自 然 言 語 処 理 140−21
（2000. 11. 22）

自然言語を用いた対話形式による文書検索における事典情報の利用

高野 敦子* 平井 誠** 北橋 忠宏***
兵庫大学　経済情報学部*
大阪市大　工学部**
大阪大学　産業科学研究所***

計算機による情報検索が一般に広がり，精通していない領域での検索をユーザ自身で行うといったケースも増えてきている．それに伴い，検索を支援するシステムは，ユーザが求める情報を具体化するという検索作業の初期の段階での支援が求めらてきている．我々はそのような状況を踏まえて，ユーザの要求情報の具体化作業を支援するインタフェースとして，自然言語を用いた対話形式の枠組みを提案している．本稿では，検索領域の事典情報を用いることによって，ユーザの入力から得られる索引語の拡張を図り，ユーザとの相互作用によって要求を段階的に具体化する手法を提案する．自然言語によるユーザの入力はユーザが漠然と持っている情報要求を説明していることが多い．そこで，各領域での重要語句である見出し語とそれに対する説明記述からなっている事典情報において，入力中の索引語が説明記述の中に頻繁にでてくる見出し語がユーザの情報要求を具体化する手がかりとなり得ると考えられる．さらに，この見出し語をユーザに提示し，その適合性をユーザに問うことによって，ユーザ自身が要求を具体化することの支援を目指す．

情報検索　対話　事典情報　索引語　索引語拡張　ブーリアンモデル

A Framework of the information retrieval through man-machine dialogue

Atsuko Takano Hirai Makoto Tadahiro Kitahashi
Faculty of Economics Information Science, Hyogo University,
Faculty of Engineering, Osaka City University,
The Institute of Scientific　and Industrial Research Osaka University

In this paper, we propose a framework for making use of cyclopedias for system-user dialogue for newspaper accounts retrieval. We have been pursuing a dialogue control scheme which aims to satisfy incompletely defined users' needs through a dialogue between users and system. In this scheme, the system does not require users to formulate a queries. The system constructs an image of the view of users' interests based on their inputs and responds with the information according to the image through system-user interaction. In order to realize the control scheme, we introduce a "knowledge base" approach based on users' interests and a method which reconstructs the knowledge base by grouping analogous concepts in view of users' interests.

information retrieval, cyclopedia, dialogue Boolean retrieval, index term

1 はじめに

一般に情報検索作業において，要求の具体化が最も困難な作業の１つであることは既に報告されている [Oddy 77]．計算機による情報検索が一般に広がり，目的が明確化していない段階で検索者自身が計算機に向かって検索を行うといったケースが増えてきている近年，要求の具体化におけるシステムの支援が求められるようになってきた．つまり，検索を支援するシステムに，検索者が求める情報を具体化するという従来より初期の段階での支援機能が求めらてきている．我々はそのような状況を踏まえて，自然言語を用いた対話形式による情報検索支援システムを提案してきた．[高野 96] 自然言語を使った自由な入力を許すことは，要求が具体化できていない検索者の負担を大幅に削減できると考えられる．

Oddy の THOMAS システム [Oddy 77] は，１度の受け答え毎に検索処理を完結させるのでなく，対話という形式をとることにより，段階的に検索条件の具体化を図る．本研究も，システムは検索者の要求を修正しながら蓄積し，検索者はシステムが提供する情報によって要求情報の具体化を進めるという相互作用を行いながら検索者の要求を満足することを目的として対話形式のインタフェースを採用した．

THOMAS システムでは対話形式をとっているものの，検索者の入力に自然言語を用いることは，現在の言語処理技術では困難なことから，検索者の入力として索引語の集合を採用している．しかし，自然言語による入力の意味解析は困難であるが，文書の索引付けと同じ手法を用いて索引語の集合に変換することは可能である．検索者の入力は情報要求が具体化されていない段階で行われており，検索者の要求が的確に表現されていることが期待できないことを考慮すると，意味処理はあまり必要ないと考える．むしろ，我々は自然言語による入力を採用することによって，検索者が索引語を生成する作業の負担を削減するだけでなく，その作業によって失われる検索者の漠然とした要求を具体化する手がかり語の利用を図る．

しかし，一般に検索者が入力するテキストの量は多くないので，そこから抽出できる索引語の数は少なくなる．したがって，抽出した検索語を元に，検索者の要求をより的確に表現する概念を検索者と対話しながら導く必要がある．そのためには，検索者の示した索引語の拡張とその表現の最適化が必要である．

索引語を拡張する手法としては，シソーラスや概念体系の知識を用いる方法が知られている．特に元になる索引語が検索者の要求を的確に表現している場合には，各分野専用シソーラスを使った索引の拡張は，その効果が期待できる．しかし，本研究が対象とする検索者の要求が具体化されていない場合にはシソーラスなどの利用はあまり効果が期待できない．

我々は，検索者の求める情報が漠然としているため，その入力は漠然とした要求情報をなんらかの方向から説明しようとする内容となることが多いことに着目した．これは，各領域事典の見出し語とその説明という基本的な構造と類似している．さらに，事典情報の電子化が急速に進んでいる現状を鑑み，事典情報を使った索引語の拡張を提案する．本稿では，検索者の入力から得られる索引語から，関連する見出し語を抽出することによって，ユーザの得たい情報の具体化を支援し，より有効な検索式を構成する方法について述べる．

本研究では 99 年度の日経新聞から環境問題に関する記事を検索する作業を例題として取り上げ，その支援システムについての実験を行った．その際に利用する事典情報は，日外アソシエーツが提供する環境問題情報事典である．この事典は 871 の見出し語から成る．実験の際に用いた検索者の入力例は環境に関する冊子に寄せられた読者の質問などをアレンジしたものである．

2 事典情報の利用

2.1 検索者の情報要求表現の拡張

検索者は要求する情報が具体化されていないため，それを直接表現することができず，漠然と持っている情報要求を説明するような内容を入力することが多い．したがって，検索者の要求は比較的一般的な表現の組み合せによって表されていることが多く，索引付けを行い，索引語の集合に変換してしまうと，個々の索引語は検索者の要求と直接結びつく情報を十分持たないものになって

しまう可能性が高い．例えば，
検索者入力１：「回収された新聞や広告の紙はどうなるのか」
という入力から，索引語として，「回収」，「新聞」，「広告」，「紙」が抽出できる．この４個の語句それぞれは使用範囲の広い一般的なものなので，これらのみをキーワードとしたのでは検索者の目的はとうてい果たせない．実際，上記４個それぞれが単独に抽出する記事数は，1221件，998件，1456件，953件である．そのうちで“リサイクル”という検索語を含む記事はそれぞれ，212件，18件，12件，113件であり，”リサイクル”が抽出する記事は1241件であることを考え合わせると，リサイクル問題以外に関する記事の割合が高いことも推測できる．

したがって，検索者の入力から得られた索引語を拡張する必要がある．先に述べたように，索引語の拡張にはシソーラスや概念体系と呼ばれる知識を利用する方法が知られている．これは，検索質問中に用いられる検索者の語彙と文書中で用いられる語彙の不一致を解消することが目的であった．しかし，ここでは，入力１の例でもわかるように，単なる語彙選択の問題ではなく，いくつか語彙の組み合せから全く新しい概念を導入する必要がある．したがって，一般的なシソーラスなどから得られる概念の抽象化や具体化，また類似性や直接的な関連性などの知識では不十分といえる．さらに，入力１の場合には，索引語が一般的な語彙ばかりで，一般的には有効性の認められている分野専用シソーラスも利用できない．

一方，各分野の事典の基本的な構造は，その分野での重要語句である見出し語とそれに対する説明記述からなっている．その記述では，比較的領域専門用語を用いない一般的な表現を用いて専門性の高い見出し語を説明している．そこで，検索者の入力から抽出された索引語が説明記述の中に多く出現する見出し語を抽出すれば，それが，検索者の要求を的確に表現している可能性がある．そうでなくても，検索者の情報要求を具体化する手がかりとなり得ると考えた．

入力例１の場合，環境問題情報事典情報において，説明文に４個の語句が出現する回数が最も多い見出し語は“再生紙”であり，これはユーザの情報要求を最も的確に表した概念の１つと言える．実際，“再生紙”が抽出する記事数は55件であり，そのうち，“リサイクル”を含むものが24件である．

このような元もとの索引語からかなり離れた概念への拡張は従来の索引語の拡張では扱ってこなかったものであるが，検索者が要求を具体化する段階での支援を行うためには必要な機能だと考えられる．

本手法では，このようにして導いた見出し語を検索キーワードとして使用する前に検索者に提示して，検索者の得たい情報との関係の有無を確認する．これによって，単にシステムが検索者の要求を具体化するだけでなく，検索者自身が要求情報を具体化したり，潜在的に持っていながら，表現できなかった情報要求を発見することを支援することができる．

2.2 領域語と一般語の扱い

先に述べたように検索者の入力が自然言語で表現されている場合，抽出できる索引語の個数は少なく，重みに関する有力な情報は得られない．一方，入力文の意味解析は困難なため，意味から索引語間の論理関係を導くこともできない．しかし，次の入力を考えた場合，全ての索引語を同等に扱うことは望ましくない．
入力例２：「市によって分別収集の方法が違うのか.」
ここで，索引語として，{市，分別収集，方法}が得られるが，”分別収集”は単独で有効な検索キーワードと成り得ると考えられるが，”市”や”方法”は単独では検索者の要求に関する有力な情報を持たない．実際“分別収集”が抽出する記事数は39件，“市”，“方法”が抽出する記事数はそれぞれ234件，1909件であり，３個の索引語が共通に抽出する記事数は０件である．

このような区別を個々の入力の解析によって行うことは困難なため，本手法では，領域の用語とそれ以外の一般的な語句として区別し，前者を領域語，後者を一般語と呼ぶことにする．一般語は，それに類似する語句が検索者の要求に適合する文書に出現する可能性が高いが，多様性が高いと考えられる．一方，領域語は，それに類似する

語句が出現する文書が検索者の要求に適合する文書である可能性が高く，多様性は比較的低いと考えらる．本手法ではこのような現象を論理式で表すことにする．したがって，検索モデルとしては，運用レベルのほとんどの情報検索システムが採用しているブーリアンモデルを用いる．

まず，索引語のうち，領域語は単独で1個の検索キーワードと見なす．一方，全ての一般語をOR演算子で結んだものを1個の検索キーワードと対等とみなし，背景キーワードと呼ぶこととする．論理式は検索キーワードおよび背景キーワードをAND演算子で結んだものとする．一般に検索語の集合を用いるシステムは，それらをOR演算子で結合することが多いが，検索語の数が少ないことが予想されること，5.2でのべるように，1件も抽出できない場合は論理式を修正する仕組みを取り入れていることから，ここではAND演算子を採用する．

索引語を領域語と一般語に分類するために，事典の見出し語を利用する．まず，見出し語が領域語であることは議論の余地が無いが，領域語を見出し語のみとするのでは不十分と考える．全見出し語を形態素解析すると平均4.2個の形態素に分解されることからもわかるように，見出し語は，一般的に使われる表現に比べると専門的過ぎる場合や通常は省略した形で使われる名称の正式名称であるなどの理由から通常使われる表現より冗長になっている場合が多い．例えば，一般的には「環境家計簿」と呼ばれる語彙が見出し語では「地球環境家計簿」という表現になっている．「環境家計簿」を検索キーワードとして検索すると，6件の記事が抽出されるのに対して，「地球環境家計簿」を検索キーワードとして検索すると1件も抽出されない．したがって，見出し語の部分表現も領域語と見なす必要がある．ただし，単一の形態素をみると，“運動”や“会”など一般語と見なせるものが多い．そこで，2個以上の形態素の結合とする．例えば「地球環境家計簿」の場合，(地球，環境，家計簿）という形態素のリストで表されるので，それから得られる領域語は次の3語となる．“地球環境”，“環境家計簿”，“地球環境家計簿”

領域語及び一般語を事典情報の見出し語を利用して以下のように手続き的に定義する．

領域語 ：

- 見出し語に一致する索引語
- 見出し語を形態素niのリスト（n1，n2，…，nm）で表したとき，連続する2個以上の形態素の並びで最後が名詞になっているものをすべて取りだし，それぞれその順に結合して生成される名詞句

一般語 ：領域語以外の索引語

3 対話の流れ

検索者とシステムの対話は以下のような流れになる．ここで1度に提示する抽出文書の最大値をMAXと設定する．例えば一画面に表示可能な文書数などを指定する．

1. 検索者：自然言語による自由な入力を行う．
2. システム：検索者の入力から初期検索論理式を生成する．
3. システム：ブーリアンモデルを用いた検索を行う．
4. システム：抽出された文書数がMAX以上の場合：
   (a) 検索者の入力と類似度の高い見出し語を抽出する．
   (b) 抽出できない場合は8へ
   (c) 検索者に見出し語とその説明文を提示し関連の有無を尋ねる
   (d) 検索者：関連有りと判断した場合：
      i. その見出し語を用いて検索キーワードを追加
      ii. 3へ
   (e) 検索者：関係無しと判断した場合：
      i. 次に類似度の高い見出し語を抽出
      ii. 抽出できた場合4（b）へ
      iii. 抽出できない場合8へ

5. 適合文書がない場合：

   (a) 背景キーワードの拡張

   (b) 検索キーワードの修正または削除

   (c) 検索キーワードの OR 結合

   (d) 3へ

6. システム：抽出結果を表示.

7. システム:抽出された文書の有効性を尋ねる

8. システム：検索の続行／終了を尋ねる.

9. 検索者：終了を指定した場合終了

10. 検索者：有効な文書がある場合

    (a) システム：有効な文書の見だしの索引付け処理を行う.

    (b) システム： 索引語を抽出

    (c) システム：索引語のうち領域語をユーザに提示し, 関連の有無を尋ねる.

11. 検索者：追加入力があれば入力

12. 検索者：追加入力した場合

    (a) 追加入力を索引付けし，１０で抽出した索引語とあわせて論理式を生成する.

    (b) 検索者：追加入力しない場合：
    １０で抽出した索引語から論理式を生成する.

    (c) 3へ

4 情報の表現（検索対象・事典情報）と検索者の入力の解析

以降の説明は，具体的に検索対象として 99 年日経新聞の記事, 事典としてと環境問題情報事典を用いて行うことにする.

4.1 情報の表現

環境問題情報事典では，見出し語とその説明記述を用いる．共にまず形態素解析システム「茶筅」[Matsumoto 97] を用いて，形態素を抽出する．見出し語は形態素のリストとして表す．説明記述は簡便な方法で索引付けを行い，索引語の集合としてあらわす．簡便な方法とは，形態素を抽出した後，連続した名詞の結合など可能な結合を施して名詞句を抽出し，その中で代名詞などの不要語を除去したものを索引語とする方法である．その結果を用いて，説明記述は，索引語とその出現回数の組みの集合として表す．例えば，以下に再生紙の説明記述とその表現を示す.

説明記述：

「回収された古紙を原料に配合して再生した紙。配合の割合は様々である（新聞用紙で約４割)。初めて使用されるパルプ（バージンパルプ）のみの上質紙よりやや高価となるが、森林保護と紙ゴミ対策として重要である。日本の回収・再生率は約５０％で先進国でもトップレベルである。」

表現：

((回収 2) (古紙 1) (原料 1) (配合 2) (再生 1) (紙 1)(割合 1) (様々 1) …)

日経新聞記事データは，既に索引付けが行われているので，それを利用する．ブーリアンモデルを用いた検索には予め作成されている転置ファイルを用いる．抽出プログラムは言語処理学会が提供しているドライバーを用いて作成した.

4.2 検索者の入力の論理式への変換

1. 前節で示した方法で索引語の集合に変換する

2. ２．２で示した方法で一般語と領域語を分類する

3. 一般語を OR 演算子で結合して背景キーワードを生成

4. 背景キーワードと領域語を AND 演算子で結合する.

5 論理式の更新

検索者の入力などから作成した検索論理式を使ってブーリアン検索をした結果，抽出された文書が設定した上限値MAXを超える場合，逆に1件の文書も抽出されない場合，以下のような方法で検索論理式を更新する．

5.1 抽出文書数がMAXを超える場合

事典情報を用いてAND演算子によって結合する検索キーワードを追加する．追加する検索キーワードの生成は，検索者の入力から生成された索引語の集合と事典情報中の各見出し語に対する説明記述から生成された出現回数つき索引語の集合との照合を行い，以下の方法で検索者の入力との類似度の高い見出し語を抽出することによって行う．最も類似度の高い見出し語が複数抽出された場合はヒューリスティックを用いて絞り込む．

1. 各見出し語と検索者の入力との類似度を計算し，最も類似度の高い見出し語を求める．類似度は，対応する説明記述と入力から共通して得られる索引語が説明記述中に出現する回数の総和とする．

2. 最も類似度が高い見出し語が複数存在する場合：
「入力から得られる索引語を含んだ見出し語の方が検索者の要求情報を表している可能性が高い」というヒューリスティックスに基づいて，入力から得られる索引語を含んだ見出し語があれば，それらに絞り込む．

3. 絞り込まれた見出し語をその部分表現のOR演算子による結合に変換する．
2.2の領域語の定義と同じ根拠で，絞り込んだ見出し語をそのまま検索キーワードにするのではなく，部分表現をOR演算子で結んだものとする．部分表現の生成方法は領域語の生成方法と同様とする．

5.2 抽出文書がない場合

文書が抽出されない主な原因を分析した結果，次の3項目にまとめた．

1. 背景キーワードの制約が強すぎる
先に述べたように，背景キーワードを構成している一般語は多様性が高いので，検索者の用いる語彙と新聞記事で用いられる語彙の不一致から本来適合すべき文書を排除している可能性がある．

2. 単独でも適合する文書の存在しない検索キーワードがある
次の入力例3において，検索キーワード：“使い捨て文化”は単独でも適合する文書がない．
入力例3：「使い捨てカメラのような使い捨て文化をくいとめるアイデア」
このような検索キーワードは，見出し語の説明記述を用いて，出現する文書が存在する関連概念に変換することができる．

3. 共通する適合文書を持たない検索キーワードの組みがある
次の入力例4において検索キーワード“エコストア”と“エコ商品”はそれぞれ単独では1件，4件の文書に出現するが，同時に出現する文書は存在しない．
入力例4：「エコストアやエコ商品などに基準はあるのか」
関連のある検索キーワードの組みあわせでも，それぞれが出現する文書が少ない場合などにこのようなことがしばしば起きる．このような場合には，それぞれが出現する文書を抽出することが望ましい．

以下にそれぞれの場合に対する論理式の修正方法を述べる

1. 背景キーワードの拡張または削除
論理式中の検索キーワードに対応する見出し語の説明記述のなかに出現する一般語を背景キーワードに加えることにより，背景キーワードの出現領域を拡張する．ここで，論理式が背景キーワードのみの場合は，最も類似度の高い見出し語を抽出して，その見出し語に対して，同様の処理をする．それでも1件も文書が抽出できない場合は，背景キーワードを削除する．

2. 検索キーワードの変換または削除
1件も文書を抽出できない検索キーワードは対応する見出し語の説明記述から得られる索引語をOR演算子で結合した式に変換する．それでも文書を抽出できない場合はその検索キーワードを削除する．

3. 検索キーワードのOR結合
単独では文書を抽出できるが，AND演算子で結合すると，

6 実験の過程で得られた知見

6.1 実験の過程で得られた知見

類似度の高い見出し語を見つけて用いることによって効果的な検索キーワードが得られる場合が以下のように分類できることがわかった．

1. 求める概念自体を知らなかった場合：
入力例1では，検索者は要求を満たすと思われる“再生紙”の概念自体を知らなかったと考えられる．

2. 求める概念自体は知っているが表現方法がわからなかった場合：
例えば「容器の代金をはじめに払う制度」からは，“デポジット”という見出し語を抽出することができる．この1つの索引語は，入力（容器，代金，はじめ，制度）に比べて，端的に要求する情報を表し得る．検索者はこの概念自体は知っていたが，名称を知らなかったために，有効な表現ができなかったと考えられる．

3. 使用した語彙が記事中で用いられているものと一致しなかった場合：入力例3の“使い捨て文化”がその例である．

6.2 本手法の問題点

1. 類似度の高い見出し語抽出処理における問題点
先に述べたように，検索者の入力から抽出される索引語の数が少ないため，1個の索引語の影響で無効な見出し語が多く抽出される可能性がある．抽出された見出し語は検索者に提示して，有効と判断したもののみ採用するため，検索処理自体にはあまり影響がないが，検索者にとって無駄な作業が増大する．

2. 背景キーワードによる再現率の低下．
背景キーワードを構成する索引語は非常に多様性が高いため，検索者の語彙と記事の語彙の不一致による検索漏れが起こる可能性が高い．

7 まとめ

要求情報を十分に具体化できていないユーザを支援して，ユーザの潜在的な要求も含めて具体化し，より適した索引語を導き，それを使って文書を検索するインタフェースシステムの枠組みについて述べた．そのために，対象領域の事典情報の利用を提案した．

検索者の入力から得られる索引語が少なく，重み付けなどの情報も得られないため，事典情報を利用することによって，索引語の拡張と論理関係の生成を行った．定量的な分析までは至らなかったが，システムが自動で行う処理は非常に有効な場合と全く見当はずれの場合があり，それに検索者の判断を有効に組み入れることによって，見当はずれの処理を早い段階で排除できることがわかった．今後はその仕組みを検討していきたい．

参考文献

[Salton 88] Salton,G. Buckley,C. : Term-weighting approaches in automatic text retrieval. Information Processing Management 24(5),513-512(1988)

[Matsumoto 97] 松本, 北内, 山下, 平野, 今, 今村：日本語形態素解析システム「茶筅」 version 1.0 使用説明書 (1997)

[Tokunaga 99] 徳永：情報検索と言語処理 (第5巻)．東京大学出版会 (1999)

[Oddy 77] R.N.Oddy, : Information Retrieval through Man-Machine Dialogue, Journal

of Documentation, Vol.33, No.1, .pp.1-14(1977)

[高野 96] 高野, 平井, 北橋, : ユーザの興味を考慮し発想支援を目指した応答生成手法について, 電子情報通信学会技術報告, Vol.96 No4.4,.pp21-26(1996)

[Frakes 92] Frakes,W.B. Baeza-Yates,R. : Information Retrieval: Data Structures Algorithms. Prentice Hall(1992)