4N-1

ユーザ編集型のシステムを利用した単語同士の関連語の抽出

The extraction of related words from a system that a user can edit.

宮田 祐輝† 小瀬木 浩昭‡ 松澤 智史† 武田 正之†
Yuuki Miyata Hiroaki Ozeki Tomofumi Matsuzawa Masayuki Takeda

1. はじめに

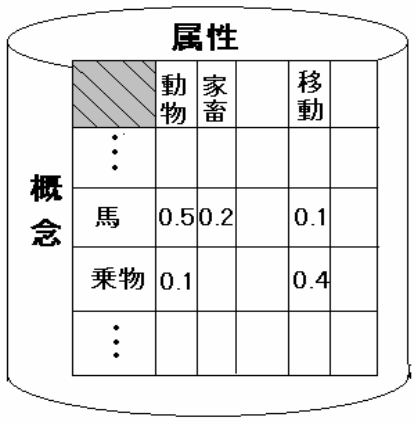
Google などの検索エンジンで検索するときに，自分で適切な単語が思いつかなかったり，AND 検索で類似語や関連している語を知りたいというのはよくある状況である．その場合，ある単語と関連性のある単語を抜きだすことのできるシステムがあれば，便利である．

現在，国語辞典または Web ページで文章を集めて，その情報からシソーラス（単語を同義語、広義語、狭義語、関連語に分類し、整理したもの）を作る試みもされているが，まだ精度が高いとはいえない．そこで，一般の利用者が単語の説明を書いていくシステムとして，はてなダイアリーから概念ベースを構築すると，どのような性能のものができるのかを検証してみることにした。この方法が有効であることを示すことができれば，利用者の書き込みのみで自動的に精度の高い関連語抽出システムを構築していくことが可能となる．

2. 概念ベースとは

概念ベースでは，各単語（以下、概念と呼ぶ）は複数の属性と属性値のペアで表現される．属性はその概念に関連した概念であり，属性値はその関連の度合いである．この場合，概念間の類似性の度合いは，属性の共通性の度合いを利用して定義すれば単純に計算することが可能となる．

それぞれの概念は下の**図1**のように，いくつかの属性でその属性とどのくらい関連しているかの値を持っている．



（**図1**） 概念ベースのイメージ

図 1 では，馬の概念は動物の属性が強く，移動の属性がやや小さいことを表していて，乗り物の概念では動物の属性が低く，移動の属性が高くなっていることが分かる．

†東京理科大学 理工学部 情報科学科，
Dept. of Information Sciences, Tokyo University of Science
‡東京理科大学大学院 理工学研究科 情報科学専攻，
Graduate School of Science and Technology,
Tokyo University of Science

3. 何を情報源（コーパス）とするか

情報源を Web ページ全体としてしまうと，Web ページ内に書かれている文章に内容の一貫性がないページがあるため，情報量が多くても精度が落ちてしまう．そのため，一般の人が情報を書くときに少し制約のある，はてなダイアリーのキーワード欄の情報を使うこととした．さらに，はてなダイアリーはユーザ主導型のシステムのため，新しく出てきた単語にも即座に対応できるメリットもある．

4. 概念ベース構築アルゴリズム

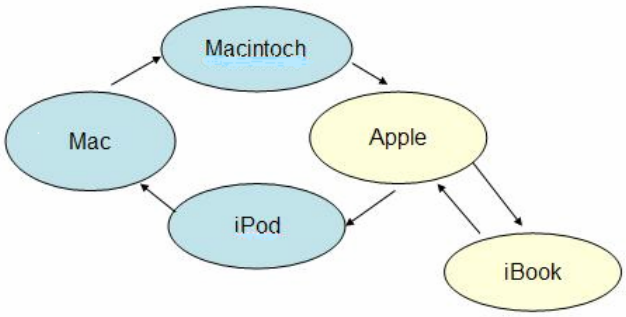
概念ベースの構築は大きく分けると， (1) 属性の抽出、(2)属性値の計算，の２つのフェーズに分けられる．

4.1. 属性の抽出

概念ベースを作るためには，まずその元となる属性が必要である．はてなダイアリーに出てくる全ての単語をそれぞれ属性としてしまう方法もあるが，そうすると「意味」や「意義」など同じような概念が存在してしまうことになる．

そこで，リンクをたどっていくとループとなるキーワード組み合わせをすべて抜き出す．下の**図2**の例では，「iPod」「Mac」「Macintosh」「Apple」と「iBook」「Apple」がそれにあたる．ただし，あまりにループする数が長いと関連度が低下すると考えられるので，小さく巡回しているものを集めていく．

さらに，このようにループとなる組み合わせの中で，共通の概念を含むグループ同士をまとめることで（**図2**参照），1つの同義語グループとする．ただし，まとめるときにそのグループがある一定以上の数を超えないように調整をする．



（**図2**） ループしている概念の組

ただし，調整の際に複数のグループに所属する概念も存在することになるが，今回はその重複を許すことにしている．そのようにしてできたグループの中でリンクされている数の一番多い概念を，グループの代表の属性名とする．

4.2. 属性値の計算方法

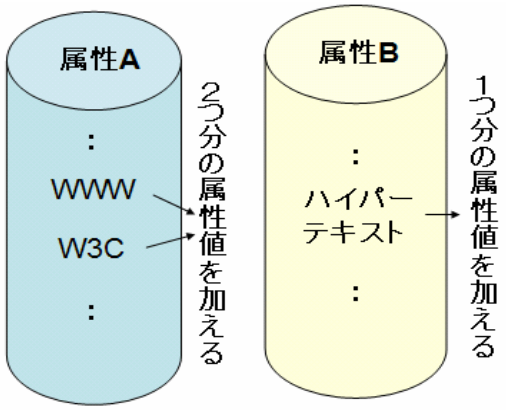
ある概念において，それぞれの属性の値を出すには，以下の3つからのアプローチを取ることにする．

① 説明文の中にでてきた属性
② 孫引きした概念の説明内の属性
③ 逆引き参照した概念の説明内の属性

この3つの中にでてきた属性をもとに，それぞれの属性値を決めていく。なお、孫引きした概念とは、直接説明文の中に出てきた概念について，さらに調べて見つけた概念のことで、逆引き参照した概念とは、その概念を説明文の中で逆に使用された概念のことである．

では，具体的なやり方を説明する．例えば，「HTML」の説明文の中に，「WWW」「ハイパーテキスト」「W3C」の3つのキーワードが入っていたとする．

ここで、属性 A に所属しているのが「WWW」「W3C」，属性 B に所属しているのが「ハイパーテキスト」であるとする．このとき，概念 HTML の属性として，A と B に値が加えられることとなる（**図 3** 参照）．



(**図 3**)属性値の基本計算手法

この操作を，孫引きした概念，逆引きした概念に対しても，それぞれ対応した属性に値を追加していくと，最終的な属性値がそれぞれ計算できる．ただし，直接出てきた単語に比べて、孫引きの概念や逆引き参照の概念は価値が違うと考えられるので，それぞれ補正をかける必要がある.すなわち，属性 i の値は以下の式で計算できる．

$$q_i = \alpha K_i + \beta \sqrt{K_i'} + \gamma K_i^R$$

ここで,q_i は i 番目の属性の値を表す．また，K_i は①で計算した値，K_i' は②で計算した値，K_i^R は③で計算した値である．α 、β 、γ は実験的に決定する補正値(>0)である．

5. 類似度の計算方法

4章で作成した概念ベースを基にして類似度を計算していく．概念のそれぞれの属性をベクトルの要素とみなすと，2つの概念A, Bの類似度はAとBの属性ベクトルの方向が近ければ関連度が高いと判別できる．つまり，内積から cos の値を計算することになる．

Aのそれぞれの属性の値を、$q_i(1 \le i \le n)$ であらわし、（n は総属性数とする），比べる対象 B の属性のそれぞれの値を $r_i(1 \le i \le n)$ と定めると，

$$\cos\ \theta = \frac{\sum_{i=1}^{n} q_i r_i}{\sqrt{\sum_{i=1}^{n} q_i^{\,2}}\sqrt{\sum_{i=1}^{n} r_i^{\,2}}}$$

が関連度の値を計算する式となる．この値が1に近ければ近いほど，関連性が高いことを示すことになる．

6. 評価法

このはてなダイアリーから作成した概念ベースの評価方法としては，類語辞典を利用する方法と，一般のユーザに評価させる手法の2つのアプローチからやることにする．ある2つの概念が関連しているかどうかは，人の感覚によるところが大きいため,コンピュータのみでは正確に判断できないためである．

具体的な方法は，まず対象概念 G に対し，それに関連する概念 G_1，比較的関連している概念 G_2，全く関連していない概念 G_3 を類語辞典からそれぞれ用意しこれらを1組とし，これを N 組つくり評価データとする．ここで，$G - G_1$ 間、$G - G_2$ 間、 $G - G_3$ 間の類似度をそれぞれ r_1 、r_2 、r_3 とし、評価指数として以下を設定する．

$$F_1 = \frac{\bar{r_1} - \bar{r_3}}{1 + \sigma_1 + \sigma_3}$$

ここで，$\bar{r_1}$ ，$\bar{r_3}$, σ_1 , σ_3 は，それぞれ r_1 ，r_3 の平均，標準偏差である．この値が大きければ，関連した概念と関連していない概念の類似度の差が大きいことがいえる．

さらに r_1 と r_2 の関係としては基本的には $r_1 > r_2$ でなければならない．したがって，評価指数として以下を設定する．

$$F_2 = \frac{m}{N}$$

ここで m は，$r_1 > r_2$ の関係が成立した組の数である．この値が大きければ，2つの類似する概念が存在する場合，どちらがより類似しているかを正確に識別できていることになる．

この2つの評価指数を総合した指標を F とすると，

$$F = F_1 \times F_2$$

と定める．この値は,理想的な関連度判別システム下で1となる．

また，一般の人に対しても複数の概念組に対して人間の感覚により 0～4 の 5 段階で関連度を与え，その値と評価対象となる類似度計算法で得られた類似度との相関をとることも予定している．

参考文献

[1] 川島 貴広，石川 勉： 言葉の意味の類似性判別に関するシソーラスと概念ベースの性能評価，人口知能学会論文誌 20巻 5 号B (May. 2005).

[2] 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 (July. 1997).

[3] グエン・ベト・ハー，帆莉 譲，石川 勉，笠原 要:単語の意味の類似性判別のための大規模概念ベース，情報処理学会論文誌 Vol.43 No.10(Oct. 2002).

[4] 平川 秀樹，木村 和弘:概念体系を用いた概念抽象化手法と語義判定におけるその有効性の評価，情報処理学会論文誌. Vol.44, No.8,pp.2230-2243 (2003)