システムソフトウェアと
オペレーティング・システム 77－4
マルチメディア通信と分散処理 87－4
（１９９８．２．２６）

想起型情報検索システムについて

飯田 敏幸*1　松澤 和光*1　池上 徹彦*2　石野 福弥*3　今井 賢一*4
*1 日本電信電話株式会社コミュニケーション科学研究所
*2 ＮＴＴアドバンステクノロジ株式会社　*3 一橋大学　*4 スタンフォード日本センター

概要：情報洪水の中から有用な情報を抽出するためには、効率のよい情報検索システムが必要である。従来のシステムが抱える問題点を解決するために、①利用者の意図に合致した連想の実現、②人間のような状況に応じたダイナミックな連想の実現、③利用者の曖昧な要求を具体化する発想支援タイプの情報検索の実現を目指した想起型情報検索システムの研究を進めている。検索対象としては新聞記事を設定している。辞書の見出し語と説明文中の語との関係、及び、新聞記事中に現れる語の共起頻度を基に作成した単語の意味に関する知識（概念ベースと呼ぶ）を用いて単語間の類似度が計算できる。本システムはこの類似度に従い、語の連想や新聞記事の検索、検索結果のクラスタ化による発想支援をすることに特徴がある。

Associative Information Retrieval System

Toshiyuki IIDA*1, Kazumitsu MATSUZAWA*1, Tetsuhiko IKEGAMI*2,
Fukuya ISHINO*3, Ken-ichi IMAI*4

*1 NTT Communication Science Laboratories　*2 NTT Advanced Technology Corporation
*3 Hitotshubashi University　*4 Stanford Japan Ceter

Abstract: We are promoting the research and development of *Associative Information Retrieval System*, making newspaper articles as the object, to achieve 1) association with higher precision, 2) dynamic association like human beings, and 3) association support type information retrieval which gives an specific form to an ambiguous user's request. Its main characteristics are 1) measuring the semantic similarity between words using *concept base* (knowledge about word's meaning) and 2) grouping newspaper articles by their similarity.

１．はじめに

印刷物やインターネットを通じて提供される情報の量は日々増加し、我々人間が利用できる能力をはるかに越えている。情報洪水の中から有用な情報が抽出できれば、これを知識として蓄積できる。このために、種々の情報検索システムが使われているが、検索結果が余りにも多量であったり、要求通りの結果が得られなかったりする。そこで、このような従来の情報検索システムの問題点を解決するために、我々は想起型情報検索システムと呼ぶ新しい情報検索システムの研究を進めている[1]~[3]。以下、本システムの目的、基本的な考え方、システムの概要等について述べる。

2．目的

テキストを対象とする従来の情報検索システムでは、利用者により指定されたキーワード（KW）を基に、シソーラス（類義語辞書）等により関連するKWを持つ情報を検索し、利用者に提示する形態が主流である。

さて、情報検索に関する我々の行動で最も身近なものである本屋で本を探す行動を考えてみる。すると、少なくとも以下に示す２つの心の状態があることに気付くであろう。

タイプ１：探したい本が明確な状態である。例えば、題名、著者等が分かっている場合である。あるいは、探したい本は具体的に決まっている訳ではないが、おおよその見当がついている状態である。例えば、題名は分からないが、こんなことが書いてある本が欲しいという場合である。

タイプ２：書架の間を歩き回り、目に付いた本を手に取り、何か面白そうな本はないかと探しているような状態である。

上記に照らし合わせてみると、質問KWの羅列、あるいは、それぞれに重みをつけた論理式を与える従来の情報検索システムは、上記タイプ１の利用者を前提としたシステムであると言える。換言すれば、タイプ２の利用者には余り役に立たないシステムである。

確かに、タイプ１のように対象が比較的クリアな場合には、検索対象を代表するようなKWを明示的に宣言しながら、少しずつ対象を絞り込むことにより検索することができる。しかし、システムに与える質問KWの個数が少ない場合には、関連する対象が得られないために再現率（有用な結果が実際に検索される割合）が低くなる。再現率を向上するためには沢山のKWが必要となるが、人手で沢山のKWを指定するには限界がある。そこで、シソーラスを使って関連するKWをシステムが選び（KW連想）、このKWを使って検索がなされる。しかしながら、必ずしも検索の意図が明確になる訳ではなく、かえって適合率（検索結果が有用である割合）が低下してしまうことがある。特に、KWが多義語の場合には、そのKWはシソーラス上の複数のパスに対応するために、多義性を解消せずにKW連想を行うと不必要な語が連想されてしまう。その結果、適合率が低下する。そこで、利用者の意図に合致した連想の実現が必要である。

また、KW連想には与えられたシソーラスに固定された連想しかできないという問題もある。即ち、シソーラスを用いた連想では、ある語から連想される語は常に同じである。一方、人間の連想は状況に応じてダイナミックに変化するものであることから、従来の情報検索システムで使われる「連想」は人間の連想とは全く異なるものであると言うことができる。そこで、このような人間らしい連想をシステムに取り入れることが必要である。

さらに、タイプ２のような使い方は、ネットサーフィンに象徴されるように、当初は対象が漠然としているが、次々と現れる情報に触れているうちに徐々にイメージが明らかになり、あるいはある情報が急に記憶に蘇り、得たかった情報に遭遇するような使い方である。このような情報への接近が、人間の発想を支える重要な部分と想像される。

上記課題を整理すると以下のようになる。

（１）利用者の意図に合致した連想の実現

（２）人間らしい連想の実現

（３）発想支援インタフェースの実現

想起型情報検索システムの目的はこの課題の達成である。

なお、本システムは①情報入力にコストがかからないこと、②情報が追加され固定的でないこと、③多くの利用者が見込めることから、日本語と英語の新聞記事を当面の検索対象としている。

3．基本的な考え方

3．1　概念ベース

本システムの技術的なポイントは単語や文章の持つ意味をどう捉えるかにある。単語の表す意味はその単語の使われ方によって異なる。単語が持つ意味の広がりの中で、文章中で使われている意味を特定するために、以下の意味モデルを考える。

辞書には単語の意味が書かれているが、実際の意味の解釈は辞書を見る人に委ねられている。解釈は人により多少異なるが、平均的には以下のような形式的モデルで表現できると考える。

$$\text{単語Wiの意味} \Leftrightarrow \boldsymbol{m}_i = (m_{i1}, m_{i2}, \cdots, m_{im})$$

但し、m_{ij}は意味素e_jに対する重みである。しかし、意味素e_jを具体的に表現することができないので、現実世界にこのモデルをマッピングし、以下のようにモデル化する。

$$\text{単語Wiの意味} \Leftrightarrow \omega_i = (\omega_{i1}, \omega_{i2}, \cdots, \omega_{in})$$

但し、ω_{ij}は単語Wjに対する重みで、具体的には、辞書の見出し語Wiの説明文に単語Wjが入っている程度、あるいは、新聞記事中での単語WiとWjの共起の程度を表している。このモデル化は、辞書の見出し語とその説明文中に現れる単語とは同じような意味素を有するであろうという仮説と、１つの新聞記事中に現れる単語は同じ話題の中で使われているために、同じような意味素を有するであろうという仮説に基づいている。ベクトルω_iを単語Wiの属性ベクトルと呼ぶ。ここでは、ノルムが１になるように正規化されているものとする。各単語の属性ベクトルを並べた行列

$$\Omega = (\omega_1, \omega_2, \cdots, \omega_n)^T$$

を概念ベースと呼ぶ。このｎ次元空間を意味空間と呼ぶ。

上述のように辞書の見出し語と説明文中の語との関係、及び、文章中での単語の共起関係に基づく２種類の概念ベースがあり、前者を「辞書に基づく概念ベース」、後者を「新聞記事に基づく概念ベース」と呼ぶ。２種類の概念ベースを用意したのは方式比較のためである。

3．2　観点を考慮した語の類似度[4]

一般に単語の意味をベクトルで表す方式では、単語W1とW2の類似度Sim(W1, W2)は、属性ベクトルの内積を用いて、

$$\mathrm{Sim}(W1, W2) = \omega_1 \cdot \omega_2$$

のように定義される。しかし、この方法では多義性の問題が残ってしまう。そこで、以下に示す観点という考え方を導入する。例えば、動物の話をしていると、馬は自動車よりも豚の方に似ていると感じるが、乗り物の話をしていると馬は自動車の方に似ていると感じる。動物や乗り物は文脈や状況に相当し、これを観点と呼ぶ。観点も語であり、属性ベクトルを持つ。ある観点から語を見ることを観点の属性ベクトルの特徴的な成分を重視することに対応させる。これは具体的には、観点の属性ベクトルのうち値の大きい成分について語の属性ベクトルを強調すること（変調と呼ぶ）により実現している。このようにして２つの語をある共通の観点から見ることにより、語の多義性を解消した類似度が計算できる。即ち、観点Wvにより変調された属性ベクトルω_1', ω_2'を用いて、観点を考慮した語の類似度を

$$\mathrm{Sim}(W1, W2, Wv) = \omega_1' \cdot \omega_2'$$

のように定義する。これにより、観点Wvから見た語W1から連想される語の集合Φは

$$\Phi = \{ x \mid \mathrm{Sim}(W1, x, Wv) > \theta \}$$

で求められる。但し、θ （$0 < \theta \leq 1$）は閾値である。これにより確度の高い連想や人間らしい連想が可能となる。

なお、実際には各属性ベクトルの要素のうち大半は０であり、類似度はほとんどの場合

０になってしまう。そこで、属性ベクトルの次元を減らす工夫をしている。

３．３　類似度を用いた記事検索

文章を単語の羅列と捉えると、文章、即ち、記事も意味空間上のベクトルとして表現できる。そこで、語と語の類似度と同様に、語と記事の類似度、記事と記事の類似度も属性ベクトルの内積として定義できる。これにより、KW検索、質問文検索、あるいは類似記事検索が実現できる。概念ベースを用いるKW連想に必要となる観点は、入力されたKWあるいは質問文中のKWの合成ベクトルを用いる。

類似度の高い記事同士をクラスタとしてまとめ、各クラスタの特徴を提示できれば、この情報を使った絞り込みが可能となる。クラスタの特徴としては、重心ベクトルとの類似度が高い語が候補となる。この方法は発想支援にも使える。

３．４　未知語の扱い

概念ベースにない語（未知語と呼ぶ）が質問KW、質問文、あるいは類似記事検索の元となる記事に現れた場合の対処が問題になる。辞書に基づく概念ベースを用いる場合には、概念ベース構築に利用した辞書の見出し以外の語は未知語となる。新聞記事に基づく概念ベースを用いる場合には、概念ベース構築に利用した記事に出現しない語、あるいは、出現頻度が低い語は（効率上、概念ベースには入れていない）未知語となる。未知語は０ベクトルに対応するので、３．２で定義した類似度は使えない。そこで、未知語が質問中に入った場合には、以下により類似度を計算する。

$$類似度=\alpha \cdot S+(1-\alpha)\cdot n/N$$

但し、αは重み（$0 \leqq \alpha \leqq 1$）、Sは３．２で定義した類似度、Nとｎはそれぞれ質問中の未知語KW数と記事中に含まれる未知語KW数である。

４．システムの概要

４．１　システム構成

本システムは情報管理プログラムと検索処理プログラムから構成されている。情報管理プログラムは概念ベースと新聞記事DBを構築するバッチ処理プログラムである。検索処理プログラムは主に概念ベースを用いた新聞記事の検索機能、方式比較のためのシソーラスを用いた従来型の検索機能、評価データを収集する機能から構成されている。検索処理プログラムの利用者インタフェースはWindows９５上のクライアントとしてWWWブラウザ上にJAVAを用いたオブジェクト指向型の処理として実現され、サーバ側の機能はUNIXワークステーション上にCGIによって呼び出されるモジュール群として実現されている。システム構成を図４．１に示す。

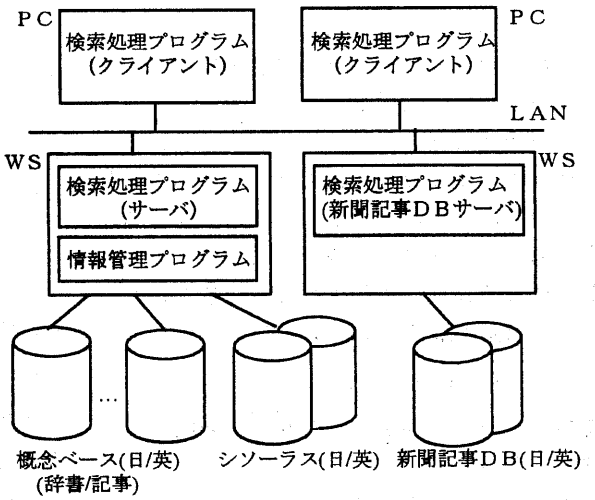

図４．１　システム構成

４．２　機能概要

情報管理機能と検索処理機能のうち主要な機能について簡単に説明する。

（１）情報管理機能

ａ）日本語と英語の概念ベース構築機能

電子化された複数の国語辞書を入力とし、説明文を解析して抽出された単語と見出し語の関係から辞書に基づく概念ベースを構築する[4]。同様に、電子化された新聞記事を入力とし、記事中の文毎に抽出された単語の共起頻度を計算し、これを元に新聞記事に基づく概念ベースを構築する[5]。各概念ベースは日本語、英語の２種類があり、データベースとして管理される。

ｂ）新聞記事データベース構築機能

見出しと本文を単語分けし、概念ベースに基づき見出しベクトルと本文ベクトルを計算し、見出し、本文等と一緒にデータベースに格納する。

（２）検索処理機能

ａ）条件設定機能

日本語の新聞記事を検索対象とする時には日本語の概念ベース／シソーラスを、英語の新聞記事を検索対象とする時には英語の概念ベース／シソーラスが使われる。辞書に基づく概念ベースと新聞記事に基づく概念ベースのどちらを使うかは利用者の選択による。当然、検索インタフェースは日本語と英語の切り替えが可能である。

検索対象が膨大であるので、日付、ジャンル等を指定して検索範囲を設定できる。

ｂ）記事検索機能

KWや質問文を入力し、概念ベースやシソーラスを用いたKW連想によりKWが得られる。必要であればKWを追加／削除し、得られたKWを基に記事検索を行うことができる。概念ベースを用いた場合には、３．３で説明した方法が使われる。また、記事間の類似度計算により検索された記事を指定して類似記事を検索することができる。検索された記事の記事id、類似度、見出しが表示され、見出しからのハイパーリンクにより本文を表示できる。実行例を図４．２に示す。更に、検索された記事を相互の類似度に従ってクラスタ分けし、各クラスタの特徴語を提示することができる。これらの処理は繰り返し行うことができ、また、処理の途中で概念ベースの種類を変更することも可能である。検索処理の過程は履歴情報として管理され、これを参照することにより記事間の渡り歩きの際に迷子になる問題を解決している。

ｃ）評価機能

記事検索の評価指標としては再現率と適合率を使うことにしている。ベンチマークを用いて、辞書に基づく概念ベース、新聞記事に基づく概念ベース、シソーラスのそれぞれについて自動的に再現率、適合率を計算することができる。また、検索要求から得られた記事それぞれに対して、その検索要求に適合し

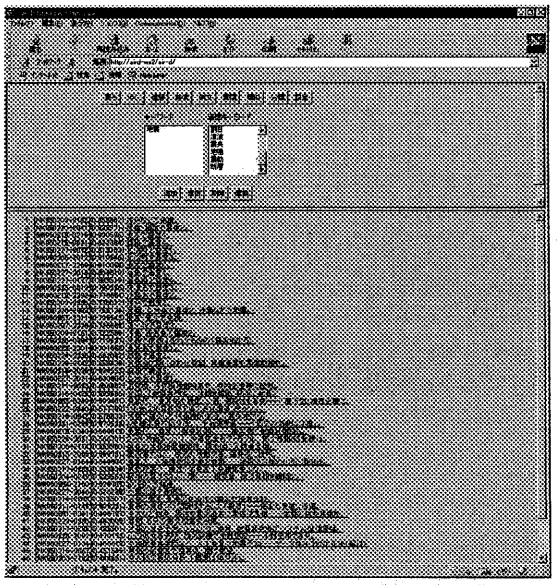

図４．２　実行例

ているか否かを入力するインタフェースがあり、この情報を用いて適合度を計算する。一方、発想支援に関しては利用者の満足度を指標とすることを考えている。

5．ユーザインタフェースの改良

本システムのユーザインタフェースはテキストベースであり、一見したところ従来の情報検索システムと余り相違が見えない。特に、タイプ２の利用者にはまだ使いにくそうである。記事の関係が視覚的に捉えられればこの問題は解決できると考え、３次元表示を利用した情報検索インタフェースを開発した[6]（図５．１参照）。今後、本システムとの結合を予定している。

6．おわりに

本システムを試用してみた感想を簡単に述べる。

①辞書に基づく概念ベースを用いた場合には、連想KWとして思ってもいない語が得られることがあり、発想支援に向いていそうである。

②新聞記事に基づく概念ベースを用いた場合には、連想KWとして時事的な語が得られるので、KWの入力漏れが防げそうである。

③新聞記事よりも辞書に基づく概念ベースを用いた方がクラスタが鮮明に分かれる。

今後、本格的にシステム評価を行う予定である

想起型情報検索システムはＩＰＡ創造的ソフトウェア育成事業による。

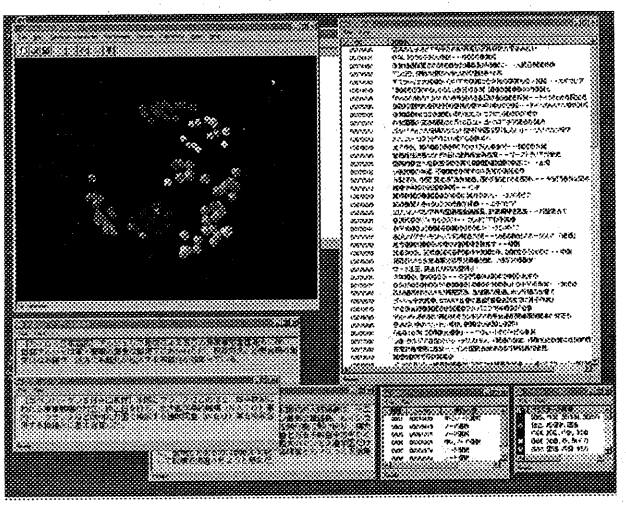

図５．１　３Ｄ表示インタフェース

【参考文献】

[1]飯田敏幸, 松澤和光, 松田晃一, 池原悟, 石野福弥, 今井賢一：想起型情報検索方式の提案, 情報処理学会第５３回全国大会, 1T-01, pp. 3-153-154(1996).

[2]松澤和光, 飯田敏幸, 松田晃一, 今井賢一：想起型情報検索システムの基本構想, 情報処理学会第５３回全国大会, 1T-02, pp. 3-155-156 (1996).

[3]飯田敏幸, 松澤和光, 遠藤隆也, 上田洋美, Stanley Peters, 石野福弥, 今井賢一:想起型情報検索と概念創造のソフトウェアの開発, 創造的ソフトウェア育成事業 及び エレクトロニック・コマース推進事業 中間成果発表会論文集, 創造的ソフトウェア育成事業編, pp. 383-387(1997).

[4]笠原要, 松澤和光, 石川勉:国語辞書を利用した日常語の類似性判別, 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272-1283(1997).

[5]Schütze, H. and Pedersen, J. O.: Information Retrieval Based on Word Senses, 4th Annual Symposium on Documets Analysis and Information Retrieval, pp. 161-176(1995).

[6]飯田敏幸, 熊本睦, 松澤和光, 今井賢一:情報検索のための３Ｄインタフェース, 情報処理学会第56回全国大会, 4Z-04, (1998)(予定).