．

- 閾値法（thresholding strategy）
- 定数法（k-per-doc strategy）
- 比例配分法（proportional assignment strategy）

閾値法では一定の閾値を超えるスコアをもつカテゴリをすべて文書に付与する．閾値法は，すべての文書とカテゴリの組についてスコアが比較可能な値におさまることを前提としている．また，閾値を変化させることによって各文書に付与されるカテゴリ数を制御することができる．

定数法ではスコアの高いものから一定数（k個）のカテゴリを文書に付与する．定数法はスコアが各文書に対して比較可能な値におさまることを前提としている．定数kを変化させることによって各文書に付与されるカテゴリ数を制御することができる．

比例配分法では訓練データのカテゴリ分布を反映するようにスコアの高いカテゴリから文書に付与する．比例配分法はスコアが各カテゴリに対して比較可能な値になることを前提としている．各文書に付与されるカテゴリ数は比例定数と呼ばれる値を変化させることによって制御する．

文書の自動分類の評価は情報検索の評価と同様に再現率と精度によって行うことが多い．再現率と精度を計算するためにはあらかじめ正解カテゴリが付与された文書集合が必要である．

6. QAシステム

QA（質問応答）システムとはユーザの自然言語による問い合わせに対して，システムがもつ知識を使って回答するシステムである．したがって，広い意味では情報検索システムなどもQAシステムといえなくはないが，通常は情報検索システムのように漠とした文書を返すのではなく，ユーザの質問に直接的な答えを回答するシステムのことを指す．

QAシステムは言語処理に関するほとんどすべての要素技術を必要とするので，機械翻訳とならび，言語処理技術の代表的な応用システムとして長く研究されてきた．1970年代から1980年代にかけては人手で記述した規則に基づく手法が中心的であり，多くのシステムが試作された．特定の専門分野を対象としたQAシステムは特にエキスパートシステムと呼ばれている．WinogradのSHRDLU[4]は，この時代の代表的なシステムで，ユーザとの対話を通して積木を操作したり，積木の状態に関してユーザの質問に答えたりすることができた．しかし，このようなアプローチは，知識の記述がボトルネックとなり，対象領域を狭く限定しないと，システムがうまく動作しないという限界があった．そのため「おもちゃのシステム」であるという批判を受けることになる．

1990年代に入って情報検索の分野では，大規模なテストコレクションを用いた評価型の会議TRECが米国で始まった．当初は情報検索のタスクのみであったが，1999年のTREC-8から始まったQAタスクでは，従来の情報検索システムのようにユーザの検索質問に対して文書の集合を返すのはなく，直接的な答えを返すことを目的としている．例えば，「日本で一番高い山は？」という問いに対して，「富士山」のように，より直接的な答えを要求される．現在では，QAシステムあるいは質問応答システムといえば，TREC型のシステムを指すことが多い．

TREC型のQAシステムでは，まず，情報検索の手法を用いて，大規模な文書の集合から答えを含んでいそうな箇所を検索する．この場合，必ずしも文書全体を検索対象とするのではなく，文書の一部を検索対象とすることもある．このような検索をパッセージ検索と呼ぶ．検索されたパッセージに対し情報抽出の要素技術である名前の同定技術を使い，文書中に現われるオブジェクトのタイプ（例えば，人名，企業名，日時，時間，場所など）を同定する．一方，ユーザが入力した質問が何を聞いているかを分類することも重要である．現在のQAシステムは，上述した富士山の例のように事実を問うものが中心であり，質問がいわゆる5W1Hのどのタイプかを分類することになる．例えば，富士山の質問は「何であるか」を聞いているのでwhat型の質問に相当する．質問のタイプとパッセージ中に現われるオブジェクトのタイプの間の関係はあらかじめ簡単な規則として記述しておき，パッセージの中でも答えとなりそうなオブジェクトの周辺を切り出して回答する．

TREC型のQAシステムは，それ以前の規則に基づく古典的な質問応答システムに比べ，情報検索や情報抽出の技術を取り入れ，対象領域を限定しない点が特徴である．特に対象範囲をさらに拡大するために，最近では，Webページを知識源とするQAシステムも研究されている．しかし，現在のところこのような手法で回答できる質問は限られており，複雑な回答を必要とするようなhow型やwhy型の質問の扱いに関しては盛んに研究されている状況である．また，古典的なQAシステムと違い，対話の文脈を考慮しない点も大きな違いである．

7. おわりに

本稿では，情報アクセス技術の代表である情報検索技術とその関連技術について紹介した．情報検索に関する網羅的な教科書としては，[1][6]などがある．また，最近ではBlogやインターネット上掲示板などを分析し，ユーザの意見や評判などを抽出する技術の研究も盛んに行われている[5]．

参考文献

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[3] C. D. Paice. Constructing literature abstracts by computer: Techniques and prospects. *Information Processing & Management*, Vol. 26, No. 1, pp. 171-186, 1990.

[4] T. Winograd. *Understanding Natural Language*. Academic Press, 1972.

[5] 大塚裕子, 乾孝司, 奥村学. 意見分析エンジン—計算言語学と社会学の接点—. コロナ社, 2007.

[6] 徳永健伸. 情報検索と言語処理. 東京大学出版会, 1999.