自然言語入力と目次との柔軟な照合による図書検索システム

白木 伸征　黒橋 禎夫

京都大学大学院工学研究科 電子通信工学専攻
{nshiraki, kuro}@kuee.kyoto-u.ac.jp

1 はじめに

現代の高度情報化社会において，情報検索は重要な位置を占めるようになってきている．その一分野である図書検索は，電子図書館などの普及により高精度なシステムが必要となっている．

図書検索は，図書のどのような情報を用いるかによって次のように分けることができる：1. 書名，2. 索引やキーワード，3. 目次，4. 本文テキスト．

1の方法はOPACなどで用いられているが，書名だけでは明らかに情報不足といえる．2の方法は，図書に索引やキーワードがあまりついていないため，実用的ではない．4の方法は，図書の本文がほとんど電子化されていないため，実用的ではない．これに対して3の方法は，目次が比較的よく図書の内容を表現していること，データ量的にこれまでの冊子体の図書の目次を電子化することも可能であること，などから最も有効な検索手法になると考えられる．

我々はこれまでに目次の構造を用いる図書検索の手法を提案している [1]．本研究は，この図書検索システムを基本とし，情報検索で用いられてきたいくつかの手法を組み合わせ，さらに名詞句の柔軟な照合を行うことによる高精度な図書検索システムを提案する．

2 本研究の方法

2.1 目次の階層構造の利用

目次のテキストには，章や節の構造による階層構造が存在する．そのため，目次中に2つの単語 (以下AとBとする) が現れる場合，それらの語の間の関係の強さは，階層構造中でのそれらの2単語の位置関係により以下のように順序づけできると考えられる．(タイトルは書名や章名や節名を表す)

- A = B : AとBが同じタイトル中に存在
- A > B : AがBより上位のタイトル中に存在
- A < B : AがBより下位のタイトル中に存在
- A ∥ B : AとBが同じ階層のタイトル中に存在
- A # B : 上の4つ以外の関係でタイトル中に存在

このためAとBという入力の場合には，この順で検索意図に対して適当であろうと推測される [1]．

2.2 ベクトル空間モデルと wtf・idf

情報検索で良く用いられる手法に，ベクトル空間モデルがある．ベクトル空間モデルは，ドキュメントと質問を多次元ベクトルで表現し，それらのベクトルの内積を質問とドキュメントとの類似度とし，それにより検索結果を得る方法である．

そのベクトルの要素として，単語の重要度を表す指標である tf・idf が多く用いられる [3]．tfはドキュメント中の単語の出現頻度で，多く現れる単語が重要であることを表す．またidfは単語が出現する文書の数の逆数で，特定の文書に現れる単語が重要であることを表す．

目次はその階層構造により，上位のタイトルの方が下位のタイトルよりも重要な意味を持つと考えることができ，そこに含まれる単語も同様のことがいえる．そこで，tfのように単に単語の現れた回数を数えるのではなく，単語の現れる目次階層の深さによりその単語の重要度に差をつけ，次の式のように計算する．

$$wtf = \sum_i \frac{1}{(depth_i \times w_{depth} + 1)}$$

ここで，iは単語の出現位置を表し，$depth_i$は出現位置iにおける目次階層の深さ(書名を0，章名を1とする)とし，w_{depth}は重みの値である．この重みは予備実験により$w_{depth}=1.0$とした．

このような重みを考慮した単語の出現頻度を**wtf** (weighted tf)と呼び，これに通常のidfをかけあわせたものを，**wtf・idf**と呼ぶことにする．

2.3 自然文による入力とその解析

本研究の図書検索システムは自然文による入力を用いる．自然文による入力の利点は以下の2つである．

- 検索者の検索意図をそのままの形で入力できる．
- 単なる単語の入力では検索意図を十分に表現できないが，自然文による入力を用いることにより，検索意図を表す名詞句を取り出し活用することができる．

入力された自然文(以下**検索式**と呼ぶ)を活用するために形態素解析を行い，検索式を単語に区切りその品詞を得る．そこから名詞，助詞「の」，助詞「と」からなる単語列とそれ以外とを分け，名詞，助詞「の」，助詞「と」のみからなる句を対象に以下の処理を行う．

助詞「と」はその前後の名詞句に関して並列であることを示すが，その区切りは曖昧である．そこで，句の中に助詞「と」が含まれる場合，その句を「$\cdots$のN_{-2}のN_{-1}とN_1のN_2の$\cdots$」(N_iは名詞)として以下のような処理を行う．

- N_{-1}とN_i $(i \geq 1)$との類似度[1]を比較し，類似度の最も高いものを$N_a(a \geq 1)$とする．
- N_1とN_i $(i \leq -1)$との類似度を比較し，類似度の最も高いものを$N_b(b \leq -1)$とする．

これにより，

「$\cdots \underline{N_{b-1}}$の$\underline{N_b \cdots N_{-1}}$と$\underline{N_1 \cdots N_a}$の$\underline{N_{a+1} \cdots}$」

のようになり，「$\underline{N_b \cdots N_{-1}}$」と「$\underline{N_1 \cdots N_a}$」が並列であると見なされる．その結果この句は

「$\cdots \underline{N_{b-1}}$の$\underline{N_b \cdots N_{-1}}$の$\underline{N_{a+1} \cdots}$」

「$\cdots \underline{N_{b-1}}$の$\underline{N_1 \cdots N_a}$の$\underline{N_{a+1}}$の$\cdots$」

の2つの意味を持つと考えることができる．

例えば，「現代の日本の経済とアメリカの政治の腐敗」という句は，「経済」に最も類似した語は「政治」，「アメリカ」に最も類似した語は「日本」となるので，「現代の日本の経済の腐敗」と「現代のアメリカの政治の腐敗」と解釈することができる．

このようにして得られた名詞と助詞「の」からなる句を**ノ名詞句**と呼ぶ．検索式中に現れるノ名詞句は検索意図を強く表すものと考えられる．たとえば，検索意図が「日本の公害」の場合，「日本の公害の与える影響」という句が存在する目次と，「中国の公害と日本の技術提供」という句が存在する目次では，前者の方が検索意図に一致しているといえる．

そこで，検索式中に現れるノ名詞句が目次中にも現れる場合，その図書が検索意図に一致している可能性が高いと判断する．

2.4 検索式の拡張

検索式の拡張についてはこれまでにも様々な研究が行われているが，単に類義語を用いて拡張すると，検索精度が悪くなることが報告されている[2]．そこで本研究では，検索語のひらがな形とカタカナ形に注目し，それにより検索式の拡張を行う．

例えば，「癌」という語は「がん」や「ガン」と表現されることが多い．これらは検索語と全く同じ意味を持つため，拡張を行っても検索精度を悪化させることはない．逆に，ひらがなやカタカナで入力された検索語を漢字に変換するものよいと考えられるが，それは非常に難しいので本研究では考慮しない．

また，複合名詞はその中に現れる名詞の意味も含むため，検索語が複合名詞中にあればその意味を含むと考えることができる．

以上より，検索語T_tのひらがな形とカタカナ形を$T_{t0}, T_{t1}, \cdots, T_{tn}$ (ただし$T_{t0}=T_t$とする)とし，また，XをT_tを除いた0個以上連続する名詞とすると，検索語T_tは，$XT_{ti}X$ $(i=0,1,\cdots,n)$と拡張される．

[1]名詞の類似度の計算には，NTTの日本語語彙大系[4]の単語の意味属性により体系化された木構造を用いる．単語xとyの類似度sim_{xy}は以下の式で求められる．

$$sim_{xy} = \frac{2L}{l_x + l_y}$$

ここで，l_x, l_yはxとyの意味属性の木構造の根(Root)からの階層の深さを表し，Lはxとyの意味属性で一致している階層の深さを表す．

例えば，「癌の告知」というノ名詞句を拡張すると，「肺ガンの告知問題」にも一致することが可能となる．

3 本研究の図書検索システム

本研究の図書検索システムは，2章で述べた方法を組み合わせて，次のように2段階に検索を行う．

3.1 第1検索図書の選択

入力された検索式を形態素解析し，そこから名詞を検索語として抜き出す．ベクトル空間モデルの質問ベクトルのベクトル要素として検索語とそのひらがな形とカタカナ形の現れる回数を用い，ドキュメントベクトルのベクトル要素として図書の目次中に表れる名詞のwtf・idfを用いる．ベクトルの内積の値をその図書の基本点とし，基本点が0より大きく，基本点の上位100までの図書を**第1検索図書**と呼ぶ．このように図書を制限するのは，この段階の選択である程度適当な順位づけが行われていると考えられるためと，また次の3.2節の処理には時間がかかるためである．

3.2 最終検索図書の選択

さらに第1検索図書のそれぞれに対し，検索式中の全てのノ名詞句に関して以下のように目次の内容を調べ，基本点へのスコアの加算を行う．

1. 拡張されたノ名詞句の存在

   2.3節で述べたように，検索式中のノ名詞句が目次中に存在する場合，その図書は検索意図に一致する可能性が非常に高いといえる．そのためノ名詞句が目次中に存在する場合，スコアを大きく加算する．ノ名詞句が「AのBのC」の場合，「AのBのC」「AのB」「BのC」の3種類全てを調べる．

2. 2単語の目次階層の5種類の関係が存在

   ノ名詞句の中の2つの名詞に対して，2.1節で示した目次の階層関係を調べる．加算するスコアの大きさは，目次の階層関係の強さの順とする．ノ名詞句が「AのBのC」の場合，AとB，BとCの2種類の目次階層の関係を調べる．

表 1: スコア加算の重み

	重み
ノ名詞句	20.0
A = B	5.0
A > B	3.0
A < B	3.0
A ‖ B	2.0
A # B	1.0

表 2: 実験結果

検索式総数	202
最終検索図書総数	16443
(1検索あたり	81.4)
評価3の最終検索図書総数	755
(1検索あたり	3.7)
評価2の最終検索図書総数	1008
(1検索あたり	5.0)
評価1の最終検索図書総数	14680
(1検索あたり	72.7)

ここで，ノ名詞句が存在する場合と5種類の階層関係が存在する場合のスコアを加算するための重みは，予備実験により表1の通りとなった．

第1検索図書のすべてに対してスコアの加算を行った結果をその図書の最終的なスコアとし，スコアの降順に並べた図書を**最終検索図書**と呼ぶ．

4 実験

本研究の実験は，岩波新書1211冊の目次を対象に，202の検索式を用いて検索実験を行った．

4.1 正解データの作成

初めに，検索式とそれに対応する評価済みの図書のデータを作成する．3節の方法により202の検索式を入力して検索をし，その結果得られた最終検索図書を全て三段階 (3: 満足，2: 普通，1: 不満) で評価する．その結果を表3に示す．

4.2 評価基準

ここで，評価基準として以下のような計算方法を用いる．

1. 最終検索図書数を n とする．

表 3: 実験結果

	かな拡張	複名拡張	ノ名詞句	目次階層	wtf idf	tf idf	RP
1	○	○	○	○	○	○	59.9%
2	×	○	○	○	○	○	57.0%
3	×	×	×	○	×	×	48.2%
4	×	×	×	×	○	×	52.1%
5	×	×	×	×	×	○	49.4%

2. 順位1からnまでの図書を対象に，RecallとPrecisionを計算する．Recallを横軸に，Precisionを縦軸にしてグラフにプロットする．RecallとPrecisionは以下のように定義する．

$$\text{Recall} = \frac{\text{順位1から}n\text{の中の正解図書数}}{\text{全正解図書数}}$$

$$\text{Precision} = \frac{\text{順位1から}n\text{の中の正解図書数}}{n}$$

全正解図書数は，4.1節の結果から得られる正解図書の数である．

3. nを一つ減らして2.のステップを$n=1$になるまで繰り返す．

4. グラフにプロットした点を線で囲み，その面積を求め，Recall 100%, Precision 100%の正方形の領域に占める面積比を求める．

5. 202の検索全てに対して面積比を計算し，その平均を求める．

正解図書を評価3のみとした時，正解図書を評価3と2とした時の平均面積比をそれぞれRP3，RP2とし，その平均をRPとしてそれを評価基準とする．

4.3 実験結果と考察

4.1節で得られた正解図書のデータをRPの計算の基準として，表3のような実験を行った．表中「かな拡張」はひらがな形とカタカナ形による検索式の拡張を表し，「複名拡張」は複合名詞による検索式の拡張を表す．

この結果から，本研究で考えた各種の手法を組み合わせる方法が最も検索精度が高いことがわかる．また，tf・idfに対するwtf・idfの有効性や，かな拡張や複合名詞拡張の有効性なども確かめられた．

表 4: 「日本の福祉」の検索結果

点数	評価	書名	主な関係
31102.1	3	体験ルポ	ノ名詞句
18579.1	3	高齢者医療と福祉	ノ名詞句
2631.4	3	体験ルポ	A=B
518.7	2	現代日本の民主主義	A>B
480.4	1	会社本位主義は崩れるか	A>B
429.4	3	社会保障	A‖B
369.8	2	現代日本社会と民主主義	A>B
301.8	3	議会	A#B
286.9	3	障害者は，いま	A‖B
255.1	3	GHQ	A‖B

表3の1の方法による検索で，「日本の福祉」という検索式の検索結果でスコアが上位のものを表4に示す．

5 おわりに

本研究では，図書検索の高精度化のために，これまでに情報検索で用いられてきた手法を組み合わせ，さらにノ名詞句を用いる検索の方法を新たに提案した．この方法は簡単であるが有効であるため，今後様々な応用ができると期待される．

参考文献

[1] 黒橋 禎夫, 萩原 典尚, 長尾 眞: 目次情報を利用した図書検索システム，情報処理学会 情報学基礎研究会, 45-5, 1997.

[2] 栗山 和子: シソーラスを用いた検索式拡張の評価，情報処理学会 情報学基礎研究会, 52-1, 1998.

[3] Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company, 1983.

[4] 池原，宮崎，白井，横尾，中岩，小倉，大山，林編: 日本語語彙大系，岩波書店, 1997.