Web上における単語の意味関係の変化を把握する手法の提案

関　良太　　岸　義樹
茨城大学大学院理工学研究科　茨城大学工学部情報工学科

1. はじめに

近年，単語の意味関係を考慮したWeb検索に関する研究が盛んである．しかし，そこで利用される意味資源は予め作成されたものであり，更新されることも少ないので，極めて静的な性質を持つ．一方でWebは刻々と変化するので，逆の性質を持つ．この性質の違いから，上記のような意味資源をWebに利用した場合，その有効性に限界があると考えられる．意味資源をWebに対して有効なものとするためには，Web上における単語間の意味関係の変化も知識として把握しておく必要がある．

そこで本研究では，単語間の意味関係の変化を，意味構造の1つである意味ネットワークの変化から把握する手法を提案して，その有効性を検証した．

2. 提案手法

我々が提案する手法は次の通りである．

1. 検索期間 $P=\{p_1,p_2,\cdots,p_l\}$ と単語の集合 $T=\{t_1,t_2,\cdots,t_n\}$ を与え，T からクエリの集合 $Q=\{q_1,q_2,\cdots,q_{n^2}\}$ を作る．
2. P の要素 $p_i(i=1,2,\cdots,l)$ において，Q の要素でWeb検索を行い，Web情報 wi_i を取得し，Web情報の集合 WI に加える．
3. WI の要素 $wi_i(i=1,2,\cdots,l)$ それぞれから，GAで意味ネットワーク $sn_i(i=1,2,\cdots,l)$ を作り，意味ネットワーク集合 SN に加える．
4. 3.で作成した SN 中の要素を比較し，T 中の要素の意味関係の変化を把握する．

2.1 検索クエリ

Q の要素 $q_{n\times(i-1)+j}(1\le i,j\le n)$ の定義は下式の通りである．式中のANDはAND検索を行うことを意味している．

$$q_{n\times(i-1)+j}=\begin{cases}t_i \text{ AND } t_j & (i\neq j)\\ t_i & (i=j)\end{cases}$$

2.2 取得するWeb上のデータ

取得するWeb情報 $wi\in WI$ について説明する．まず，Q の要素 $q_{n\times(i-1)+j}(1\le i,j\le n)$ それぞれを検索クエリにして検索したときのHit件数と上位50件分のレスポンスを，Yahoo! API([1])を用いて取得する．さらに，そこから上位5件分(1位から5位)と下位5件分(46位から50位)のURLを抽出して，それが示すWebコンテンツをダウンロードし，テキスト部分のみを抽出する．取得したHit件数を $hit_{n\times(i-1)+j}$，そしてテキスト情報を $txt_{n\times(i-1)+j}$ として，wi を以下で定義する．

$$wi=\{(hit_1,txt_1),(hit_2,txt_2),\cdots,(hit_{n^2},txt_{n^2})\}$$

2.3 GA（遺伝的アルゴリズム）

2.3.1 コード化手法

遺伝子 $g_{n\times(i-1)+j}(1\le i,j\le n)$ を以下で定義する．

$g_{n\times(i-1)+j}$ ＝単語t_iと単語t_jの関係の強さ

関係の強さは，弱い方から強い方へ順に，0から5までの6段階の重みを割り当てることで表現する．

そして個体は，この遺伝子の並びで表現される．すると単語の関係を全て表現していることになるので，個体そのものを意味ネットワークと見なすことが出来る．

2.3.2 個体評価手法

構成する意味ネットワークは，Web上での単語の関係を表現している必要がある．そのためには個体を評価する指標が，Web上の情報に基づいていなければならない([2], [3])．

個体中の各遺伝子 $g_{n\times(i-1)+j}(1\le i,j\le n)$ における単語 t_i と単語 t_j の関係について4つの指標に基づいたスコア付けを wi の要素 $(hit_{n\times(i-1)+j},txt_{n\times(i-1)+j})$ で行い，その結果から個体の適応度を計算する．

A proposal of method for understanding change of the semantic relationship between words in Web

Ryota Seki, Yoshiki Kishi
Ibaraki University
4-12-1 Nakanarusawa, Hitachi, Ibaraki, 316-8511, JAPAN

1．シンプソン係数に基づいたスコア付け; sim_k

シンプソン係数([2])は，キーワードXとYの共起の強さを表す指標である．キーワードX，YのHit件数をそれぞれ$|X|$と$|Y|$，XとYのAND検索でのHit件数の結果を$|X \cap Y|$としたとき，このスコアは次式で与える．

$$sim_k = |X \cap Y| / \min(|X|, |Y|)$$

2．初出単語距離に基づいたスコア付け; FTD_k

txt_kから名詞，動詞，形容詞を抽出し，リスト化する．距離を測りたい単語をXとYとしたとき，初出単語距離([3])はリスト中で最初に現れたXとYの距離，つまりインデックスの差で表現される．この距離をdとしたとき，このスコアは次式で与える．

$$FTD_k = 1/d$$

3．最小単語距離に基づいたスコア付け; MTD_k

使用リストは初出単語と同じである．最小単語距離([3])は，リスト中のXとYの距離の中で最小のものである．この距離をDとしたとき，このスコアは次式で与える．

$$MTD_k = 1/D$$

4．単語の出現密度に基づいたスコア付け; AD_k

リスト中の全単語のうち，目的の単語X，Yが出現する割合のことを表す．このスコアは次式で与える．

AD_k = XとYの出現回数/リストの総単語数

以上の評価に加え，意味ネットワーク中でリンクを結ぶには相応のコストが必要であると仮定し，結ぼうとするリンクの重みが大きければ大きいほど，コストがかかるように設定した．

これらの要素を考慮して，個体の適合度fを次式で定義する．

$$f = \sum_{i=1}^{n^2} \{(sim_i + FTD_i + MTD_i + AD_i) \times g_i - \text{コスト}_i\}$$

2.4 変化の把握手法

作成した意味ネットワークから，1週間ごとの各遺伝子の重みの平均をとる．その平均値が4以上であれば強い関係，1以下であれば弱い関係であるとする．その関係の変化を見ることにより，単語間の意味変化を把握する．

3 実験

3.1 内容・データについて

平成13年度版情報通信白書の文章から「ネットワークインフラ」，「高品質音声サービス」，「次世代携帯電話」，「本格サービス開始」，「ITU世界無線通信会議」，「商用サービス」，「試験サービス」，「ジェイフォングループ」，「デジタル方式」，「総務省」という10語を抽出して，関係の変化を見た例を示す．Web情報の収集期間は2009年11月15日から12月12日までである．GAの交叉は1点交叉，選択はルーレット戦略を用い，交叉率は0.8%，突然変異率は0.3%としている．コストは，0から5のリンクの重みそれぞれに対して，0.00，0.01，0.02，0.03，0.04，0.05と設定した．

3.2 結果

実験で得られた，強い関係，弱い関係を示した遺伝子の数の推移を表1に示す．

表1：強い関係，弱い関係の遺伝子数の推移

期間 / 関係	第1週	第2週	第3週	第4週
強い	9	9	7	11
弱い	1	0	0	0

また，4週を通して関係の変化が起きなかった遺伝子数は強い関係を持つもの，弱い関係を持つもの，ともに0であった．

4 考察

表1において強い関係を持つ遺伝子の数の変化や，4週間ずっと同じ関係を持ち続けた遺伝子は無いことからWeb上における単語の関係が変化していることが分かる．また，弱い関係を持つ遺伝子数が1や0であり少なかった．これは，弱い関係の基準が厳しく，本来ならば該当すべきものまで除いてしまっていることが考えられる．

5 まとめ

本研究ではWeb上における単語間の関係の変化を把握する手法を提案して，実際に実験した．その結果，本手法で関係の変化を把握できることが分かった．

参考文献

[1] Yahoo! デベロッパーネットワーク http://developer.yahoo.co.jp/

[2] 松尾 豊，友部 博数，橋田 浩一，中島 秀之，石塚 満：Web上の情報からの人間関係ネットワークの抽出，人工知能学会論文誌，Vol.20，No.1E，pp.46-56，2005

[3] 田 馳，手塚 太郎，小山 聡，田島 敬史，田中 克己：質問キーワードの意味的関連と近接性に着目したウェブ検索の精度改善，電子情報通信学会第17回データ工学ワークショップ(DEWS2006)論文集，5A-o1，2006