# Web 上の情報を用いた関連語のシソーラス構築について

榊　剛史[†]　松尾　豊[††]　石塚　満[†]

本論文では Web 上の情報を利用し、自動的に関連語のシソーラスを構築する手法を提案する。検索エンジンを利用し、$\chi^2$ 値による語の関連度の指標を用い、従来の Web を用いた関連度の指標の問題点を解決する。また、新しいクラスタリング手法である Newman 法を用いて語のネットワークをクラスタリングすることで、従来手法より適切に関連語を同定する。コーパスおよび既存のシソーラスから生成した関連語正解セットを用い、提案手法の効果についての検証を行う。

**キーワード**:　シソーラス、クラスタリング、共起情報、 *Web* マイニング

# Construction of related terms thesauri from the Web

Takeshi Sakaki [†] , Matsuo Yutaka [††] and Mitsuru Ishizuka [†]

This paper describes a method to costruct related terms thesauri automatically based on Web information. We utilize Web search engine to obtain word co-occurrence information and propose a new efficient similarity metrics applying $\chi^2$ value to solve problems of the existing methods. We also introduce a new method to identify related terms using word-clustering. We do word-clustering on that assocative network to identyfy related terms using latest clustering methods, "Newman method". We make evaluations and show the effectiveness of our approach using sets of related terms extracted from a corpus and a current thesaurus.

KeyWords:　*thesaurus, co-occurence, word-clustering, Web mining*

## 1　はじめに

シソーラスは、機械翻訳や情報検索のクエリー拡張、語の曖昧性の解消など、言語処理のさまざまな場面で用いられる。シソーラスは、WordNet(Miller 1990) や EDR 電子化辞書 (日本電子化辞書研究所 1996)、日本語語彙大系 (池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 1997) など、人手で長い年月をかけて作られたものがよく用いられている[1]。しかし、こういったシソーラスを作成するのは手間がかかり、また日々現れる新しい語に対応するのも大変である。一方で、シソーラスを自動的に構築する研究が以前から行われている (Crouch and Yang 1992; Grefenstette 1994)。Web ページをはじめとする大規模で多様な文書を扱うには、シソーラスを自動で構築する、もしくは既存のシソーラスを自動で追加修正する手段が有効である。

† 東京大学 情報理工学系研究科　電子情報学専攻, Graduate School of Information Science and Technology
†† 産業総合研究所 情報技術研究部門, National Institute of Advanced Indutrial Science and Technology

1 2003 年からは WordNet だけに焦点を当てた International WordNet conference も開催されている。

シソーラスの自動構築は、語の関連度の算出と、その関連度を使った関連語の同定という段階に分けられる (Curran and Moens 2002)。2 語の関連度は、コーパス中の共起頻度を用いて求めることができる (Church and Hanks 1990)。これまでの研究では、コーパスとして新聞記事や学術文書が用いられることが多かった。それに対し、近年では Web をコーパスとして用いる手法が提案されている。Kilgarriff らは、Web をコーパスとして用いるための手法やそれに当たっての調査を詳細に行っている (Kilgarriff and Grefenstette 2003)。佐々木らは Web を用いた関連度の指標を提案している (佐々木靖弘, 佐藤理史, 宇津呂武仁 2005)。

Web には、新聞記事や論文といった従来からある整形された文書のみならず、日記や掲示板、ブログなど、よりユーザの日常生活に関連したテキストも数多く存在している。世界全体で 80 億ページを超える Web は、間違いなく現時点で手に入る最大のコーパスであり、今後も増え続けるだろう。Kilgarriff らが議論しているように、Web の文書が代表性を持つのかといった議論はこれからも重要になるが、Web はコーパスとしての大きな可能性を秘めていると著者らは考えている。Web をコーパスとして扱う際にひとつの重要な手段になるのが、検索エンジンである。これまでに多くの研究が検索エンジンを用いて、Web 上の文書を収集したり、Web における語の頻度情報を得ている (Turney 2001; Heylighen 2001)。しかし検索エンジンを用いる手法とコーパスを直接解析する手法には違いがあるため、従来使われてきた計算指標がそのまま有効に働くとは限らない。

本論文では、Web を対象とし、検索エンジンを用いて関連語のシソーラスを構築する手法を提案する。特に、検索エンジンを大量に使用すること、統計的な処理を行うこと、スケーラブルなクラスタリング手法を用いていることが特徴である。ただし、類義・同義語に加え、上位・下位語や連想語など、より広い意味である語に関連した語を関連語とする。

まず、２章で関連研究について述べる。そして、３章で検索エンジンを用いた関連度の指標を提案し、さらに４章では関連語ネットワークをクラスタリングする手法について紹介する。そして、５章では評価実験を行い、この手法の効果について議論を行う。

## 2　関連研究

語の関連性を自動的に得る方法は、これまでにさまざまな研究が行われている。コーパス中での語の共起情報をもとに語の関連度を測る指標として、様々なものが提案され用いられており (Church and Hanks 1990; Wettler and Rapp 1993; Sanderson and Croft 1999; Curran and Moens 2002)、それらは大きく 2 つに分けられる。１つは単語ベクトルを用いたベクトル空間手法である。これは、単語を多次元ベクトル空間の単語ベクトルで表現し、それぞれの単語ベクトルを比較することで関連度を測る手法である。ベクトル空間手法では、表１のようにベクトルの内積をもとにした計算指標が用いられている。表１において、$x_i, y_i$ はそれぞれ単語ベクトル $\vec{x}, \vec{y}$ の $i$ 番目の要素を表す。なお、overlap 係数はバイナリベクトルにしか用いることはで

表 1　類似度の計算指標

| ベクトル空間手法 | | 確率手法 | |
|---|---|---|---|
| cosine | $\frac{\vec{x}\cdot\vec{y}}{\lvert\vec{x}\rvert\lvert\vec{y}\rvert}$ | 相互情報量 | $\log\left(\frac{p(w\cap w')}{p(w)p(w')}\right)$ |
| dice | $\frac{2(\vec{x}\cdot\vec{y})}{\sum(x_i+y_i)}$ | dice | $\frac{2p(w\cap w')}{p(w\cup w')}$ |
| Jaccard | $\frac{\vec{x}\cdot\vec{y}}{\sum(x_i+y_i)}$ | Jaccard | $\frac{p(w\cap w')}{p(w\cup w')}$ |
| overlap | $\frac{\lvert\vec{x}\cap\vec{y}\rvert}{min(\lvert\vec{x}\rvert,\lvert\vec{y}\rvert)}$ | T 検定 | $\frac{p(w\cap w')-p(w')p(w)}{\sqrt{p(w')p(w)}}$ |
| Lin[2] | $\frac{\sum(x_i+y_i)}{\lvert\vec{x}\rvert+\lvert\vec{y}\rvert}$ | Lin98A[3] | $\log\left(\frac{f(w,r,w')f(*,r,*)}{f(*,r,w')f(w,r,*)}\right)$ |

きない。単語ベクトルの要素の取り方は研究によって様々であり、各文書への出現頻度を要素とするベクトルや各単語との共起頻度を要素とするベクトルなどが考えられる。ただし、独立な事象の確率は足し合わせることができないため、内積を用いる関連度では、語の出現確率を単語ベクトルの要素とすることは不適切と考えられる。

もう 1 つはコーパス中での確率を用いる確率手法である。この手法では、2 語がコーパス中で共起する確率をもとに関連度を算出している。確率手法で用いられている計算指標を表 1 に示す。表 1 において、$p(w \cap w')$ は語 $w, w'$ の共起確率を表し、$p(w \cup w')$ は語 $w, w'$ のどちらかが出現する確率を表す。また $f$ は (Lin 1998) で定義されている関数であり、$f(w, r, w')$ は語 $w, w'$ が $r$ の関係を持って出現する頻度を、$f(*, r, w')$ は語 $w'$ がいずれかの語と $r$ の関係を持って出現する頻度を表す。これらの計算指標は、ベクトル空間手法で用いられている指標を書き換えたものが多い。また、単語同士の共起確率ではなく、各単語が他の語と共起する確率の確率分布関数の類似性を用いて関連度を算出する研究も数多く行われている (Brown, Pietra, deSouza, Lai, and Mercer 1992; Baker and McCallum 1998; Slonim and Tishby 2000)。確率分布関数を用いた類似度は、確率分布類似度 (Distributional Similarity) と呼ばれる。類似した名詞は共通した動詞と共起すると仮定し、動詞との共起分布の類似性から関連度を算出している。

語の関連度が得られれば、関連度に基づいて語をクラスタリングすることで関連語が得られる。実際には、同じクラスタに分類された語同士を関連語や同義語であるとしている。語のクラスタリングには分布クラスタリング (Distributional Clustering) が用いられることが多い。分布クラスタリングとは、類似した名詞は共通した動詞と共起すると仮定し、各語の動詞との確率分布の類似度に基づいて、データを結合もしくは分割していくクラスタリング手法である (Pereira and Lee 1993; Li and Abe 1998; Dhillon 2002)。

これらコーパスから関連度を自動的に算出する手法では、コーパス内に出現する語しか扱えないという欠点がある。そのため、広範囲の語をカバーするためには、広範囲の内容をカバーするコーパスが必要となる。

2 (Lin 1998) で提案されている手法
3 (Lin 1998) で提案されている手法

近年では、より広範囲の語をカバーするために Web をコーパスとして用いることが提案されている。しかし Web 上の文書は莫大であり、直接収集し、解析するためには非常に大きな時間コストと設備コストがかかる。そのため、Web 全体での語の出現頻度や 2 語の共起頻度を獲得するためには従来のコーパスを用いたシソーラス構築とは異なる工夫が必要である。そのような工夫の一つとして Kilgarriff らは検索エンジンを用いた手法を紹介している (Kilgarriff and Grefenstette 2003)。「語 $w_a$」をクエリーとして検索エンジンを利用すると、語 $w_a$ の Web 上でのヒット件数が得られる。検索エンジンは非常に多くのページをクローリングしているため、このヒット件数を語 $w_a$ の Web 全体での出現頻度と近似できる。同様にして、「語 $w_a$ and 語 $w_b$」をクエリーとすれば、Web 上での語 $w_a$ と語 $w_b$ の共起頻度を獲得することができる。

検索エンジンから獲得できる頻度情報を用いて関連度を算出する手法としては、次のようなものがある。Heylighen は検索エンジンのヒット件数を用いた語の関連度の尺度により、語の分類や語の曖昧性解消、より優れた検索エンジンの開発の可能性を示唆している (Heylighen 2001)。Baroni や Tuerney は、類義語を同定するために、検索エンジンを用いた語の関連性の尺度を提案している (Baroni and Bisi 2004; Turney 2001)。Turney はその結果を用いることで TOEFL のシソーラスの問題で平均的な学生よりもよい得点を挙げたことを報告している。佐々木らは検索エンジンの上位ページとヒット件数を利用した専門用語集の自動構築を行っている (佐々木靖弘他 2005)。Szpektor は名詞ではなく動詞の関連度を検索エンジンを用いて定義している (Szpektor, Tanev, Dagan, and Coppola 2004)。これら検索エンジンを用いて関連度の計算を行っている研究では、条件付き確率や表 1 の確率手法で定義されているような相互情報量、Jaccard 係数が計算指標として用いられている。

# 3　検索エンジンを用いた関連性の測定

本章では、Web 上の情報を用いて語の関連度を測る手法を提案する。

## 3.1　検索エンジンのヒット件数の利用と従来手法の問題点

検索エンジンのヒット件数を用いて 2 語の関連度を計算する手法について説明する。ここでは、従来研究で用いられている相互情報量を計算指標として関連度を算出する。そして、その関連度を検証し、従来手法の問題点について述べる。

具体的な例を使って説明しよう。ここで用いられている手法は、(Baroni and Bisi 2004) のものと同一である[4]。関連度を測りたい語を、例えば「インク」「インターレーザー」「プリンタ」「印刷」「液晶」「Aquos」「テレビ」「Sharp」の 8 語とする。これらの語群は、Epson のプリンタであるインターレーザーに関する語と、Sharp の液晶 TV である Aquos に関する語であり、

4 ただし、Baroni らは検索エンジンとして Altavista を用いているが、Altavista は日本語に正式に対応していないため、検索エンジンは Google を用いた。

各語の関連度を得ることで、2 つのグループを適切に分けたいと仮定する。

表 2 に示しているのは、語群の各語に対して、検索エンジンによって得られたヒット件数である。表 3 には、語群中の 2 語を検索エンジンのクエリーとしたときのヒット件数を行列形式にしたものを示す。例えば、「インク」と「プリンタ」であれば、

“インク”　“プリンタ”

をクエリーとして検索エンジンに入力し、そのヒット件数を調べる[5]。8 語に対してこの行列を得るには、${}_8C_2 = 28$ 回のクエリーが必要となる。

Baroni らは、この 2 つの情報を使って求めた相互情報量の値が、語の関連度を示すよい指標になると述べている。相互情報量は、語 $w_a$ の出現確率を $p(w_a)$、語 $w_b$ の出現確率を $p(w_b)$、語 $w_a$ と語 $w_b$ の同時出現確率を $p(w_a \cap w_b)$ とすると、

$$\begin{aligned} MI(w_a, w_b) &= \log \frac{p(w_a \cap w_b)}{p(w_a)p(w_b)} \\ &= \log \frac{Nn(w_a, w_b)}{n(w_a)n(w_b)} \end{aligned} \tag{1}$$

と表される。ここで $n(w_a)$ は語 $w_a$ をクエリーとしたときのヒット数、$n(w_a, w_b)$ は「語 $w_a$ 語 $w_b$」をクエリーとしたときのヒット数であり、また、$N$ は検索エンジンのクロールした全ページ数である。Baroni らは $N$ を 3 億 5 千万ページとしているが、2006 年末現在では、Google は約 150 億ページ、AltaVista は約 120 億のページである。ここでは $N = 100 \times 10^8$ とした。

表 4 に相互情報量を示す。「液晶」の行に注目すると、「液晶」と関連が強いとあらかじめ想定している語は「テレビ」「Aquos」「Sharp」であるが、「プリンタ」や「インターレーザー」との相互情報量が大きく、「テレビ」や「Sharp」との値は小さくなっており、適切な関連度が算出されていない。

この原因は、相互情報量が「出現確率の影響を受ける」という特徴を持つためである。この特徴は式 (1) を次式のように書き換えるとわかりやすい。

$$MI(w_a, w_b) = \log p(w_a|w_b) - \log p(w_a) \tag{2}$$

$p(w_a|w_b)$ は語 $w_b$ が出現するときに語 $w_a$ と語 $w_b$ が共起する条件付き確率を表す。$p(w_a|w_b)$ が等しい場合は、$p(w_a)$ の出現確率が小さいほど相互情報量は大きい値になる。この特徴自体は「共起する確率が同じなら、出現確率の低い語と共起する方が関連性が強い」と考えられるので、問題がない。しかし、検索エンジンにおいては語によって出現頻度に大きなばらつきがあり、また全事象を表す $N$ が非常に大きいために出現確率の違いによる影響が大きくなり過ぎて

5 ダブルクオーテーションで囲んでいるのは、2 単語以上からなるフレーズに対しても適切に処理するためである。

表 2　語単独でのヒット件数

| プリンタ | 印刷 | インターレーザー | インク | 液晶 | テレビ | Aquos | Sharp |
|---|---|---|---|---|---|---|---|
| 17000000 | 103000000 | 215 | 18900000 | 69100000 | 192000000 | 2510000 | 186000000 |

表 3　2語でのヒット件数の行列

| 語/語 | プリンタ | 印刷 | インターレーザー | インク | 液晶 | テレビ | Aquos | Sharp | 合計 |
|---|---|---|---|---|---|---|---|---|---|
| プリンタ | 0 | 4780000 | 273 | 4720000 | 4820000 | 5090000 | 201000 | 990000 | 20601273 |
| 印刷 | 4780000 | 0 | 183 | 4800000 | 6520000 | 11200000 | 86400 | 1390000 | 28776583 |
| インターレーザー | 273 | 183 | 0 | 116 | 176 | 91 | 0 | 0 | 839 |
| インク | 4720000 | 4800000 | 116 | 0 | 3230000 | 4950000 | 144000 | 656000 | 18500116 |
| 液晶 | 4820000 | 6520000 | 176 | 3230000 | 0 | 18400000 | 903000 | 4880000 | 38753176 |
| テレビ | 5090000 | 11200000 | 91 | 4950000 | 18400000 | 0 | 840000 | 2830000 | 43310091 |
| Aquos | 201000 | 86400 | 0 | 144000 | 903000 | 840000 | 0 | 1790000 | 3964400 |
| Sharp | 990000 | 1390000 | 0 | 656000 | 4880000 | 2830000 | 1790000 | 0 | 12536000 |
| 合計 | 20601273 | 28776583 | 839 | 18500116 | 38753176 | 43310091 | 3964400 | 12536000 | 166442478 |

しまう。例えば、「テレビ」のように出現確率の極端に大きい語と他の語の相互情報量が小さくなる。表 4 の「テレビ」の列に注目すると、いずれの語においても「テレビ」との相互情報量が小さくなっていることが分かる。実際に表 2 の語のヒット件数と表 4 の各行との相関係数 (式 3) は −0.35 となり、相互情報量と語の出現確率にやや強い負の相関があることが分かる。それに対し、表 3 の共起ヒット件数と表 4 の相互情報量のとの相関係数は 0.06 となり、ほとんど相関がないことが分かる。

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \ (\bar{x} : x_i\text{の平均値}) \tag{3}$$

このように、従来用いられてきた相互情報量は語の出現確率に影響を受けるため、関連度を測る際に各語の出現確率に数千倍、数万倍といった開きがある場合、値の信頼性は低くなるという問題がある. これは、Jaccard 係数や dice 係数など他の類似度の指標についても当てはまる。

## 3.2　$\chi^2$ 値を用いた関連度の指標

本論文では、$\chi^2$ 値を使った関連度の指標を用いる。$\chi^2$ 値は、あるデータ集合内での統計的な偏りを表す指標であり、機械翻訳やコロケーション処理など、多くの手法で用いられている。語の関連度としては Curran らが用いている (Curran and Moens 2002)。

$\chi^2$ 値を関連度に用いるのは、語の出現頻度のばらつきによる影響を排除するためである。相互情報量や Jaccard 係数を関連度に用いる場合の問題点は、語の出現確率に大きな影響を受ける点である。この問題の解決策として、出現確率を適切に正規化するというアプローチが考え

表 4　相互情報量行列

| 語/語 | プリンタ | 印刷 | インターレーザー | インク | 液晶 | テレビ | Aquos | Sharp |
|---|---|---|---|---|---|---|---|---|
| プリンタ | 0 | 4.195 | 7.504 | 5.878 | 4.602 | 3.635 | 4.740 | 2.029 |
| 印刷 | 4.195 | 0 | 5.302 | 4.093 | 3.103 | 2.622 | 2.094 | 0.567 |
| インターレーザー | 7.504 | 5.302 | 0 | 6.542 | 5.663 | 3.981 | 0.000 | 0.000 |
| インク | 5.878 | 4.093 | 6.542 | 0 | 4.096 | 3.501 | 4.301 | 1.512 |
| 液晶 | 4.602 | 3.103 | 5.663 | 4.096 | 0 | 3.518 | 4.840 | 2.222 |
| テレビ | 3.635 | 2.622 | 3.981 | 3.501 | 3.518 | 0 | 3.746 | 0.655 |
| Aquos | 4.740 | 2.094 | 0.000 | 4.301 | 4.840 | 3.746 | 0 | 4.534 |
| Sharp | 2.029 | 0.567 | 0.000 | 1.512 | 2.222 | 0.655 | 4.534 | 0 |

られる。$\chi^2$ 値では、語群を構成する語の出現頻度を正規化要素とし、値の正規化を行ったうえで、共起の偏りを算出するので、出現確率のばらつきによる影響を抑えることができる (Yang and Pedersen 1997)。このため、値のばらつきが大きい検索エンジンのヒット件数を用いて関連度を算出する場合、$\chi^2$ 値を計算指標として用いることが適切であると考えられる。

対象とする語群の中で、共起の偏りを統計的に調べるために、1 つ 1 つの語について、語群内の他の語との共起頻度を標本値とし、「$w_i, w_j \in G$ が共起する確率は、語 $w_i$ と語群 $G$ 内の語が共起する確率と等しい」という帰無仮説をおいて検定を行う。語 $w_i$ と語 $w_j$ の実際の共起頻度を $n(w_i, w_j)$、語 $w_i$ と語群 G の語との共起頻度の和を $S_{w_i} = \sum_k n(w_i, w_k)$、全ての共起頻度の和を $S_G = \sum_{w_i \in G} S_{w_i}$ とするとき、語 $w_i$ と語 $w_j$ に関する $\chi^2$ 値は次式で表される。

$$\begin{aligned} \chi^2(w_i, w_j) &= \frac{n(w_i, w_j) - E(w_i, w_j)}{E(w_i, w_j)} \\ E(w_i, w_j) &= S_{w_i} \times \frac{S_{w_j}}{S_G} \end{aligned} \tag{4}$$

$E(w_i, w_j)$ は語 $w_i$,$w_j$ の共起頻度の期待値を表している。例えば、語 $w_i$ を「プリンタ」、語 $w_j$ を「インターレーザー」とすると、$n(w_i, w_j)$ は 273、$S_{w_i} = 20601273$、$S_{w_j}/S_G = 839/166552478$ となる。表 5 は、表 3 から計算された $\chi^2$ 値行列である。表 5 では、「プリンタ」は「印刷」や「インク」と偏って共起している。また、「インターレーザー」は「プリンタ」との共起が、「Aquos」は「Sharp」との共起が強いなど、良好な結果となっている。

また、表 6 のような、「プリンタ」「液晶」との関連が低いと考えられる 4 語と「プリンタ」「液晶」の 2 語で構成される計 6 語の語群を与えた場合を考える。この語群では、表 2 の語群と違い、「プリンタ」と「液晶」の関連性が強いと考えられる。「プリンタ」の行に注目すると、確かに「プリンタ」と「液晶」の $\chi^2$ 値が大きくなっており、語群に基づいた適切な結果が得られている。

表 5　$\chi^2$ 行列

| 語/語 | プリンタ | 印刷 | インターレーザー | インク | 液晶 | テレビ | Aquos | Sharp |
|---|---|---|---|---|---|---|---|---|
| プリンタ | 0.000 | 416649 | 275.5 | 2579092 | 113.8 | 0.000 | 0.000 | 0.000 |
| 印刷 | 416649 | 0.000 | 9.925 | 801848 | 0.000 | 1840173 | 0.000 | 0.000 |
| インターレーザー | 275.5 | 9.925 | 0.000 | 5.548 | 0.000 | 0.000 | 0.000 | 0.000 |
| インク | 2579092 | 801848 | 5.548 | 0.000 | 0.000 | 3846 | 0.000 | 0.000 |
| 液晶 | 113.8 | 0.000 | 0.000 | 0.000 | 0.000 | 6858012 | 0.000 | 1317796 |
| テレビ | 0.000 | 1840173 | 0.000 | 3846 | 6858012 | 0.000 | 0.000 | 0.000 |
| Aquos | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 7449430 |
| Sharp | 0.000 | 0.000 | 0.000 | 0.000 | 1317796 | 0.000 | 7449430 | 0.000 |

表 6　$\chi^2$ 行列-2

| 語/語 | プリンタ | 小説 | 液晶 | 紅茶 | バイオリン | 化粧品 |
|---|---|---|---|---|---|---|
| プリンタ | 0.000 | 0.000 | 2402760 | 0.000 | 0.000 | 0.000 |
| 小説 | 0.000 | 0.000 | 0.000 | 277513 | 712208 | 19024 |
| 液晶 | 2402760 | 0.000 | 0.000 | 0.000 | 0.000 | 116983 |
| 紅茶 | 0.000 | 277513 | 0.000 | 0.000 | 11149 | 597032 |
| バイオリン | 0.000 | 712208 | 0.000 | 11149 | 0.000 | 0.000 |
| 化粧品 | 0.000 | 19024 | 116983 | 597032 | 0.000 | 0.000 |

# 4　関連度を用いたネットワークに基づくクラスタリング

従来は、確率分布の類似度に基づいた分布クラスタリングの方法を用いて、関連語をクラスタに分けることが多かった。本研究では、語の関連度からネットワークを構築し、ネットワークに基づく新しいクラスタリングの方法を適用する。関連語ネットワーク上で Newman 法によりクラスタリングを行い、その結果、同じクラスタに分類されたもの同士を関連語として取り出す。このクラスタリング法は、語の数が大規模になったときにでも適用でき、対象によってはよいクラスタを生成するので近年注目を集めている。

## 4.1　関連語ネットワークの構築

まず、語の関連性を用いて、語のネットワークを構築する。ノードが語、エッジが強い関連を表す。本論文では、これを関連語ネットワークと呼ぶ。

関連語ネットワークは次のように構成される。

(1)　語群 $G$ を与える。

(2)　次式により 2 語 $w_i$、$w_j \in G$ の関連度 $\chi^2_{w_i,w_j}$ を計算する。

$$\chi^2_{w_i,w_j} = \frac{n(w_i,w_j) - E(w_i,w_j)}{E(w_i,w_j)}$$

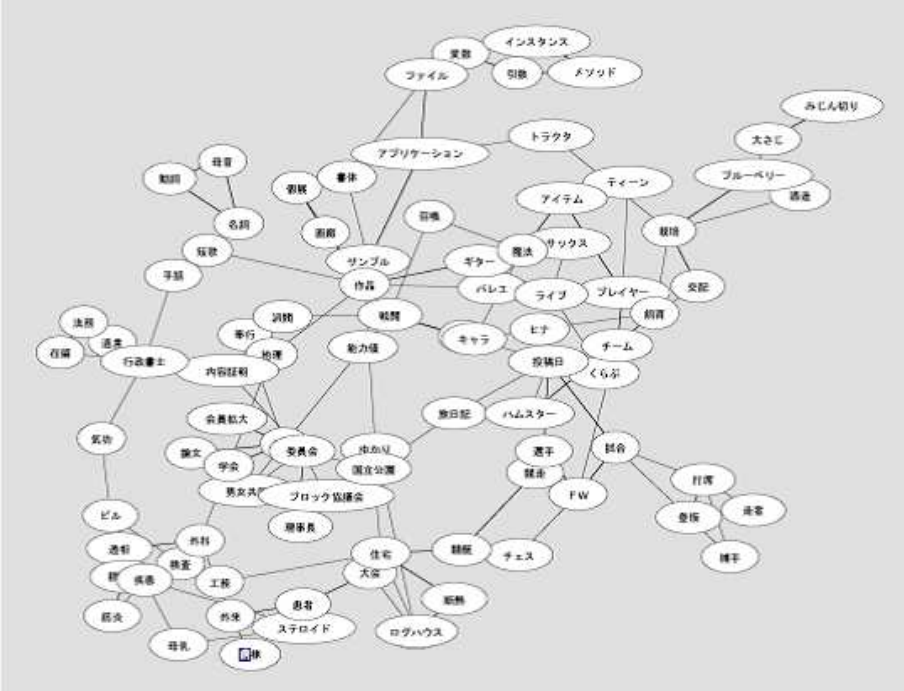


図 1　関連語ネットワーク

$$
\begin{aligned}
E(w_i, w_j) &= S_{w_i} \times \frac{S_{w_j}}{S_G} \\
S_{w_i} &= \sum_k n(w_i, w_k) \\
S_G &= \sum_{w_i \in G} S_{w_i}
\end{aligned}
\tag{5}
$$

(3)　　各語 $w_i \in G$ をノードとして配置する。

(4)　　$\chi^2_{w_i,w_j} > 0$ のとき、ノード $w_i$、$w_j$ 間にエッジを張る。

例を図 1 に示す。これは、Web から獲得したコーパス中に高頻度に出現する計 90 語をこのネットワークの構成語として用い、ヒット件数を得る検索エンジンとして Google を用いた関連語ネットワークである。この関連語ネットワーク上では、関連の強い語同士が近く配置されている。例えば、図 1 の左下には「疾患」「患者」などの医学関連の語が密集している。また、上部では「アプリケーション」「ファイル」などのコンピュータ関連の語が密集している。このように関連語ネットワーク上では、関連の強い語同士が密集して存在している。

表 7　階層的クラスタリングで用いられる距離関数 $D(c_i, c_j)$

| 手法 | 最大距離法 | 最小距離法 | 群平均法 |
|---|---|---|---|
| $D(c_i, c_j)$ | $\max_{w_k \in c_i, w_l \in c_j} Sim(w_k, w_l)$ | $\min_{w_k \in c_i, w_l \in c_j} Sim(w_k, w_l)$ | $\frac{1}{n_i n_j} \sum_{w_k \in c_i} \sum_{w_l \in c_j} Sim(w_k, w_l)$ |

## 4.2　ネットワークに基づくクラスタリング

従来のシソーラス構築における語のクラスタリングには確率分布を用いた分布クラスタリング手法が一般的に用いられている。(Pereira and Lee 1993; Dhillon 2002)。また情報検索の分野では、語を属性とする高次元のベクトルを用いた語のクラスタリング手法も多く、LSA や random projection といった次元を圧縮する手法も有効である (Deerwester, Dumais, Landauer, Furnas, and Harshman 1990; Papadimitriou, Tamaki, Raghavan, and Vempala 1998)。

一方で、近年ではデータをネットワークとして表した上で、それを分析する手法が提案され、着目を集めており、語の関係性の分析にも用いられている (Widdows and Dorow 2002; Motter, Moura, Lai, and Dasgupta 2002; Palla, Derényi, Farkas, and Vicsek 2005)。Sigman は WordNet がネットワーク構造としての性質を持っていることを示し、WordNet にネットワーク分析の手法を適用できることを示している (Sigman and Cecchi 2002)。

ネットワークのクラスタリングには、従来、表 7 のように距離関数 $D(c_i, c_j)$ を定義し ($n_i$ はクラスタ $c_i$ に含まれる語の数、$Sim(w_k, w_l)$ は語 $w_k, w_l$ の類似度を表す)、距離の近い順に各クラスタをマージしていく階層的クラスタリング手法や、EM アルゴリズム、NaiveBayes といった機械学習の手法を用いたクラスタリング手法が一般的に用いられてきた。しかし、ここ数年で新たなクラスタリング手法がいくつも提案されている。代表例としては、betweenness クラスタリングがあげられる。betweenness クラスタリングは、グラフ[6]の betweenness というエッジの媒介性を表す指標 (あるエッジが他のエッジの最短パスにどの程度の割合で含まれているか) に注目し、できるだけ部分グラフをつなぐような betweenness の高いエッジを削除していくことにより、密度の濃いサブグラフを同定する手法である (Girvan and Newman 2002)。

これらの手法は高次元のベクトルに対しても有効であり、以前の手法と比べて高い精度で現実のクラスタ構造を再現することができる。その反面、時間計算量が大きく、大規模なネットワークに適用することは難しい。例えば、ネットワークのノード数を $n$、エッジ数を $m$ とするとき、betweenness クラスタリングの時間計算量は $O(n^3)$ または $O(m^2 n)$ であり、ノード数が多いネットワーク上で betweenness クラスタリングを行うことは困難である。そこで、本研究では大規模なネットワークにも適用可能なクラスタリング手法である Newman 法を用いる。

6 ネットワークは、エッジに重みや長さなどの数値が付加されているのに対し、グラフはエッジに数値の付加されていない、接続関係だけを表すものである。

Newman 法は、階層的クラスタリング手法の一つであるが、クラスタリングを評価関数 $Q$ の最大値導出問題に置き換えた手法である (Newman 2004)。評価関数 $Q$ とは、各クラスタの結合度を表す関数であり、$Q$ が大きいほど各クラスタ内の結合が強いことを表している。Newman 法では、$Q$ の高い状態がより適切にクラスタリングされた状態であると定義している。そして、$Q$ の最大値を求めることで、そのネットワークに最適なクラスタリング結果を得ることを目標としている。

評価関数 $Q$ は次式で表される。

$$Q = \frac{1}{2m}\left[\left(\sum_{v,w} A_{vw}\delta(c_v, c_w)\right) - \left(\sum_{v,w} \frac{k_v k_w}{2m}\delta(c_v, c_w)\right)\right] \tag{6}$$

$k_v$ は頂点 v が持っているエッジの本数、$m$ は全エッジ本数の合計、$c_v$ は頂点 $v$ が属しているクラスタを表している。$\delta(c_v, c_w)$ はクロネッカーの $\delta$ である。式 (6) の第 1 項において、$A_{vw}$ は頂点 $v, w$ 間のエッジの有無を表しており、また頂点 $v, w$ が同じクラスタのときのみ、$\delta(c_v, c_w) = 1$ となる。つまり、第 1 項は各クラスタ内に含まれるエッジの本数の合計を表している。同様に第 2 項においては、$\frac{k_v k_w}{2m}$ は頂点 $v, w$ 間にエッジが引かれる確率を表しているため、第 2 項は、各クラスタ内に含まれるエッジの本数の合計の期待値を表している。

すなわち、評価関数 $Q$ とは、クラスター内に存在するエッジの本数の合計が期待値からどの程度ずれているかを相対的に表した値である。クラスター内のエッジ本数の和が期待値と同じなら $Q = 0$、それより強いクラスターなら $Q > 0$ であり、弱いクラスターなら $Q < 0$ となる。$Q$ が最大であるとき、各クラスター内での結合度が最大であるので、ネットワーク全体として最も良くクラスタリングされた状態であると考えられる。

しかし $Q$ の最大値を求める場合、エッジ数 $m$、ノード数 $n$ のとき、計算量が $O(n^3)$ もしくは $O(m^2 n)$ となり、大きくなってしまう。そこで Newman 法では Greedy アルゴリズムを用いて $Q$ の値が極大値をとるようにクラスタリングを行う。Greedy アルゴリズムなので、「$Q$ の変化量 $\Delta Q$ が最大になるようにクラスタ、もしくはノードをマージする」という手順を繰り返していく。そして「$\Delta Q$ の最大値 $< 0$」となった時点でクラスタリングを終了とする。このようにして $Q$ の極大値を求めている。この際、常に「$\Delta Q$ が最大になるような 2 つのクラスタを選んでマージ」するため、クラスタがマージされていく順序は一意であり、初期条件によってクラスタリングの結果は変化しない。また、クラスタ数を任意に制御したい場合は、終了条件を $\Delta Q < 0$ ではなくクラスタ数にすることも可能である。

Newman 法と betweenness クラスタリングを比較すると、Newman らにより Newman 法は betweenness クラスタリングとほぼ同じ精度のクラスタリング結果が得られることが示されている。また、Newman 法の時間計算量は $O((m+n)n)$ もしくは $O(n^2)$ であり、時間計算量が $O(m^2 n)$ あるいは $O(n^3)$ である betweenness クラスタリングと比べ、計算量が少なく、高速な手法となっている。そのため、Newman 法はノード数やエッジ数が大きい大規模ネットワーク

に適用可能である。

### 4.3 Newman 法による関連語の獲得

語群 $G$ を用いてシソーラスを構築する場合、Newman 法を用いて関連語を同定する手順は次のようになる。

(1) 検索エンジンのヒット件数と $\chi^2$ 値を用いて語群 $G$ の語の関連度を算出する。
(2) 関連度をもとに語群 $G$ を構成語とする関連語ネットワークを構築する。
(3) 1 つの語を 1 つのクラスタとする。
(4) ある 2 つのクラスタが 1 つのクラスタになったと仮定して、$Q$ の変化量 $\Delta Q$(式 7) を計算する。
(5) (4) を全てのクラスタの組み合わせについて行う。
(6) $\Delta Q$ が最大となるような 2 つのクラスタをマージし、1 つのクラスタとする。ただし、最大の $\Delta Q < 0$ なら (8) へ。
(7) マージしたクラスタの $e_{ij}, a_i$ を再計算し、(4) に戻る。
(8) 同じクラスタに属している語を関連語とみなす。

$$\begin{aligned} \Delta Q_{ij} &= 2(e_{ij} - a_i a_j) \\ e_{ij} &= \text{クラスタ } i, j \text{ 間のエッジの本数 (割合)} \\ a_i &= \sum_i e_{ii} \end{aligned} \tag{7}$$

## 5 評価

### 5.1 評価実験の概要と正解セットの作成

シソーラスを評価する手法として、WordNet や EDR など人手で構築された既存のシソーラスと比較する方法 (Jarmasz and Szpakowicz 2003; Curran 2002)、綿密に作られたアンケートや語の分類タスクを人が行い、その結果と比較することでシソーラスの適切さを評価する方法 (Sanderson and Croft 1999; Hodge and Austin 2002) がある。前者の手法は WordNet に出現する語しか評価できないため語の範囲が限られてしまい、後者はコストがかかるのが問題である。

本研究では、提案手法で構築されたシソーラスを、2 種類のシソーラスを用いて提案手法の評価を行う。一つ目は Web より収集したコーパスから作成したシソーラスであり、これを関連語の正解セット作成用のデータとして用いることで提案手法と従来のコーパスを用いた手法との比較を行う。二つ目は既存のシソーラスであり、これから作成した関連語の正解セットを用いて、人手によって構築されたシソーラスと提案手法との比較を行う。

また、1 つ目の正解セットには Web に特徴的な語が多く含まれるのに対し、2 つ目の正解

セットでは、既存のシソーラスに含まれるような、いわゆる汎用的な語が多く含まれる。そのため、それぞれの正解セットを評価実験に用いることで、Web に特徴的な語に対する提案手法の有効性、汎用的な語に対する提案手法の有効性を検証することにもなる。

OpenDirectory を用いた正解セットの作成

シソーラスを作成するコーパスとして OpenDirectory[7]を用い、あらかじめ各カテゴリに特徴的な語を抽出することで、正解となるシソーラスを模擬的に作成する。OpenDirectory は、ボランティア方式で運営される世界最大のウェブディレクトリであり、各カテゴリは、担当のエディタによって管理されている。Web ディレクトリの中では、カテゴリ分類の信頼性が高いもののひとつである。各カテゴリに特徴的に出現する語は互いに関連しているという仮定のもとで、提案手法および比較手法による語の関連性の適切さを評価する。

OpenDirectory の 14 個のカテゴリの中から、「アート」、「スポーツ」、「コンピュータ」、「ゲーム」、「社会」、「家族」、「科学」、「健康」、「レクリエーション」の 9 つのカテゴリを用いた[8]。各カテゴリ内に含まれる Web ページを用い、次のようにカテゴリに特徴的な語を抽出する。

(1) 各カテゴリ $C_i(i = 1...9)$ ごとに登録順に 1000 ページの文書を取得する。

(2) 全ての文書に形態素解析[9]を行う。そして連接する名詞 5-gram までを単語として取り出す (Manning and Schutze 1999)。

(3) カテゴリ $C_i$ 内で、単語 $w_a$ が含まれる文書の数を $f^i_{w_a}$ とする。また、全てのカテゴリで語 $w_a$ が含まれる文書数を $f^{all}_{w_a}$ とする。

(4) カテゴリ $C_i$ における語 $w_a$ の重みを次のように計算する。

$$score^i_{w_a} = f^i_{w_a} \times \log(N/f^{all}_{w_a}) \tag{8}$$

ただし、$N$ は全文書数である。

(5) カテゴリ $C_i$ ごとに score の高い語 $w_a$ を取り出し、それらをそのカテゴリに特徴的な語群 $R_{C_i}$ とする。すなわち、$R_{C_i} = \{w_k | rank_i(w_k) \leq 10\}$ である ($rank_i(w_k)$ は、カテゴリ $C_i$内での語 $w_k$の $score$ の順位を表す)。また、$A = \{w | w \in R_{C_i}, i = 1...9\}$ とする。

ここでは、各カテゴリごとに特徴的に現れる語を、tfidf の考え方を用いて重み付けしている。また上記説明の (1) において「登録順に」とあるが、これは OpenDirectory のサイトから文書データを収集する際に、データが得られる順番を意味している。この順番は、文書の内容に関係なく無作為に並んでおり、特定のルールはないと考えられるため、ランダムな順番と考えても問題ないと言える。

7 http://dmoz.org/World/Japanese/

8 なお、「ニュース」、「キッズティーンズ」、「ビジネス」、「オンラインショップ」、「各種資料」は、他のカテゴリとの重複が大きいため除いた。

9 茶筌。http://chasen.aist-nara.ac.jp/.

表 8　OpenDirectory から取り出した関連語群

| カテゴリー | 関連語群 |
|---|---|
| アート | 画廊, 作品, 劇場, サックス, 短歌, ライブ, ギター, 披露, バレエ, 個展 |
| コンピュータ | 掲示板, ソースコード, 無料レンタル, アクセス数, 文字コード, 初期値, 拡張子 |
| 科学 | 情報処理, 実証, 方法論, 社会科学, 研究対象, 格差, 研究員, 専門, 専攻, 討論 |

得られた語の一部を表8に示す。例えば「アート」カテゴリから取り出された語に注目すれば、「画廊」「作品」「個展」は絵画関連の語、「サックス」「ライブ」「ギター」は音楽関連の語、「バレエ」「披露」「劇場」はパフォーミングアート関連の語、「短歌」は文芸関連の語となっており、いずれも「アート」に関連した語が取り出されている。こうして得られたカテゴリごとの特徴的な語を用いて、

- ある 2 語が同一カテゴリ内に含まれれば、関連している
- ある 2 語が異なるカテゴリであれば、関連していない

と見なす。

ここでの評価法は、カテゴリごとの特徴語の抽出に基づいている。各カテゴリに特徴的に現れる語を重み付けする方法は、(長尾, 水谷, 池田 1976) や (Xu, Kurz, Piskorski, and Schmeier 2002) で用いられている。後者では、各カテゴリに特徴的な語を tfidf で重み付けし、tfidf 値の高い語をカテゴリに特徴的な語として抽出している。さらに (Chang 2005) では、OpenDirectory のカテゴリ分類を用いて各カテゴリに特徴的な語を取得し、その結果、人手による評価で平均 65%、最大で 81%の正解率を得ている。もちろん、ここでの関連語の正解セットは完全ではなく、異なるカテゴリに含まれていても関連している場合もあるかもしれないし、同一カテゴリ内であっても、その関連の度合いは程度の差が大きいかもしれない。しかし、本研究では、このデータを手法の比較を行うための目安として用いており、比較手法の優劣を示すには十分であると考えている。

図 2 に全体の概要を図示する。OpenDirectory から獲得したカテゴリ分類されたコーパスを用いて関連語の正解セットを作成する。その正解セットの語を用いて提案手法および比較手法によって関連語を出力する。その際、比較手法はコーパス内の共起情報を用いて関連度の算出を行う。そして出力結果と正解セットを比較し、手法の評価を行う。

図 2 に示すとおり、本評価実験では正解セット作成用コーパス、比較手法で用いる関連度学習用コーパスの 2 種類のコーパスが必要となる。そこで、全部で各カテゴリから 5000 ページずつ計 4 万 5 千ページの文書をコーパスとして用意し、1/5 を正解セット作成用に、4/5 を関連度の学習用に用いて 5 分割交差検定を行った[10]。正解セット作成用のコーパスを変えたそれぞれの正解セットを $Ao_i(i = 1, 2, 3, 4, 5)$ とする。

10 ただし、関連度の学習を行う際はコーパスの持つカテゴリ分類は無視し、flat なコーパスとして扱った。

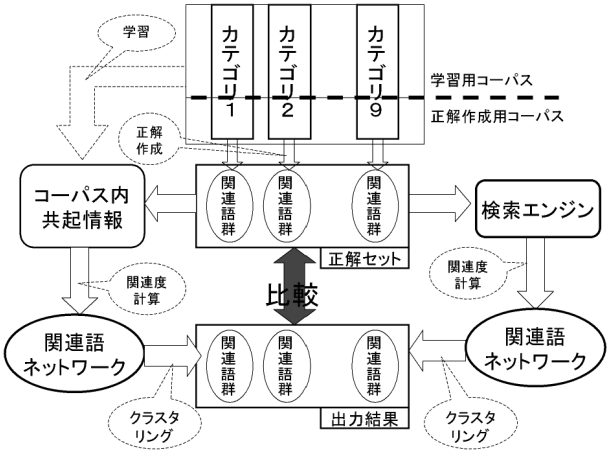


図 2　評価実験の概略図

表 9　評価実験の例 (ヴァイオリン)

| | 関連語 | 適合率 | 再現率 | InvR |
|---|---|---|---|---|
| 正解セット | ビオラ, チェロ, 笛, ギター | | | |
| 手法 1 | 1 位:ビオラ 2 位:チェロ 3 位:ビール 4 位:ピック | 0.5 | 0.5 | 1.50 |
| 手法 2 | 1 位:ピック 2 位:ビール 3 位:ビオラ 4 位:チェロ 5 位:ギター | 0.6 | 0.75 | 0.78 |

関連度の評価は、適合率、再現率、Inverse Rank Score によって測る。Inverse Rank Score とは正解とマッチした語の順位の逆数の合計値であり、正解となる語が上位にランクされる程大きい値となる。この値を用いることで、順位を考慮した比較を行うことができる。

簡単な算出例を表 9 に示す。この場合、手法 1 による出力は 4 語中 2 語が正解であるので適合率 $= \frac{2}{4} = 0.50$、正解セット 4 語のうち 2 語が手法 1 により出力に含まれているので、再現率 $= \frac{2}{4} = 0.50$ となる。同様に手法 2 では 適合率 $= \frac{3}{5} = 0.60$、再現率 $= \frac{3}{4} = 0.75$ となり、手法 2 の方が優位となる。しかし、正解の語が上位にランクされている手法 1 の方が手法としての実用性が高い、とも考えられる。このような場合に各手法の Inverse Rank Score を求めると手法 1 では、$\frac{1}{1} + \frac{1}{2} = 1.50$、手法 2 では $\frac{1}{3} + \frac{1}{4} + \frac{1}{5} = 0.78$ となり、手法 1 の方が優位となる。

このように適合率、再現率に加え、Inverse Rank Score を用いることで、順位を加味した評価を行うことができる。(Widdows and Dorow 2002; Curran 2002)。

既存シソーラスを用いた正解セットの作成

本論文では、Curran ら (Curran 2002) の手法を元にして、提案手法と既存のシソーラスを比較を行う。そのために、正解セット作成用シソーラスと比較用シソーラス、2 種類のシソーラスを用意する。まず、正解セット作成用シソーラスから関連語を取り出し、正解セットを作成する。この正解セットの語群に対して提案手法と比較用シソーラスを適用して関連語の分類を行う。その結果、どの程度正しく語群が関連語群に分類されるかによって提案手法と既存シソーラスとの比較を行う。本論文では、Curran らが用いた Roget's Thesaurus の最新版である Roget's Millenium Thesaurus(Barbara Ann Kipfer 2006) を正解セット作成用のシソーラスとして用い、WordNet 及び Moby Thesaurus(Ward 1996) を比較用のシソーラスとして用いる。

Roget's Millenium Thesaurus は見出語を持ち、その見出語がそれぞれ関連語群を持つ、という 2 層構造をしたシソーラスである。本実験においては、一つの見出語から取り出される関連語群をそのまま正解の関連語群とした。ただし、比較用シソーラスに含まれない語は除くものとする。関連語の例は表 10 のようになる。今回、見出語としては TOEIC 最頻出英単語リスト[11]に含まれる名詞の計 220 語を用いる。これらの見出語から無作為に 10 語選び、その 10 語からそれぞれ関連語群を取り出し、1 組の関連語正解セットとする。本実験では、計 10 組の正解セット $Aw_i(i = 1, 2, ..., 10)$ を作成した。

また比較用のシソーラスを用いた関連語群の分類では、算出した関連度に基づいて行うのではなく、各 2 語が比較用シソーラスで関連語とされているか、いないかの 2 値的な判定によって行うものとする[12]。この際、どの 2 語を関連語とみなすかは、シソーラスの構造によって違う方法を用いた。Roget's Thesaurus と同様に見出語と関連語群の 2 層構造を持つ Moby Thesaurus においては、見出語とその関連語群同士、及び同じ見出語を持つ語同士を関連語とみなす。木構造を持つ WordNet においては、見出語と Hyponyms(下位語)、見出語と Hypernyms(上位語) 及び見出語と Coordinate Terms(共通の上位語を持つ語) 同士を関連語とみなす。

関連度の評価指標としては、OpenDirectory を用いる場合と同様に、適合率と Inverse Rank Score を用いる。

## 5.2　関連度の指標の評価

関連度の指標に関する評価を行う。提案手法では、関連度の計算に $\chi^2$ 値を用いているが、この有効性を示すため、相互情報量、Jaccard 係数を用いた関連度と比較する。検索エンジンを利用する際、日本語のみを扱う OpenDirectory による正解セットでは、検索時のオプション

11 http://www.linkage-club.co.jp/ExamInfoData/toeic.htm

12 WordNet を用いて 2 語の関連度を算出する方法もあるが、予備実験により関連語の分類には適さないことが判明したので、本論文では採用しなかった

表 10　正解用シソーラスから取り出した関連語群

| 見出語 | 関連語群 |
|---|---|
| access | admission, contact, door, entrance, entree, ingress, introduction, open door, road, route, |
| election | acclamation, appointment, by-election, referendum polls, primary , selection, voting |
| pollution | abuse, contamination, corruption, decomposition, uncleanness dirtying, impurity, infection, rottenness, spoliation |
| agriculture | agronomy, culture, horticulture, tillage, husbandry |

表 11　OpenDirectory を用いた正解セットでの適合率

| 正解セット/指標 | cosine | 相互情報量 | Jaccard 係数 | $\chi^2$ 値 |
|---|---|---|---|---|
| セット $Ao_1$ | 0.557 | 0.447 | 0.424 | 0.567 |
| セット $Ao_2$ | 0.513 | 0.406 | 0.389 | 0.493 |
| セット $Ao_3$ | 0.519 | 0.396 | 0.376 | 0.539 |
| セット $Ao_4$ | 0.561 | 0.404 | 0.417 | 0.569 |
| セット $Ao_5$ | 0.529 | 0.421 | 0.404 | 0.519 |
| 平均 | 0.535 | 0.415 | 0.402 | 0.538 |

として「日本語のページを検索」を選択した値を用い、英語のみを扱う既存のシソーラスによる正解セットでは検索時のオプションとして「ウェブ全体から検索」を選択した値を用いる[13]。

また、コーパスを用いて学習する手法との比較も行う。コーパスを用いる手法では、tfidf 値を要素とする単語ベクトルを用い、計算指標としては cosine を用いた。実験の手順を以下に示す。

(1)　正解セット $A_i$ に含まれる全ての語について、各指標ごとに 2 語の関連度を計算する (比較用シソーラスを用いる際はこの手順は省略)。

(2)　各指標ごとに語 $w_i$ と関連度の高い上位 9 語を $A_i$ から選び、それを語 $w$ の関連語群 $G_w$ とする (比較用のシソーラスを用いる場合は、比較用シソーラスにおいて語 $w_i$ の関連語とされる語を全て取り出し、$G_w$ とする)。$G_w$ と正解セットを比較し、適合率を計算する。

(3)　(2) を語 $w_i \in A_i$ 全てについて行い、指標ごとに適合率の平均値を算出する。

(4)　(1) から (3) を正解セット $A_i(i = 1\sim n)$ について行う。

OpenDirectory から作成した正解セットの適合率の平均値を表 11 に、Inverse Rank Score の平均値を表 12 に示す。

まず、検索エンジンを用いた手法同士で比較すると、どの正解セットにおいても $\chi^2$ 値が他

13 Google ではオプションによって検索するページの対象範囲をコントロールできる

表 12　OpenDirectory を用いた正解セットでの Inverse Rank

| 正解セット/指標 | cosine | 相互情報量 | Jaccard 係数 | $\chi^2$ 値 |
|---|---|---|---|---|
| セット $Ao_1$ | 2.42 | 1.58 | 1.63 | 2.36 |
| セット $Ao_2$ | 2.41 | 1.90 | 1.43 | 2.75 |
| セット $Ao_3$ | 1.97 | 1.09 | 1.02 | 1.64 |
| セット $Ao_4$ | 2.29 | 2.04 | 1.84 | 2.52 |
| セット $Ao_5$ | 1.70 | 2.17 | 1.83 | 2.20 |
| 平均 | 2.16 | 1.76 | 1.55 | 2.29 |

の 2 つの計算指標よりもよい適合率、Inverse Rank Score を示している。これより、$\chi^2$ 値が検索エンジンを用いる手法の関連度の指標として有効であることが分かる。また、コーパスを用いて学習した手法である cosine と検索エンジンを用いた手法を比較すると Jaccard 係数、相互情報量は cosine よりも低い適合率, Inverse Rank Score である。cosine と $\chi^2$ 値を比較すると正解セットによって 2 つの評価指標の優劣が変化している。しかし、平均ではほとんど差がないことから、$\chi^2$ 値と cosine はほぼ同じ適合率であると考えられる。ただし、コーパスから学習する手法ではコーパス中に出現する語しか扱えないという欠点を持つのに対し、検索エンジンを用いる手法では Web 上に出現するほとんどの語を扱うことができる。そのため同じ適合率ならば、$\chi^2$ 値を計算指標として検索エンジンを用いる手法の方が優れていると言える。

また、表 11,12 において、5 つの正解セットにおける標準偏差 (式 9) を求める。

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(\bar{x} : x_i\text{の平均値}) \tag{9}$$

すると適合率の標準偏差は 0.014、Inverse Rank Score の標準偏差は 0.12 であり、いずれも標準偏差は 10%以内に収まっている。このことから、正解セットによるばらつきによる影響はあまり大きくないと考えられる。

次に Roget's Thesaurus から作成した正解セットを用いた既存シソーラスと提案手法の比較実験の結果を表 13 と 14 に示す。ただし、比較用シソーラスにおいては、全ての関連語が等価に扱われており順位が存在しないため、Inverse Rank Score の算出は省略する。

まず、Web を用いる手法同士を比較すると、OpenDirectory を用いた正解セットと比べて適合率の差が小さくなってはいるが、シソーラスを用いた正解セットにおいても、$\chi^2$ 値が他の計算指標よりよい数値を示している。これより、提案手法の優位性は小さくなるものの、Web 上での出現頻度にばらつきの少ない汎用的な語に対しても、提案手法が有効であることがわかる。

次に $\chi^2$ 値と既存のシソーラスを比較すると、既存のシソーラスの精度の方が若干高い数値を出してはいるものの、ほぼ同程度の精度・適合率が得られている。これより、関連語を同定するタスクにおいて、提案手法を用いることで既存シソーラスと同程度の効果が得られると言える。

表 13　既存シソーラスを用いた正解セットでの適合率

| 正解セット/指標 | シソーラス | 相互情報量 | Jaccard 係数 | $\chi^2$ 値 |
|---|---|---|---|---|
| セット $Aw_1$ | 0.385 | 0.324 | 0.374 | 0.405 |
| セット $Aw_2$ | 0.375 | 0.322 | 0.311 | 0.353 |
| セット $Aw_3$ | 0.342 | 0.365 | 0.402 | 0.411 |
| セット $Aw_4$ | 0.370 | 0.291 | 0.295 | 0.320 |
| セット $Aw_5$ | 0.438 | 0.459 | 0.487 | 0.515 |
| セット $Aw_6$ | 0.339 | 0.390 | 0.369 | 0.374 |
| セット $Aw_7$ | 0.391 | 0.287 | 0.335 | 0.345 |
| セット $Aw_8$ | 0.290 | 0.337 | 0.330 | 0.339 |
| セット $Aw_9$ | 0.468 | 0.295 | 0.279 | 0.316 |
| セット $Aw_{10}$ | 0.444 | 0.375 | 0.368 | 0.390 |
| 平均 | 0.392 | 0.345 | 0.349 | 0.369 |

表 14　既存シソーラスを用いた正解セットでの Inverse Rank

| 正解セット/指標 | 相互情報量 | Jaccard 係数 | $\chi^2$ 値 |
|---|---|---|---|
| セット $Aw_1$ | 1.260 | 1.277 | 1.441 |
| セット $Aw_2$ | 1.145 | 1.263 | 1.290 |
| セット $Aw_3$ | 1.329 | 1.390 | 1.477 |
| セット $Aw_4$ | 1.184 | 1.006 | 1.144 |
| セット $Aw_5$ | 1.526 | 1.453 | 1.572 |
| セット $Aw_6$ | 1.498 | 1.273 | 1.241 |
| セット $Aw_7$ | 1.250 | 1.183 | 1.255 |
| セット $Aw_8$ | 1.337 | 1.354 | 1.432 |
| セット $Aw_9$ | 1.033 | 1.101 | 1.298 |
| セット $Aw_{10}$ | 1.320 | 1.313 | 1.301 |
| 平均 | 1.288 | 1.255 | 1.3321 |

以上より、提案手法を用いることで、検索エンジンを用いた既存手法やコーパスから学習する手法よりも適切に関連度を算出することができていると考えられる。ただし、コーパスから学習する手法では cosine 以外の計算指標を用いた手法があるため、今後それらの指標とも比較する必要がある。

## 5.3　クラスタリングの評価

次に、クラスタリングの評価を行う。提案手法では Newton 法を用いているが、比較手法としては、群平均法を距離関数とする階層的クラスタリングを用いる。

クラスタリング手法の評価手法を以下に示す。

(1)　正解セット $A_i$ に含まれる全ての語について、2 語の関連度を計算する。

(2)　関連度をもとに関連語ネットワークを構築する。その際、ネットワークの密度が 0.3

表 15　関連語抽出実験結果（上段：適合率　中段：再現率 下段：F 値）　OpenDirectory 使用

| クラスタリング | | cosine | 相互情報量 | Jaccard 係数 | $\chi^2$ 値 |
|---|---|---|---|---|---|
| 群平均法 | 適合率 | 0.772 | 0.864 | 0.848 | 0.812 |
| | 再現率 | 0.209 | 0.222 | 0.208 | 0.221 |
| | F 値 | 0.328 | 0.353 | 0.333 | 0.347 |
| Newman 法 | 適合率 | 0.815 | 0.792 | 0.797 | 0.738 |
| | 再現率 | 0.344 | 0.332 | 0.346 | 0.631 |
| | F 値 | 0.483 | 0.465 | 0.482 | 0.680 |

表 16　関連語抽出実験結果（上段：適合率　中段：再現率 下段：F 値）　Roget's Thesaurus 使用

| クラスタリング | | 相互情報量 | Jaccard 係数 | $\chi^2$ 値 |
|---|---|---|---|---|
| 群平均法 | 適合率 | 0.887 | 0.861 | 0.852 |
| | 再現率 | 0.174 | 0.186 | 0.184 |
| | F 値 | 0.291 | 0.305 | 0.302 |
| Newman 法 | 適合率 | 0.688 | 0.705 | 0.598 |
| | 再現率 | 0.329 | 0.302 | 0.411 |
| | F 値 | 0.440 | 0.419 | 0.485 |

[14]になるように関連度の低いエッジを切る。ネットワークの密度とは、エッジ数を存在し得る最大のエッジ数（ノード数を $n$ とすると ${}_nC_2$）で割ったものである (Scott 2000)。

(3)　提案手法及び比較手法により、クラスタリングを行う。今回は、使用したカテゴリ数が 9 であるため、群平均法はクラスタ数が 9 になった時点でクラスタリングを終了とする。また、本実験では条件を均一化するために Newman 法においても終了条件を $< 0$ ではなくクラスタ数 9 とする。

(4)　同一クラスタに属する 2 語は関連語、異なるクラスタに属する 2 語は非関連語とする。この結果を正解セットと比較し、適合率・再現率・F 値を求める。

(5)　(1) から (4) を正解セット $A_i(i = 1\sim n)$ について行う。

OpenDirectory による正解セットを用いた評価結果を表 15 に、Roget's Thesaurus による正解セットを用いた評価結果を表 16 に示す。示されている値はそれぞれ、5 個の OpenDirectory 正解セット $Ao_i(i = 1\sim5)$ と 10 個の Roget's Thesaurus 正解セット $Aw_i(i = 1\sim10)$ について実験を行った結果の平均値である。各計算指標の群平均法と Newman 法の結果を比較すると、いずれも群平均法では適合率が高く、再現率が低い。クラスタリングの評価では一般的なことであるが、これは 1 つのクラスタにほとんどの語が含まれ、残り 8 つのクラスタにそれぞれ 1〜3 語程度の語が含まれている状態と考えられる。例えば、極端な例ではクラスタ内の語数が 1 で

14 $\chi^2$ 値による関連度を用いた関連語ネットワークの密度の平均値が約 0.3 であるため。

表 17　終了条件による比較 ($\chi^2$)（上段：適合率　中段：再現率 下段：F 値）

| | OpenDirectory | | WordNet | |
|---|---|---|---|---|
| | クラスタ数指定 | 自動終了 ($\Delta Q < 0$) | クラスタ数指定 | 自動終了 ($\Delta Q < 0$) |
| 適合率 | 0.738 | 0.601 | 0.598 | 0.470 |
| 再現率 | 0.631 | 0.911 | 0.411 | 0.591 |
| F 値 | 0.680 | 0.722 | 0.485 | 0.520 |

あれば適合率が $\frac{1}{1} = 1.0$ になる。そのため、含まれている語数の少ないクラスタが多数できる手法の方が精度が上がりやすい。しかし、再現率や F 値で見ると、各クラスタに含まれる語数が均等に近くなるようなクラスタリング手法の評価が高くなる。

表 15, 表 16 から群平均法の代わりに Newman 法を用いることで、いずれの指標においても F 値が高くなっている。このことから、提案手法を用いることでより適切に語がクラスタリングされていると言える。ただし、群平均法がこの実験に適していない可能性も考えられるので、今後他の手法との比較を行う必要がある。Newman 法を用いた場合の各指標を比較すると、表 15, 表 16 いずれにおいても、$\chi^2$ 値が最も良い F 値を示している。これより、語のクラスタリングを行う関連語ネットワークの構築には $\chi^2$ 値による関連度を用いることが適切であると言える。

次に評価手法 (3) における Newman 法の終了条件を「クラスタ数 9」とした場合と「$\Delta Q < 0$」とした場合の評価実験結果を表 17 に示す。またその際の正解セットごとのクラスタ数のグラフを図 3 に示す。

表 17 より、終了条件を「$\Delta Q < 0$」とした方が「クラスタ数指定」とした場合よりも高い F 値を示している。しかし、その差は 4 ポイント程度であり、精度に大きな違いはないといえる。これより、提案手法においては、条件としてクラスタ数を与えない場合でも、与えた場合とほぼ同程度の精度で関連語のクラスタリングを行うことができることがわかる。

ただし、今回の実験ではそれぞれの終了条件によって違う傾向を持っている。「クラスタ数指定」では、適合率 > 再現率 となっているが、「$\Delta Q < 0$」では、適合率 < 再現率 となっている。これは、終了条件によるクラスタ数の違いと Web を用いて関連度を算出する際に必ずしも目的とする語関連性が得られないためである、これに関して、クラスタリング結果の具体例を表 18 に示す。ここに用いられている語は表 9 に示されている語である。

本実験では、表 9 より、表 18 の正解セットでは「科学」及び「コンピュータ」という関連性によってクラスタリングされることが想定されている。しかし、実際には「クラスタ数指定」のクラスタ $A_1$ に含まれる語は、「情報科学」及び「プログラミング」という共通の関連性を持っていると考えられる。クラスタ $A_2$ に含まれる語は「Web 掲示板」という共通の関連性を持っていると考えられる。また「$\Delta Q < 0$」では、クラスタ $A_1, A_2$ が 1 つにマージされクラスタ $B$ を構成している。

表 18　クラスタリング結果の具体例

| 終了条件 | クラスタ名 | |
|---|---|---|
| クラスタ数指定 | クラスタ $A_1$ | 情報処理, 方法論, 実証, ソースコード, 文字コード, 初期値 |
| | クラスタ $A_2$ | 無料レンタル, 掲示板, アクセス数 |
| $\Delta Q < 1$ | クラスタ $B$ | 情報処理, 方法論, 実証, ソースコード, 文字コード<br>初期値, 無料レンタル, 掲示板, アクセス |

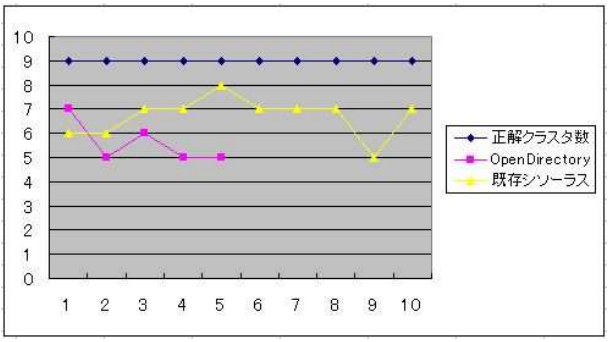


図 3　クラスタ数　横軸：正解セット　縦軸：クラスタ数

このように提案手法では、正解セットで目的としている関連性とは異なる関連性に基づいてクラスタリングされる場合が多い。これはクラスタリング手法によるものではなく、主に算出された関連度によるものである。実際には、クラスタ $A_1$ のように 2 つ以上のカテゴリの語で構成されるクラスタやクラスタ $A_2$ のように 1 つのカテゴリの語の一部のみで構成されるクラスタなど、正解セットのカテゴリ分けとは異なるクラスタができてしまっている。今回の実験においては、「クラスタ数指定」ではクラスタ $A_2$ のようなクラスタが多かったために 適合率 $>$ 再現率 となっている。また、クラスタ数の少なかった「$\Delta Q < 0$」ではクラスタ $B$ のようなクラスタが多かったために 適合率 $<$ 再現率 となっている。

以上より、提案手法の精度を高めていくためには、目的にあわせた関連度を取得する手法とより適切にクラスタ数を自動取得する手法が必要となってくる。

また、ネットワークのノード数とクラスタリングの実行時間の関係を図 4 に示す[15]。基準線は、$x$ をノード数、$z$ をエッジ数とするとき、式 $y = 1.8 \times 10^{-8} x(z + x)$ のあらわす曲線である ($1.8 \times 10^{-8}$は比例定数)。図 4 で実測値と基準線を比較するとほぼ一致しており、確かに Newman 法の計算量が $O(n(m + n))$ に比例している。そして、$n = 4029, m = 7146169$ のとき実行時間は 532 秒であり、$n, m$ が大きい大規模ネットワークにも提案手法が適用可能であると

15 実行環境　CPU:Pentium4 3.0Ghz　メモリ:1GB

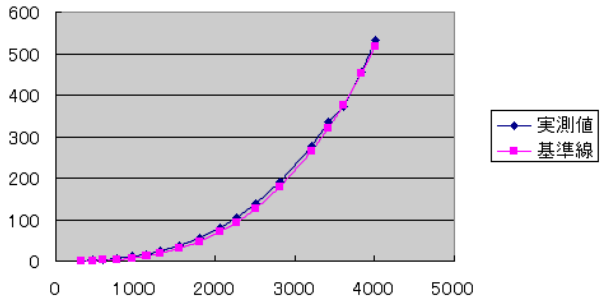


図 4　ノード数と実行時間　横軸：ノード数　縦軸：実行時間 (秒)

考えられる。

以上の評価実験の結果より、提案手法について以下の 3 点を述べることができる。

- 既存手法よりも適切に関連語のクラスタリングを行うことができる
- クラスタ数が未知の場合でも、クラスタ数が既知の場合と同程度の精度で関連語のクラスタリングを行うことができる
- 大規模なネットワークにも適用可能である

## 6　議論

語の関連は, 相対的なものである. 候補となる語群によって, あるときは関連した語同士でも, 他の場合には関連していないこともあり得る. ある語群において全ての語同士の関連度が分かっているとき, どの語とどの語を関連語と見なすかは, 関連度によって規定される語の関係性によると考えられる. 語の関連性を図 5 のようなネットワーク図（ノード間の距離を（1/語の関連度）とおく）で可視化すると, 図 5-a のような時は部分集合 A,B,C それぞれが, 関連語の集まった関連語群であると言える. 同様に図 5-b であれば, 部分集合 A,B,C,D それぞれが関連語群であると言える. このように語のネットワーク上で周囲と比べて密度が高くなっている部分を抽出することで, 各語の関連語を同定することができる.

Web は非常に多様性に富んだテキストから構成されている。したがって、目的に合わせた語の関連性を得るには、Web から適切な文書集合を切り出した上で、その文書集合内での関連度を求めるという方法が考えられる。これには、検索クエリーに特定の検索語（keyword spice）を加える方法が有効であろう (Oyama, Kokubo, and Ishida 2004)。

本論文では、関連語ネットワーク上のエッジには重みを与えていないが、語の関連性が多値

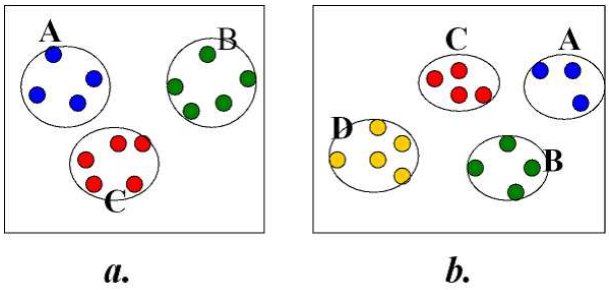


図 5　クラスタリングによる関連語の同定

的であることを考えると、重みを考慮する必要がある。ただし、既存の Newman 法は重みのあるネットワークに対応していない。そこで、重みを扱えるように Newman 法を改良することで、重みつきのネットワーク上でクラスタリングを行うことが考えられる。語の関連性を「関連がある、ない」の 2 値ではなく、重みという多値で扱うことで、クラスタ数の自動取得も含めて、より適切なクラスタリング結果が得られることが予想される。

加えて、Newman 法では 1 語が 1 つのクラスタリングにしか所属できないハードクラスタリングであるため、語の持つ多義性を解消することができない、という問題点がある。しかし、Newman 法をもとにしたソフトクラスタリングの手法も提案されており (Reichardt and Bornholdt 2004)、この手法を関連語ネットワークに適用することで語の多義性を解消できると考えられる。

また本研究では、同義・類義、上位語・下位語、連想語をすべて関連語としたが、こういった語を関係性を分類していくことも重要であろう。こういった研究には、前置詞を手がかりとして語の関係性を同定する (Litkowski 2002) の手法があるが、これを検索エンジンを利用していかに効率的に行うかは今後の検討課題のひとつである。

## 7　結論

本論文では、自動的に関連語のシソーラスを構築する手法について提案した。提案手法では、検索エンジンを利用し、Web をコーパスとして用いる。Newman 法をクラスタリング法として用いる部分が大きな特徴のひとつである。

検索エンジンを用いて語の関連度を取得する研究においては、コーパスを直接解析する手法と比べ、共起頻度以外の文法的な情報が得られないため、クラスタリングによって関連語を同定し、高い精度を得られている研究はなかった。本論文では、共起頻度のみを用いたクラスタリングで精度の高い関連語の同定に成功しており、そのような点で非常に有意義な研究だと考えられる。

また、語の関係の相対性に着目し、相対性を考慮した手法を用いた。$\chi^2$値 は語群内での相対的な偏りを示す統計的指標であり、また Newman 法はネットワーク全体で相対的に結合度の強いノードをマージするクラスタリング手法である。これらの手法を用いることにより、より適合率が高く、適用範囲の広いシソーラスの構築手法を提案することができた。

Web は重要な言語資源であり、その利用のためには検索エンジンの利用や大規模な処理への対応など、Web ならではのアルゴリズムの工夫が必要になる。今後、検索エンジンを利用した言語処理の可能性をさらに追求していきたい。

謝　　辞

株式会社社長ホットリンク下大園貞寛氏、国立情報学研究所大向一輝氏をはじめ、本研究にアドバイスをくださった全ての方に感謝いたします。

## 参考文献

Baker, D. and McCallum, A. (1998). “Distributional Clustering for Text Classification.” In *Proceedings of SIGIR-98,21st ACM International Conference on Research and Development in Information Retrieval*, pp. 96–103.

Barbara Ann Kipfer, P. (Ed.) (2006). Roget’s New Millennium ▪ Thesaurus, First Edition. Lexico Publishing Group.

Baroni, M. and Bisi, S. (2004). “Using cooccurrence statistics and the web to discover synonyms in a technical language.” In *In Proceedings of LREC2004*, pp. 26–28.

Brown, P., Pietra, V., deSouza, P., Lai, J., and Mercer, R. (1992). “Class-based n-gram model of natural language.” *Comput. Linguist.*, 18 (4), 467–479.

Chang, J. (2005). “Domain Specific Word Extraction from Hierarchical Web Documents: A First Step Toward Building Lexicon Trees from Web Corpora.” In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 64–71.

Church, W. and Hanks, P. (1990). “Word association norms, mutual information, and lexicography.” *Comput. Linguist.*, 16 (1).

Crouch, C. J. and Yang, B. (1992). “Experiments in automatic statistical thesaurus construction.” In *SIGIR ’92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 77–88.

Curran, J. (2002). “Ensemble Methods for Automatic Thesaurus Extraction.” In *Proceedings of the 2002 Conference on Empirical Methods in NLP*, pp. 222–229.

Curran, J. and Moens, M. (2002). “Improvements in Automatic Thesaurus Extraction.” In *Proceedings of the Workshop of the ACL SIGLEX*, pp. 59–66.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and Harshman, R. (1990). "Indexing by Latent Semantic Analysis." *Journal of the American Society of Information Science*, 41 (6), 391–407.

Dhillon, S. (2002). "Enhanced Word Clustering for Hierarchical Text Classification." In *Proceedings of the 8th ACM SIGKDD*, pp. 191–200.

Girvan, M. and Newman, M. (2002). "Community structure in social and biological networks." In *Proceedings of National Academic Science*, pp. 7821–7826.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers.

Heylighen, F. (2001). "Mining Associative Meanings from the Web: from word disambiguation to the global brain." In *Proceedings of the International Colloquium: Trends in Special Language Language Technology, R. Temmerman M. Lutjeharms*, pp. 15–44.

Hodge, V. and Austin, J. (2002). "Hierarchical word clustering – automatic thesaurus generation." *Neurocomputing*, 48, 819–846.

Jarmasz, M. and Szpakowicz, S. (2003). "Roget's Thesaurus and Semantic Similarity." In *Proceedings of Conference Recnet Advances in NLP*, pp. 212–219.

Kilgarriff, A. and Grefenstette, G. (2003). "Web as Corpus." In *In Proceedings. of the ACL Workshop on Intellignent Scalable Text Summarization.*

Li, H. and Abe, N. (1998). "Word clustering and disambiguation based on co-occurrence data." In *Proceedings of the 17th international conference on Computational linguistics*, pp. 749–755. Association for Computational Linguistics.

Lin, D. (1998). "Automatic retrieval and clustering of similar words." In *Proceedings of the 17th international conference on Computational linguistics*, pp. 768–774 Morristown, NJ, USA. Association for Computational Linguistics.

Litkowski, C. (2002). "Digraph Analysis of Dictionary Preposition definition." In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pp. 9–16.

Manning, C. and Schutze, H. (1999). "Foundations of statistical natural language processing." In *MITPress.*

Miller, G. (1990). "WordNet:an on-line lexical database.." In *International Booktitle of Lexicography.*

Motter, A., Moura, A., Lai, Y., and Dasgupta, P. (2002). "Topology of the conceptual network of language." *Physical Review E*, 65 (065102).

長尾真, 水谷幹男, 池田浩之 (1976). "日本語文献における重要語の自動抽出." 情報処理, 17 (2),

pp.110–117.

Newman, M. (2004). “Fast algorithm for detecting community structure in networks.” In *Phys. Rev. E 69,2004*.

Oyama, S., Kokubo, T., and Ishida, T. (2004). “Domain-Specific Web Search with Keyword Spices.” *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16 (1), 17–27.

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). “Uncovering the overlapping community structure of complex networks in nature and society.” *Nature*, 435, 814–818.

Papadimitriou, C., Tamaki, H., Raghavan, P., and Vempala, S. (1998). “Latent Semantic Indexing: A Probabilistic Analysis.”.

Pereira, F. Tishby, N. and Lee, L. (1993). “Distributional Clustering of English words.” In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pp. 183–190.

Reichardt, J. and Bornholdt, S. (2004). “Detecting Fuzzy Community Structures in Complex Networks with a Potts Model.” *Physical Review Letters*, 93.

Sanderson, M. and Croft, B. (1999). “Deriving concept hierarchies from text.” In *SIGIR ’99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 206–213. ACM Press.

Scott, J. (2000). *Social Network Analysis: A Handbook*. SAGE Publications.

Sigman, M. and Cecchi, G. (2002). “Global organization of the Wordnet lexicon.” *PNAS*, 99 (3), 1742–1747.

Slonim, N. and Tishby, N. (2000). “Document Clustring using Word Cluster via the Information Bottle neck Method.” In *Research and Development Information Retrieval*, pp. 208–215.

Szpektor, I., Tanev, H., Dagan, I., and Coppola, B. (2004). “Scaling Web-based Acquisition of Entailment Relations.” In *Proceedings of EMNLP 2004*, pp. 41–48. Association for Computational Linguistics.

Turney, P. (2001). “Mining the web for synonyms: PMI-IR versus LSA on TOEFL.” In *EMCL ’01: Proceedings of the 12th European Conference on Machine Learning*, pp. 491–502.

Ward, G. (Ed.) (1996). *Moby Thesaurus*. Moby Project.

Wettler, M. and Rapp, R. (1993). “Computation of word associations based on the co-occurrences of words in large corpora..” In *In Proceedings of the 1st Workshop on Very Large Corpora:Academic and Industrial Perspectives*, pp. 84–93.

Widdows, D. and Dorow, B. (2002). “A Graph Model for Unsupervised Lexical Acquisition..”

In *COLING 2002, 19th International Conference on Computational Linguistics*.

Xu, F., Kurz, D., Piskorski, J., and Schmeier, S. (2002). “A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping.” In *Proceedings of the 3rd International Conference on Language Resources an Evaluation (LREC'02), May 29-31*.

Yang, Y. and Pedersen, J. (1997). “A Comparative Study on Feature Selection in Text Categorization.” In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412–420. Morgan Kaufmann Publishers Inc.

佐々木靖弘, 佐藤理史, 宇津呂武仁 (2005). “ウェブを利用した専門用語集の自動編集.” 言語処理学会第 11 回年次大会発表論文集,, pp. 895–898.

池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦（編） (1997). 日本語語彙大系. 岩波書店.

日本電子化辞書研究所（編） (1996). EDR 電子化辞書 仕様説明書. 日本電子化辞書研究所.

## 略歴

榊　剛史:　2000 年東京大学工学部電子情報工学科卒業。2006 年同大学院情報理工学系研究科修士課程修了。研究分野は Web マイニング、言語処理。

松尾　豊:　1997 年東京大学工学部電子情報工学科卒業。2002 年同大学院博士課程修了。同年より産業技術総合研究所に勤務。現在同研究所情報技術研究部門勤務、スタンフォード大学客員研究員。高次 Web マイニングに興味がある。人工知能学会、情報処理学会、AAAI、各会員。

石塚　満:　1971 年東京大学工学部電子卒。1976 年同大学院博士課程修了。NTT、東大大学院生産技術研究所・助教授、同大学工学部電子情報・教授を経て、現在は 2005 年より同大学院情報理工学系研究科創造情報学専攻・教授。研究分野は人工知能、Web インテリジェンス、次世代 Web 情報基盤、マルチモーダルエージェント。人工知能学会 (会長)、IEEE、AAAI、情報処理学会、電子情報通信学会の会員。

( 受付)
( 再受付)
( 採録)