論 文 の 内 容 の 要 旨

論文題目　　オントロジーを利用した文書検索手法の研究

氏名　　尾暮拓也

たとえば図書館で何らかの事柄に関する文献を検索する必要があるとき、かつ、その事柄を概念的にしか説明できないとき、既存のキーワード検索などでは必ずしも望ましい結果が得られない一方、多くの場合図書館員や専門家に相談することによって容易に参照すべき文献に辿り着くことができる。ここで「事柄を概念的にしか説明できない」とは、知りたい内容を直接指す言語的ラベルを知らず、関連すると思われる単語を述べることしかできないという意味である。図書館員や専門家がこのような質問に答えることができるのは、質問された領域に関する概念体系を持っているためと考えられる。

領域依存の概念体系を扱う上では、明示的な領域の定義を持つオントロジーと呼ばれる知識体系を利用することができる。オントロジーという語はもともと哲学用語で「存在論」を意味するが、工学的に用いる場合は一般的には問題解決のために考慮する対象物とそれらの関係を明示的に記述したデータベースである。理想的にはオントロジーは対象とする問題領域の網羅性と、概念体系を記述したものとして人々の合意が得られていることが期待できなければならないとされる。そこで本研究では専門的な文書など対象領域が明確な検索タスクにおいてオントロジーを積極的に利用する検索モデルとしてオントロジーベクトルモデルを提案し、このモデルの性能と特性を実験で検証する。

情報検索の分野ではさまざまな手法が提案されている。ここでベクトルモデルと呼ばれる検索モデルについて紹介する。この手法は検索に使用する索引語があらかじめ用意されている場合に用いる。検索質問は索引語の任意の単語を重み付けして与える。このとき検索質問ベクトル $\mathbf{q}$ は次の式で与えられる。

$$q_i = g(term_i)$$

$$g(term_i) = \begin{cases} 0 (term_i が指定されなかったとき) \\ n (term_i が重み n で指定されたとき) \end{cases}$$

$$term_i : 索引語$$

文書 k に対応する文書ベクトル $\mathbf{d}_k$ は次の式で与えられる。

$$d_{k\,i} = f_k(term_i)$$

ここで $f_k(term_i)$ は文書 k の中で $term_i$ が出現する頻度である。文書が検索質問にどの程度適合しているかは両ベクトルの内積で定義される類似度 Sim_k によって判断する。類似度が高いほど検索質問に適合していると考える。

$$Sim_k = \frac{\mathbf{q} \cdot \mathbf{d}_k}{|\mathbf{q}\| \mathbf{d}_k|}$$

本研究ではベクトルモデルのベクトル要素を単語から概念に拡張した「オントロジーベクトルモデル」と呼ぶ情報検索モデルを提案する。オントロジーベクトルモデルの概要は次の通りである。

1. 文書集合が対象とする領域のオントロジー $O_{domain}=\{concept_h\}$ を用意する。
2. 全ての索引語 $term_i$ について、オントロジーベクトル $\mathbf{V}\,term_i$ を定義する。

$$V_{term_i\,h} = r(term_i, concept_h)$$

ここで $r(term_i, concept_h)$ は $term_i$ の $concept_h$ への関連の強さである。

3. ベクトルモデルと同様に定義された検索質問ベクトル $\mathbf{q}$ および文書ベクトル $\mathbf{d}_k$ をオントロジーベクトル $\mathbf{Q}$、$\mathbf{D}_k$ に変換する。ここで $\mathbf{M}$ は単語空間から概念空間への写像行列である。

$$\mathbf{Q} = \mathbf{q}\mathbf{M}$$

$$\mathbf{D}_k = \mathbf{d}_k\mathbf{M}$$

$$\mathbf{M} = \begin{pmatrix} \mathbf{V}_{term_1} \\ \mathbf{V}_{term_2} \\ \vdots \\ \mathbf{V}_{term_n} \end{pmatrix}$$

4. 文書が検索質問にどの程度適合しているかは両オントロジーベクトルの余弦で定義される類似度 Sim_k によって定義する。

$$Sim_k = \frac{\mathbf{Q} \cdot \mathbf{D}_k}{|\mathbf{Q}||\mathbf{D}_k|}$$

ここで重要なのは 2.の $term_i$ の $concept_h$ への関連の強さの値 $r(term_i, concept_h)$ を決める過程であり、このモデルでは「活性伝播」と呼ばれる手法を用いる。活性伝播とは人間の意味記憶における「連想」をモデル化したものであり、索引語 $term_i$ に直接関係する概念を人間が選択することにより、その概念の近傍の概念 $concept_u$ とその索引語 $term_i$ との関連の強さ $r(term_i, concept_u)$ が自動的に決定される。活性伝播ルールは一般に活性が減衰していくように設計するが、ルールは任意であり、最適化は今後の課題である。

本研究の実験には典型的な情報検索の評価方法であるテストコレクションを用いた評価を行う。テストコレクションとは次の 3 つのデータのセットである。

- 試験用の文書群.
- 複数の検索質問.
- 検索質問にそれぞれの文書が適合しているかどうかを評価した正解集合.

NTCIR[8]は国立情報学研究所が 1997 年から継続しているテストコレクション構築プロジェクトで、提供されている日本語テストコレクションは日本では標準的な存在である。このテストコレクションには同研究所が提供している「学会発表データベース」のアブストラクトの部分を試験用の文書として約 34 万文書を収録している。検索質問は 83 検索質問文が用意され、これらにとそれぞれの文書との適合性が人間によって判断され、正解集合として用意されている。

本研究で提案するモデルはオントロジーを用いるが、オントロジーは一般に領域を限定して構築されるので、対象とする特定領域として人工知能分野を選び、34 万文書のうち「人工知能学会」から提供された 2,013 文書を実験用に選んだ。この中に正解文書が 5 件以上ある検索質問は 10 件であり、実験にはこれらを用いた。正解文書の平均数は 12.5 文書である。

人工知能分野のオントロジーとしては、人工知能学会への論文投稿時に指定する階層化された分野分類を当研究室のメンバーに公開して修正の意見を集め、オントロジーとして合意を得たものを使った。最終的な概念数は 133 であった。

索引語は実験に使用する 10 検索質問に含まれる内容語（名詞、動詞、形容詞）のうち、一般的過ぎる単語（「文献」、「記述」など）を除く 21 単語と、人工知能の教科書的な書籍の索引にある単語約 2 千語のうち、一般的過ぎる単語を除いた 1,051 語を選んだ。これら合計 1,072 索引語に関して、対応する概念を先の 133 の概念からなるオントロジーに加え、活性伝播によりオントロジーベクトルを作成した。

今回の実験では単語に関係のあると思われた概念 $concept_1$ に $r(term_i, concept_1) = 1$ を設定し、活性伝播ルールとして、以下の式を適用した。

$$r(term_i, concept_h) = \mathrm{a}\, r(term_i, concept_{h-1})$$

ここで a は活性伝播率と呼ぶことにする。

このルールにより、同じ概念を指す単語同士、上位概念が同じ単語同士、上位概念も違う単語同士のオントロジーベクトルの内積をそれぞれ P_1, P_2, P_3 とすると

$$P_1 > P_2 > P_3 = 0$$

の関係が成り立ち、概念の階層構造を反映する。検索質問は検索質問文に出現した索引語を出現頻度で重み付けしたものを用いた。

これらのデータを用いてaを0.0から1.0まで推移させてパラメーターサーベイを行った。ここでa＝0.0のときは既存の単語のベクトルモデルにとほぼ同等であり、従って本研究で提案するオントロジーベクトルモデルとの性能の比較ができる。

情報検索の評価は精度(precision)と再現率(recall)を元に行われるのが通例である。それぞれの定義は以下の通りである。

$$\text{精度} = \frac{\text{正解文書の中で検索された文書数}}{\text{検索された文書数}}$$

$$\text{再現率} = \frac{\text{正解文書の中で検索された文書数}}{\text{正解文書数}}$$

計算結果を Sim_k 順に並べ、上位からn番目の文書までを検索結果とした場合の精度 Precision(n)と再現率 Recall(n)を各検索質問に関して平均した。それぞれ図1、図2に示す。また再現率が0.0から1.0までの11点での精度の平均値「11点平均精度」を図3に示す。

Sim_k が0より大きい有効な検索結果は、オントロジーベクトルモデルを用いたときで全ての検索質問に対して2031文書中1987文書であった。単語のベクトルモデルでは最大325、最小0、平均58.1文書であった。

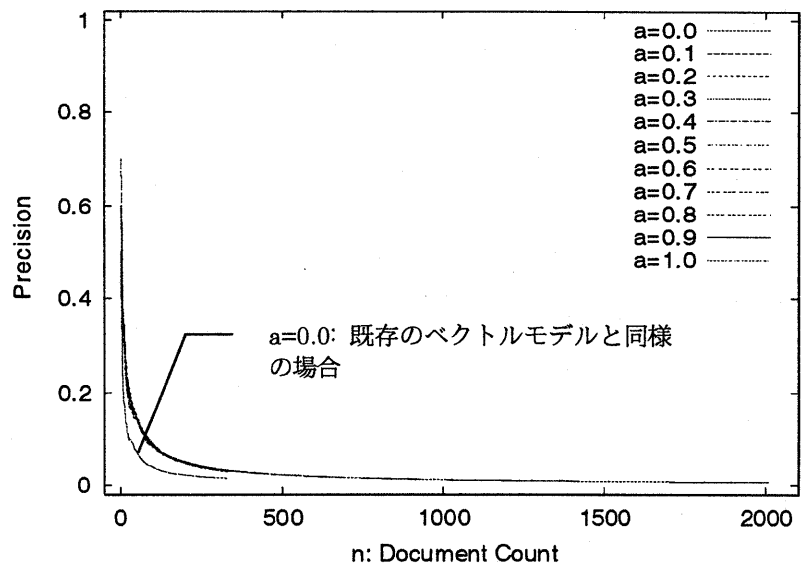

図 1　各活性伝播率における精度の平均

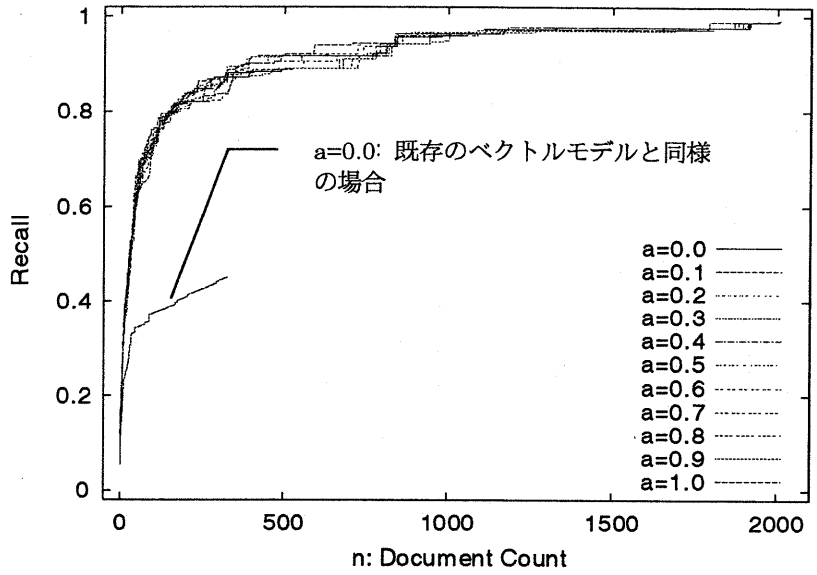

図 2　各活性伝播率における再現率の平均

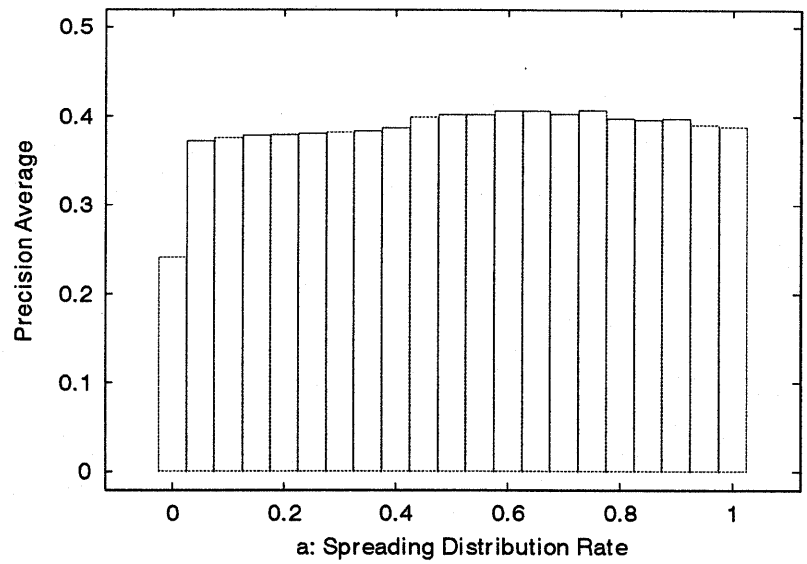

図 3　各活性伝播率における 11 点平均精度

図 1、図 2 と図 3 から、既存のモデルに比べて性能が大きく向上していることが分かる。また検索された文書数も多い。特に再現率に関しては 2 倍以上の性能となり、図 3 の１１点平均精度のグラフからは精度が 1.5 倍ほどに向上していることが分かる。また活性伝播率「a」の推移に伴って１１点平均精度は変化し、a=0.6 付近でピークが出ている。この変化はオントロジーの構造が検索性能に寄与している証拠である。また 133+1,072 概念のオントロジーと 1,072 語の索引語は人工知能分野の広さに対してカバーする範囲が狭く、この変化が小さいのは、オントロジーの概念数が少ないためと考えられる。

以上から、提案するモデルの性能とオントロジーに依存する特性を検証することができた。