

想起型情報検索システムの基本構想

1T-2

松澤 和光† 飯田 敏幸† 松田 晃一† 今井 賢一‡

† NTT(株)コミュニケーション科学研究所 ‡ スタンフォード大学

1. はじめに

個人がネットワークを通じて膨大な情報を扱えるようになりつつあるが、真に自分の望む情報を手に入れるのは簡単なことではない。そもそも望む情報が何であるかさえハッキリしない場合も多い。このため、検索対象を明確に指定する従来法に代わって、対象が漠然としていても柔軟に検索が行える新しい方式が必要とされている[1]。

こうした要求に応えるため、検索を進めながら対象を明確化する「想起型情報検索方式」を提案[2]した。本稿では、この提案を実現するための具体的な機構として、NTTが開発した「概念ベース技術」[3]を利用する方法を述べる。また、この機構を使った新聞記事の検索システムについて、その基本構想を述べる。

2. 想起型情報検索方式

従来の情報検索法はキーワードの一致を基本としているが、より柔軟な検索のため、キーワードに関連する他の単語に展開する方法がある。展開は一般にシソーラス等を用いる。しかし人間が同様のことを「連想」として行う場合、シソーラスのような固定的に設定された類語への変換ではなく、検索の状況に応じたもっとダイナミックな展開を行っていると考えられる。

また、ネットサーフィンに象徴されるような検索では、初めは対象が漠然としていても、次々に現れる情報に触れることでイメージが次第に明らかになって行く場合がある。これは、検索を行っ

ている人間が、その検索行為に触発された一種の概念創造を行っていると考えられる。

このような人間の持つ連想や概念創造の機能を機械的に実現するため、キーワードや被検索データの持つ「意味」を意味空間のベクトルでモデル化し、キーワード集合の連想による拡張と、検索結果からのキーワード生成を繰り返す。これが想起型情報検索方式の基本的考え方である。

3. 概念ベース技術による実現

3.1 概念ベース技術

概念ベースとは、NTTで研究中の「アバウト推論方式」[4]において開発した、言葉の意味(概念)に関する知識ベースである。概念ベースでは、各概念を属性と重みの対の集合で表す。属性と重みは、国語辞書等の語義文における自立語と出現頻度から獲得する。この結果、例えば、

概念「馬」= (動物, 0.4), (家畜, 0.2) ...

と言ったデータを構成できる。実際に、約4万の日常語について2百万属性から成る概念ベースを構築した。

この概念ベースを用いて、概念間の類似性をその概念が使われる状況に応じて判別する仕組みも開発した。具体的には状況を表す概念を「観点」として指定すると、観点のデータで特徴的な属性について各概念の重みを強調する(観点変調)。この変調した概念同士を属性空間中のベクトルの成す角で比較して類似度を計算する。この結果、観点「動物」の場合、「馬」と「豚」の類似度が

Recollection-based Information Retrieval System

*Kazumitsu Matsuzawa, Toshiyuki Iida, Kouichi Matsuda

NTT Communication Science Laboratories 1-2356, Take, Yokosuka, 238-03 Japan

**Ken-ichi Imai

Stanford University 52-2 Hoshoji-cho, Okazaki, Sakyo-ku, Kyoto, 606 Japan

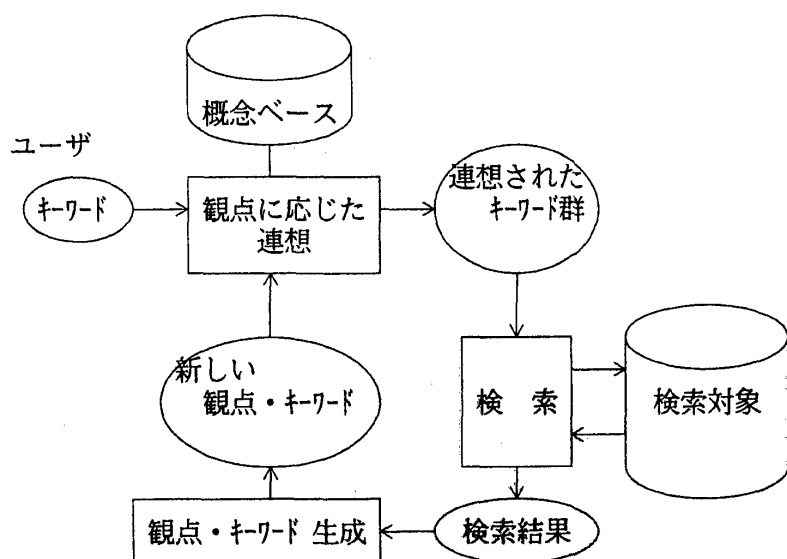


図1. 想起型検索のモデル

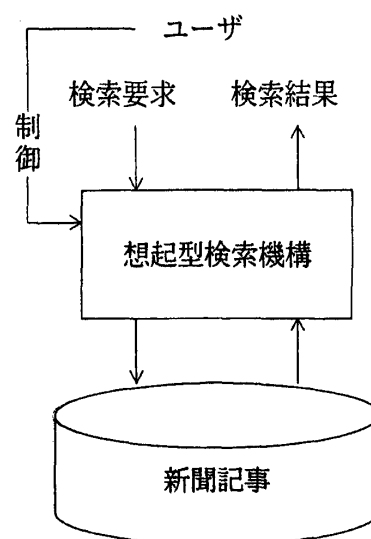


図2. 検索システムのイメージ

0.8 等と計算できる。

さらに、観点をそれら概念が使われている文章から自動生成する研究[5]も進めている。

3.2 想起型検索のモデル

前節の概念ベース技術を用い、想起型情報検索を実現するモデルとして、図1を提案する。

まず、ユーザが入力したキーワード(群)は、検索の観点に応じた類似性判別によって、関連したキーワード群に展開される。つまり、状況に応じた動的な「連想」が行われる。

次に、このキーワード群により検索された結果から、新たなキーワードや観点を生成する。これが「概念創造」を行ったことになる。

新しい観点・キーワードに基づいて、再び連想を行う。この繰り返しで想起型検索を実現する。

なおこのモデルで、観点の変調度を変化させることにより、連想を発散的にも収束的にも行える可能性があり、検索動作の制御に利用できると考えている。

3.3 検索システムのイメージ

前節のモデルに基づき、具体的な検索システムを構築して、想起型検索方式の評価を行う予定である。検索対象としては、

・電子化されている(情報入力コストゼロ)

・本方式の有効性確認後、多くの利用者を見込む

・情報の追加がある(閉じた世界は現実性なし)

等の条件から、CD-ROMで提供される新聞記事が適切と考えている。

図2に検索システムの全体イメージを示す。

4. おわりに

今後、本方式による新聞記事検索システムを実際に構築し、評価実験を行う予定である。

検索システムの開発はIPA創造的ソフトウェア育成事業による。

【参考文献】

[1] Imai, K: KB & KS as a New Economic and Social Infrastructure, Proc. of Int. Conf. on Building and Sharing of VLKB '93, p23, 1993

[2] 飯田 勲「想起型情報検索方式の提案」第53回情処全大, 1T-01, 1996

[3] 笠原 勲「観点に基づく概念間の類似性判別」情処論, Vol. 35, No. 3, pp605-609, 1994

[4] 松澤 勲「7割推論とその類似性判別機構」AI学会, SIG-J-9401-14, 1994

[5] 笠原 勲「類似語検索における観点の自動生成法」情処情報学基礎研究会, 1996. 7