日本文意味検索に必要な最小単語意味属性の組の決定

3V－6

木本 泰博† 池原 悟† 白井 諭‡
†鳥取大学工学部 ‡NTTコミュニケーション科学研究所

1 はじめに

従来の検索技術として、文書中の単語などを元としたベクトル空間モデルで文書の内容を表し、ベクトルの距離で類似性を判定し、検索するベクトル空間法がある。単語をベクトルの元とした場合では、元が多くなるため、類似度の計算量が多くなる。これに対して、ベクトルの元に単語意味属性 (2700 種) を用いる検索法[1]が提案されているが、この方法でも、ベクトルがスパースになり類似性が判定できなくなること、類似度の計算量が多いことが問題点となっている。スパース性と計算量の問題点を解決するために、本研究では単語意味属性相互の意味的関係[2]に着目し、最適な検索のための最小単語意味属性の組を決定する。

2 意味属性を用いたベクトル空間法

本手法は検索要求として文書を与え、検索要求と検索対象の文書を単語の意味によるベクトル空間モデルで表わし、ベクトルの距離で類似性を判定し、検索要求に類似した文書を検索する意味的検索法である。

(1) 文と文書の意味の表現

ベクトル空間法では文の意味を文中の単語などを元とする意味ベクトルに重みを与えて文の意味を表す。本手法では、文の意味の多くは文中の名詞から判断できるため、意味ベクトルの元は名詞の意味に相当する一般名詞意味属性を用いる。単語のそのものを用いるのではなく、意味属性を意味ベクトルの元とすることで、同義語・類義語を含む文に対して、同一の意味を表すことが可能になる。文書は文の集合であるから、文書の意味は文書中の文の意味ベクトルの和で表す。

文書 A の意味ベクトル $v_A = \sum_i v_{Ai}$

v_{Ai}:**文書 A 中の文 i の意味ベクトル**

(2) 意味ベクトルの元の値

意味ベクトルの元の値には文の意味に対して、より特徴的な元に大きな重みを与えることで文の意味を表現する。そこで、文中に多く出現する単語は特徴的であると思われるため、文中の名詞に与えられる意味属性 i の出現頻度 m_i を a 重みとし、意味ベクトルの絶対値が 1 となるように正規化した値を意味ベクトルの元 i の値 p_i とする。

$$p_i = \frac{m_i}{\sqrt{\sum_i m_i \cdot m_i}}$$

(3) 文書の類似度の判定法

ベクトル空間法では、文書の意味ベクトル間の距離が近いほど類似している文書とする。そこで本手法では、検索要求文書と検索対象文書との類似度は各文書の意味ベクトルの内積をとり、cosin の値とする。

検索要求文書と検索対象文書の類似度:$sim = v_a \cdot v_b$

v_a:**検索要求文書の意味ベクトル**

v_b:**検索対象文書の意味ベクトル**

3 最小単語意味属性の組の決定

従来の検索技術として、文書中の単語などを基底としたベクトル空間モデルで文書の内の単語意味属性を用いることで意味的な検索が可能になったが、ベクトル空間法では、ベクトルのスパース性により類似性が得られない場合があることと類似度計算量の多さが問題となっている。そこで、意味属性の意味的関係に着目して、意味属性を汎化することで、最適な検索のための最小意味属性の組を決定する。

3.1 意味ベクトルの元の汎化

(1) 意味ベクトルの元の意味的関係

単語意味属性とは対象を概念化する際の視点である。そして、意味属性間には上位-下位、全体-部分の意味的関係が存在し、意味属性全体は最大 12 段のツリーで構成されている。この意味的関係を利用すれば、下位の意味属性を上位の意味属性で代表することが可能である。

(2) 意味ベクトルの元の汎化方法

意味ベクトルの元の汎化とは、意味ベクトルの元となる意味属性のうち、下位の頻度の少ない意味属性をボトムアップし、上位属性を下位属性の代表として、意味ベクトルの元とすることである。しかし、最上位の意味属性を意味ベクトルの元として、頻度の多い意味属性を細分化して、トップダウンで意味ベクトルの元を選択しても同様の結果が得られる。今回、比較的上位の意味属性で最適な組は構成されるとの予想から計算上、効率の良いトップダウンの方法を用いる。意味属性の細分化は以下の手順で行う。

1. 新聞記事 100 件中の名詞に付与される意味属性を 4 段目まで縮退させる。
2. 意味属性の頻度統計をとる。
3. 意味属性の頻度 m が $2 \leq m \leq l(l = 100, 50, 20, 10)$ になるまで意味属性を細分化する。

(1. は ALT-JAWS により形態素解析を行い、プログラムにより自動的に名詞を抽出して、意味辞書を参照し、意味属性を付与して、4

Minimum Set of Semantic Attribute for Japanese Document Retrieval
Yasuhiro KIMOTO†,Satoru IKEHARA†and Satoshi SHIRAI‡ †Tottori University ‡NTT Communication Science Laboratories

段目まで縮退させる。)

以上の手順で細分化した意味属性を意味ベクトルの元として選択し、選択されなかった下位属性 z_j の元の値 p_{ij} は、属性 z_j を代表する属性 z_i の値 p_i に吸収する。

$$p_i = p_i + \sum_j p_{ij}$$

さらに、下位属性の元の値を上位属性の元の値へ伝搬することで、意味ベクトルの元の汎化を行う。しかし、下位属性の元の値をそのまま上位属性の元の値へ伝搬させると、下位属性の特徴がぼけるので値の一部を伝搬させることにする。**意味属性 z_j から親ノードの意味属性 z_i へ伝搬する値** $p(z_i, z_j)$ は以下に示す。

$$p(z_i, z_j) = \frac{z_j\text{の元の値}}{z_i\text{の子ノードの意味属性の総数}}$$

3.2 最適化な組の決定手順

(1) 意味属性の組の選択手順

最適な意味属性の組の決定は意味ベクトルの元を汎化して、検索実験を行い検索精度を評価して決定する。手順を以下に示す。

1. 4 段目 (21 種) の意味属性を意味ベクトルの元とする。
2. 検索実験を行い、検索精度を評価。
3. 検索精度が最大ならば、最適な意味属性の組が決定。
4. 意味ベクトルの元を汎化し、2 へ戻る。

(2) 検索精度の評価の方法

(1) で最適な意味属性の組を求める際に、検索の精度を評価する必要がある。そこで、検索精度の評価パラメータとして適合率・再現率の積を用いる。

$$\text{適合率 } P = \frac{\text{正しく検索した文書数}}{\text{検索した文書の総数}}$$

$$\text{再現率 } R = \frac{\text{正しく検索した文書数}}{\text{類似した文書の総数}}$$

$$\text{検索精度} = P \times R$$

ここで、類似した文書とする判定は人それぞれの着目点により異なる。そこで本研究では文書情報として新聞記事を扱い、類似とする基準を新聞記事に対して、3 段階の類似度で定義する。

類似度 A 同一内容の記事
類似度 B 関連する内容の記事
類似度 C 同類内容の記事

類似度 A とは、ある出来事に対する A 新聞者の記事に対して、同じ出来事について、同じ内容の B 新聞社の記事が相当する。また、類似度 B はある出来事の記事のその後の展開、批評などが相当する。類似度 C に関しては、ある出来事に対し、日時・場所・当事者は異なるが、同じ様な出来事についての記事が相当する。

4 実験と評価

4.1 最適な意味属性の組の決定

(1) 実験の条件

検索要求はランダムに選んだ 24 件の新聞記事、検索対象は検索要求記事を含む 150 件の新聞記事を用いた。また、同様の出来事を含む記事の方が客観的に見て類似していることから、検索要求に対する正解記事は類似度 A と類似度 B の類似記事とした。

(2) 結果と考察

実験結果は図 1 に示す。図 1 から検索精度は約 200 ～ 300 の意味属性を元とした時、ほぼ最大となることがわかる。

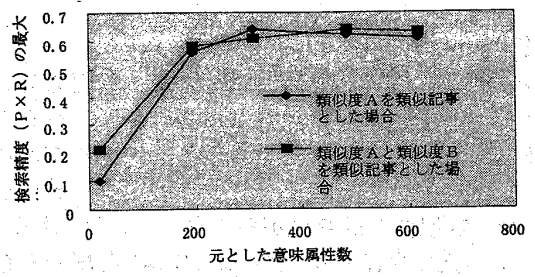

図 1: 検索精度の最大値の変化 (検索対象記事 150 件)

4.2 検索対象文書数の影響

検索対象を 150 件とした場合、200 ～ 300 の意味属性の組を意味ベクトルの元とすることで検索精度は限界に達したが、通常、検索対象となる文書は膨大な量である。ここでは、検索対象を 1000 件まで増やし、検索精度と意味属性の組との関係を調べた。実験結果を図 2 に示す。実験結果より、検索精度が限界に達するには、より多くの意味属性の組が必要となる。また、検索精度は 0.6 以上から 0.5 程度まで下がる。このように、大量の記事から検索する場合は、より多くの意味属性の組が必要かもしれない。しかし、500 種程度の意味属性の組のとき、検索精度はほぼ限界に達するので、計算量を考慮すれば、最適な検索のための最小意味属性数は 500 程度といえる。

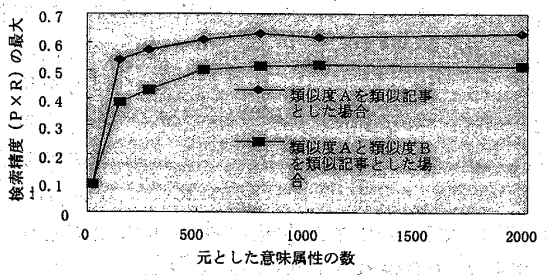

図 2: 検索精度の最大値の変化 (検索対象記事 1000 件)

5 おわりに

本論文では、単語意味属性を用いた意味的検索を最適化するために、意味属性の意味的関係を利用し、意味属性の汎化を行った。汎化により、約 500 種の意味属性の組で最適な検索を実現できることがわかった。また、検索精度も 0.5 以上であり従来のキーワード検索 (0.2 ～ 0.5) に比べて、本手法の有効であることがわかった。今後は、約 500 種の意味属性の組をベースとして、意味ベクトルの元の値に tf*idf などの適用、動詞の名詞化などにより検索精度の向上を目指す。

参考文献

[1] 松尾、内野：意味属性に基ずくテキストデータベース検索方式、情報処理学会論文誌 Vol.32.No.9(1991)

[2] 池原、宮崎、白井、横尾、中尾、小倉、大山、林：日本語語彙大系、岩波書店 (1997)