経営論集　第13巻第１号　2003年　105～114頁

情報検索システムにおける言語処理技術の利用

小　松　香　爾

1　はじめに

情報検索は，情報の大きな集合から自分の求める情報だけを探し出すことである。対象となる集合がテキストであることが多いため，テキスト検索のことを情報検索と呼ぶことも多い。情報検索の研究は古くから行われてきたが，従来，検索対象は論文のように均質なものがほとんどであった。しかし，1990年代のインターネットの爆発的な普及は，情報検索にも大きな影響を及ぼした。様々な人がインターネット上のWebサイトで情報を発信するようになり，大量でかつ多様なWebページを対象とする検索方法の必要性が生じてきた。また，新しい検索方法に対し，現実の情報検索システムに匹敵する大規模なテストコレクションを用いて有効性を評価する必要もでてきた。近年では，計算機の高性能化および記録容量の拡大にともない，画像や音声に対する検索方法も盛んに研究されている。

自然言語処理の分野では，1980年代後半以降，コーパス（電子化された言語分析用の言語の資料）をベースにした自然言語処理が中心となった。日本語コーパスに対し，統計的手法を適用した結果，Chasen，Juman，Mecabなどの形態素解析ツールやCaboCha，KNPなどの統語解析ツールが実用レベルにまで発展してきた。本論文では，既存のテキスト検索の技術を概観し，言語処理技術の情報検索への適用について述べる。また，言語処理技術を用いた柔軟なテキスト検索手法を提案する。以降，2節でテキストの索引付け，3節で索引語の重み付けについて述べる。4節で，代表的な検索モデルとして，論理モデル，ベクトル空間モデルを解説し，5節で検索システムの評価，および実際のテストコレクションについて述べる。6節でこれまでの言語処理の利用を概観し，7節で本論文が提案する手法を述べる。

2　インターネット上の文書の索引付け

索引付けとは，テキストから索引語を抽出する処理である。索引システムを作成するためには，テキスト中からそのテキストを特徴付ける索引語を抽出する必要性がある。なお，近年では，画像や動画に対し索引付けを行う研究が盛んである[(1)(2)]。有名な検索エンジンGoogle，Goo，Infoseek，Altavista等で，すでに画像の検索が実用化されているが，本論文ではテキストの索引付けのみを扱う。

2.1 自動索引付け

索引付けとは，テキストから索引語を抽出する処理である。索引語は，テキスト中からそのテキストを特徴付ける性質を持つ必要性がある。インターネット登場以前，例えば，図書館における図書の索引付けは人手で行われてきた。しかし，インターネット上の Web サイトに存在するテキストに対し，人手によって索引付けを行うことは，

① 大量のテキストデータの存在

② 各人の認識の揺れ

という二つの理由から実質的に不可能といえる。①は，ハードディスクの容量の増加やインターネット接続人口の増加等の要因により，インターネットでアクセス可能なテキストは，爆発的に増え続けている。また，今後も減少するような予兆はないということであり，②は多数の人間で索引付けを行うと，どの語を索引語として抽出するかという認識が，個人の興味，経験，教養等によって大きく異なるということを意味する。しかし，インターネット上のテキストの索引付けは，一人の人間の手におえるものではない。よって，計算機による自動索引付けの必要性が生じる。例えば，世界最大の情報量を持つサーチエンジンであるGoogle[3]では，約8000台のマシンを用いて，約14億ページが自動索引付けされている[4]（2001年 3 月時点）。

2.2 テキストを特徴付ける性質

テキストを特徴付ける性質には，大きくわけて二つの性質がある。

① 特定性

② 網羅性

①は，あるテキストには現れるが他のテキストには現れないような性質のことである。特定性を持つ索引語を多数抽出すれば，検索の精度（precision）が上がりやすい。②は，テキスト一般によく現れる性質のことである。網羅性を持つ索引語を多数抽出すれば，再現率（recall）が上がりやすい。精度と再現率は一般的にトレードオフの関係にあり，両方の尺度を同時に上げることは難しい。

2.3 検索単位の粒度

図書の索引付けなど人手による索引付けの場合，索引語は単語のみである場合がほとんどである。しかし，計算機科学においては，音声認識や全文検索の分野で，N-gram が用いられてきた。N-gram は N 文字の順列である。計算機による自動索引付けでは以下のような検索単位が考えられる。

① N-gram

② 単語

③ 複合語

④ それ以上の情報

①から④の中では，④がもっとも検索単位の粒度が荒く，一つの検索単位の中に含まれる情報量が多い。①と②や③では N＝2，3 程度ならば，①のほうが検索単位の粒度が細かく，検索単位中に含まれる情報が少ない。なお，N-gram モデルにおいて，N を大きくしすぎると，再現率が下がりすぎるため，通常は N＝5 までしか使われない。N-gram と単語の同時使用といった複合型を使用する検索システムも存在する。

2.4　単語の分類

自然言語の単語は，内容語と機能語に分類される。分類は言語の種類によって異なる。日本語の場合，内容語は，名詞，動詞，形容詞，形容動詞，連体詞，副詞であり，機能語は助詞，助動詞，接続詞，感動詞である。テキストの索引付けにおいて，通常は機能語は不必要と見なされている。機能語は不要語リストに入れられ，不要語として索引語の候補から削除される。

3　索引語の重み付け

索引語の重み付けとは，抽出した索引語がそのテキストの内容にどれだけ密接に関連しているかを基準として，重要度を索引語に付与することである。すなわち，「一つのテキスト内にいくつかある索引語の中でも，そのテキストの内容を表している度合は異なる」という仮定に基づき，各索引語をそのテキストの特徴を表している度合によって差別化することである。なお，検索語には重みを付与せずに，文書に直接重みを付ける手法もある。実用化されている Web 検索エンジンのほとんどは，この手法を用いている。

3.1　重み付けの意義

重み付けを導入すれば，検索質問に対するテキストの適合度を計算できるようになる。重みがなければ，基本的には同じ索引語を含むテキスト間の順序付けはできない。テキスト間の順序付けがない場合，ユーザ側に不都合が多い。なぜなら，大量のテキストがヒットした場合，どのテキストから優先的に目を通すべきなのかが示されないからである。検索するユーザの立場に立てば，テキスト間になんらかの順序付けが必要であり，実用検索エンジンでは，リンク元となっている Web テキストの数などで，テキスト全体に重要度を付与している。その際，単にリンク元となっているページ数だけで重要度を決めると，ダミーサイトからのリンクを故意に設定しているページが上位に来ることがある。Google ではリンク元となっているページの重要度を考慮して重要度を求める PageRank 法を用いて，ダミーサイトの影響を低く抑えている[(5)]。PageRnak 法の使用により，Google は「Scam Web ページ（汚い手口の Web ページ）に強い」という定評を得ている。なお，PageRank 法は，Google 社の登録商標，かつ，特許出願中の技術である。

3.2　重み付けの手法

索引語の重み付けの基準は以下の三種類が主に用いられる。

① 検索語頻度（term frequency : tf）

② ドキュメント頻度（document frequency : df）

③ tf･idf

①の検索語頻度は「テキスト中で何度も繰り返し言及される概念は，そのテキストの主題となる概念である」という仮定に基づいており，t をある索引語，d をテキストとすると，以下の式で定義される。

$tf(t,d)$

$tf(t,d)$ はあるテキスト d 中に出現する索引語 t の頻度を意味する。ただし，あまりに頻度の高い語は，テキストを特徴づける上では役にたたない。例えば，コンピュータ関連のテキスト集合を検索する際に，「コンピュータ」という索引語は重要ではない。索引語頻度の単独使用は，このような一般的な索引語の重みが重くなりすぎてしまうという欠点があげられる。②のドキュメント頻度は，t を検索語として以下の式で定義される。

$df(t)$

$df(t)$ は索引語 t が出現するテキスト数を意味する。ただし，テキスト検索においては，「多くのテキストに出現する索引語は，特定のテキストを特徴付けているとは言えない」という仮定に基づき，逆ドキュメント頻度 $idf(t)$ として以下のように定義され，$idf(t)$ が重要度として付与される。ただし，N をテキスト集合中のテキストの数，t を索引語とする。

$$idf(t) = \log(N/df(t))$$

$idf(t)$ は索引語 t が特定の文書に偏って出現する程度を意味する。逆ドキュメント頻度の問題として，テキスト集合が決定しなければ，求めることができないことがあげられる。その日その時で全テキストの集合が変化する Web 上での検索では，厳密な idf を求めることはできない。

実際のテキスト検索システムにおいては，索引語頻度と逆ドキュメント頻度の両方を考慮し，③の tf･idf 法を索引語の重み付けとして使用することが多い。tf･idf は，t を索引語，d をテキストとして，

$$tf\cdot idf(t,d) = tf(t,d)\cdot idf(t)$$

と定義される。

4　検索モデル

代表的な検索モデルは論理モデルとベクトル空間モデルである。他に，ファジィ集合モデル，拡張ブーリアンモデル，確率モデル，ネットワークモデル，クラスタモデルがあげられる[6]。しかし，いずれも前記二つのモデルの派生モデルであったり，高速化，高性能化が難しく発展性に乏しいモデルである。よって，本節では前記二つのモデルについて述べる。

4.1 論理モデル

現在運用レベルにある，ほとんどの情報検索がこのモデルを使用している。転置ファイルを使用するモデルであり，転置ファイルは転置リスト，文書ファイル，および辞書ファイルの三つの基本ファイルから構成されている。転置リストには，各索引語に対して索引語の出現した文書のリストが蓄積される。ある索引語に対して転置リストを探索するには辞書ファイルを通じて行う。辞書ファイルはシステム中の全ての索引語とその転置リストの場所へのポインターをソートしたリストである。検索が行われる際には，検索質問中の用語に対して転置リストが探索され，転置リスト間で検索質問にふさわしい論理操作が適用される。例えば，検索質問(文京 AND 地図）は「文京」と「地図」に対する転置リストを探索するために，辞書ファイルを用いる。得られた二つのリストは論理的に AND が適用され，最終的に得られるテキスト集合が決定される。論理モデルの特徴として，検索質問の表現形式が文書の表現形式と一致しないことがあげられる。論理モデルの構造は単純であり，各検索エンジンの優劣は主に，

① 転置ファイルを圧縮する

② 辞書ファイルをツリー構造にする

など高速化技術の優劣や，得られたテキストに順位付けをする方法，検索の対象となるページを集めるためのマシンパワーの優劣によって決まる。

4.2 ベクトル空間モデル

論理モデルと比較すると，索引語に重みを与えられることが大きな特徴である。検索質問文から抽出した質問タームとテキスト中の索引語のそれぞれに対し重要度を求め，重要度を成分とするベクトル（通常，項ベクトルと呼ばれる）を作成する。ここで，検索質問から作られた項ベクトルを q とし，テキスト j より作られた項ベクトルを Dj とすると，q と Dj が似ていれば似ているほど，テキスト j は検索質問に適合するテキストであるといえる。二つのベクトルの類似度を計算する手法はいくつかある。よく使われているのが，以下の式で表される内積をとる方法である。

（q･Dj）

ベクトル空間モデルは，適合度を文書ベクトルと検索質問ベクトル間の類似度に帰着させる手法である。ベクトル空間モデルの特徴として，適合性フィードバックが行い易い点があげられる。適合性フィードバックとは，ユーザが検索結果に対し適合性判断を行い，適合性判断の結果により検索語の重みを変化させることである。これにより検索の精度が劇的に向上する。また最近は適合性フィードバックによる学習を効率化した，サポートベクトルマシン（SVM）の利用が盛んに行われている。

5　情報検索システムの評価

検索システムを評価するためには，テストコレクションと呼ばれる，ある種のベンチマークセットが不可欠である。米国では，1992年から毎年，TREC（Text Retrieval Conference）という評価型国際会議が開かれてきた。TREC の中心となるタスクは，通常の文書検索である。TREC では，参加者には事前にテキスト集合が与えられ，評価時には検索トピック（検索要求）が与えられる。参加システムは検索トピックと関連性の高いテキストをテキスト集合から検索する。その後，精度（precision）と再現率（recall）を用いて，参加システムの検索精度を比較評価する。日本語のテストコレクションとしては，新情報処理開発機構が構築した BMIR，学術情報センターが構築した NTCIR がある。本節では，情報検索システムの評価法と，テストコレクションで要求される索引能力について述べる。

5.1　評価基準

検索結果は，「検索質問に適合する文書を漏れなく検索しているか？」という完全性と，「検索質問に適合する文書だけを検索しているか？」という正確性を同時に満たす検索システムが望ましい。そのための評価基準として，以下のように定義される精度と再現率が用いられる。

精度＝検索された文書中の適合文書の数／全文書中の適合文書の数

再現率＝検索された文書中の適合文書の数／検索された文書の数

精度と再現率を組み合わせた評価法として，横軸を精度，縦軸を適合率とした曲線である精度・適合率曲線や，以下のように定義される F 尺度が用いられる。

$$F=2/(1/P+1/R)$$

5.2　テストコレクション

一般的なテストコレクションは，

① 参加者に予め渡される検索対象となるテキスト集合

② 検索システムの評価時に用いられる検索質問文の集合

③ 各検索質問文に対し，どのテキストが適合するかを定める適合情報

の三つのデータから構成される。

例として，日本語のテストコレクション BMIR-J2[(7)] をあげる。テキストは毎日新聞94年度から，各分野まんべんなく抜き出したもので，テキストの数は5080，平均長は617.6語である。一方，質問数は60で，平均長は124語である。検索質問はシステムに要求する機能によって，以下の五つに分類できる。

① 基本機能

② 数値レンジ機能

③ 構文解析機能

④ 内容解析機能

⑤　知識処理機能

①は索引語，あるいは索引語のシソーラスによる展開語の一致，およびそれらの論理式による機能を要求するもので，最も基本的なものである。例えば「会社が不正をしたことに関する文書」のように「会社」「不正」という二つのキーワードを質問タームとすれば，適合するテキストが求めるような質問である。②は数の数え上げや，数値などの範囲を正しく解釈する機能を要求するもので，例えば「社員10人以下の企業に関する文書」などである。③は複数の索引語の間の係り受け関係を判断する機能を要求するもので,「日本が米国へ自動車を輸出することを記述した文書」のように，統語解析をしなければ正解が得られないような質問である。④は深い言語知識を利用する機能を要求するもので，文脈を理解することや，言葉の深い意味を理解することを含む。例えば「自動車市場の動向に関する文書」のように「動向」に関する意味解析を行わなければ正解が得られないような質問である。⑤は人間の常識を利用する機能を要求するもので，蓄積された事実からの推論などを含む。例えば「異業種の企業による合弁事業に関する文書」のように，「異業種」に関する一般常識がなければ正解が得られないような質問である。このような質問群に対して，現実のシステムの再現率約40％における精度は，適合性フィードバックなしの最良システムで約20％，適合性フィードバックありの最良システムで約60％であった[6]。ユーザによる適合性フィードバックがいかに有効であるかということが示されている。しかし，それがなければ，再現率40％という低い網羅性の検索結果（比較的，厳選されている）においても精度が約20％しかないのも事実である。

6　自然言語処理技術の利用

人間の言葉である自然言語は，自然発生し，変化を経て現在の形に至ったものである。人間が意図的に設計したものではないため，一種の自然現象とみなすことができる[8]。曖昧性や例外的な現象が妨げとなり，規則を完全に記述することは不可能である。したがって，言語の研究は，まず事実の収集・観察から始めなければならない。実際に使用例を集め，言葉の使用状況を具体的に分析し，個々の語の振る舞いを正確に記述し，背後の規則をモデル化することによって初めて，言葉を技術として取り扱うことができる。自然言語処理の技術は，形態素解析，統語解析，意味解析，談話解析に分類できる。近年のコーパスの充実とマシンパワーの増大により，実用に耐える頑健性を持つシステムも現れてきた[9]。本節では，情報検索における自然言語処理技術の利用について述べる。

6.1　形態素解析技術の利用

索引付けの前処理として形態素解析が利用される場合がある。それにより，自立語とそれ以外の語に分類することができ，自立語以外を不要語とすれば，わざわざ不要語リストを作る必要がないという利点がある。漢字のユニグラムに基づく索引付け（NLQ モデル）をするシステムに対し，やや性能がいいことが分かっている[6]。

例：「日本の自動車メーカは輸出規制を決めた」

漢字ユニグラム：「日」「本」「自」「動」「車」「メーカ」
「輸」「出」「規」「制」「決」

語：「日本」「自動車」「メーカ」「輸出」「規制」「決める」

6.2 統語解析技術の利用

完全な統語解析（いわゆる構文木の作成）は難しく，現時点では構文木を単純化した格フレームの利用が行われている。表層格に基づく格フレームを抽出し，格フレームの要素に重みを付けるという手法である。各フレームの重みを，単一語の索引語を用いたベクトル空間モデルに適用すると，BMIR-J2において数パーセントの精度の改善が見られた。(6)

6.3 意味解析技術の利用

文の完全な意味解析は，統語解析同様に難しく，現時点では，シソーラスや辞書の利用が現実的である。特に，シソーラスを用いて意味的に近い語を質問タームに加えるという質問拡張という手法はよく用いられてきた。索引語として表層的な語ではなく辞書の語義を使う手法は，ほとんど改善が認められないとされている。(6)

例：「保守」に対する辞書の二つの語義

保守１　＝　保守的なこと

保守２　＝　メンテナンス

6.4 談話解析技術の利用

照応関係の同定をして，代名詞を具体的な名詞に置き換える処理をして索引語の重みに加えるという手法が提案されている。しかし，現時点では照応関係の同定自体の精度が悪いため，結果が改善されないことが分かっている。(6)

7 検索質問文の柔軟な拡張

言語処理自然言語処理の情報検索への適用は，いずれも成功しているとはいえない。(10) 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つかえるようにツール化され，同時に様々な言語資源も整備されてきた。よって，言語処理技術を本格的に利用する土壌は整った(10)という主張は正しい。しかし，本論文では，シソーラスや構文木といったオーソドックスな言語資源を利用する前に，自然言語，特に日本語の持つ自由度への対策を十分に行うべきであることを主張する。以下，7.1節では，シソーラスでは現れないような類似語を利用する手法を提案する。「≒」で結ばれた表現は，表層的には異なるが意味的に非常に近く，質問タームの拡張に使用すれば有効であると考える。7.2節では従来は不要語として検索時に考慮されなかった語の情報を用いる手法を提案する。本手法を用いれば，否定文中の質問タームが精度に悪影響を及ぼ

すことを防げる。

7.1 単語レベルの自由度の吸収

以下の「≒」両辺にある単語のどちらかが現れた場合，もう一方の単語も質問タームに含めれば，精度の向上が期待できる。

例：

お母さん≒おかあさん	(かなと漢字)
コンピュータ≒コンピューター	(表記のゆれ)
登山≒山登り	(同義語)
金≒お金	(丁寧)
黒≒真っ黒	(強調)
日本車≒日本の車	(助詞)

上記の「かなと漢字」の類の類似語は既存の電子辞書から該当箇所を抽出すれば，専用の辞書が容易に作成できる。「表記のゆれ」はゆれが生じる語は少数であるため，人手で辞書が作成できる。「同義語」は既存の同義語辞典の類がそのまま使用できる。他の種類の類義語に関しては，現在までのところ使用できる言語資源がないが，将来的に辞書が整備される可能性はある。

7.2 不用語による文書の索引語の重み調整

質問文章中に存在する否定表現が，検索の精度に悪影響を及ぼすことがある[(11)]。本論文では，通常不要語と見なされる機能語を積極的に利用することにより，仮定，否定表現と共起する索引語の重要度を減らす手法を提案する。以下の例で，否定表現「ない」と共起する単語「雪印乳業」や「倒産」の重要度を減らせば，「企業が倒産したこと」に関するテキストが検索結果の上位に現れることを防げる。否定の強さの順は，「確定済みの否定」，「強い否定」，「否定」，「憶測を含む否定」で，その順に重要度を低くしなければならない。また，「二重否定」に関しては肯定になるので，無条件に重要度を減らしてはならない。二重否定を考慮した否定のパターンテンプレートを作成する必要が生じる。仮定の例も同様である。「逆説的な仮定」の場合，「雪印食品は倒産しない」など否定文が続く可能性が高い。よって，「雪印乳業」や「倒産」の重要度を減らしたほうがよい。「純粋な仮定」は，肯定文，否定文のどちらが続くかの確率が恐らくイーブンであるため（厳密にはコーパスに対し，統計的処理を行い確率を調べる必要がある），重要度は減らしすぎないほうがよいと考える。「順接的な仮定」は，肯定文が続く可能性が高いので，重要度は変えないほうがよい。

否定の例：

「雪印乳業は倒産しなかった」	(確定済みの否定)
「雪印乳業は絶対に倒産することはない」	(強い否定)
「雪印乳業は倒産することはない」	(否定)

「雪印乳業が倒産することはないであろう」　（憶測を含む否定）
「雪印乳業が倒産するということはないとはいえない」（二重否定）

仮定の例：

「雪印乳業が倒産するとしても」　（逆説的な仮定）
「仮に雪印乳業が倒産したとしても」　（逆説的な仮定）
「仮に雪印乳業が倒産した場合」　（純粋な仮定）
「雪印乳業は倒産することはないだろうが」　（順接的な仮定）

8　まとめ

情報検索を概観し，これまでの自然言語処理の適用および将来的な可能性について考察した。本論文が7節で提案した手法は，現段階では仮説に基づく提案にすぎず，テストコレクションを用いて実験されない限りその有効性は証明されない。よって，今後の課題は，実際にシステムを構築することである。

（注）

（1） Flickner, M., Sawhney, H. and Niblack, W. el al. : Query by Image and Video Content : The QBIC System, IEEE Computer, Vol.28, No.9, pp.23－32（1995）.
（2） Wactlar, H., Kanade, T. and Smith, M. : Intelligent Access to Digital Video : The Informedia Project, IEEE Computer, Vol.29, No.5（1996）.
（3） Brin, S. and Page, L. : The Anatomy of Large-scale Hypertextual Web Serch Engin, Proc. Of the 7th Int. World Wide Web Conf., pp.107－117（1998）.
（4） Brin,S. and Page, L. : Dynamic Data Mining : Exploring Large Rule Space by Sampling, http://www-db.stanford.edu/~sergey/ddm.ps.
（5） Haveliwala, T.H. : Efficient Committee of PageRank, Stanford Digital Library Technologies, Technical Rep. 1999-31（1999）.
（6） 徳永健伸：情報検索と言語処理，東京大学出版会（1999）。
（7） 木谷強，小川泰嗣，石川徹也他：日本語情報検索システム評価テストコレクション BMIR-J2，情報処理学会研究報告，データベースシステム，98-DBS-114-3（1998）。
（8） ここまできた自然言語処理，情報処理，Vol.41，No.7（2000）。
（9） 使いやすくなった自然言語処理のフリーソフト，情報処理，Vol.41，No.12（2000）。
（10） 徳永健伸：言語処理は情報検索に役立つか？，日本音響学会，2000年春季研究発表会講演論文集，pp.31－32（2000）。
（11） 白井清昭，Rila Mandala，徳永健伸他：TREC-7参加報告，電子情報通信学会，言語理解とコミュニケーション研究会，Vol.98，No.660，pp.31－38（1999）。