Web上の情報を用いた関連語のシソーラス構築

石塚 満　（協力者：榊 剛史, 松尾 豊）
情報理工学系研究科　電子情報学専攻/創造情報学専攻

概 要

情報流通，情報共有の基幹的インフラストラクチャになってきたWWW(World Wide Web)に知的能力を付与し、広域知能基盤に成長させるための研究開発を行っている。特にテキスト処理に基づく知能化を図っているが、ここではその一つの基礎技術として、Web情報を利用して自動的に関連語のシソーラスを構築する手法を研究開発したので，報告する。関連語シソーラスは、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を始め、言語処理のさまざまな場面で必要になり、Webに関するテキストの知的処理にも欠かせないデータ/知識となる。

本研究では検索エンジンを利用し、χ^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している。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っている。

1　まえがき

関連語シソーラスは、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[21] やEDR電子化辞書 [41]、日本語語彙体系など、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。WordNetやEDR電子化辞書などのシソーラスは人手で構築されている。これらのシソーラスは、有用で多くの研究で使われており、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 [6, 12]。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。

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 [8]。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 [5]。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 [16]。佐藤、宇津呂らはWebを用いた関連度の指標を提案している [40]。

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 [35, 13]。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、関連語を同定するために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、より広い意味でのある語に関連した語を関連語とする。

表 1: 類似度の計算指標

ベクトル空間手法		確率手法	
cosine	$\frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\|\vec{y}\|}$	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	$\frac{\|\vec{x}\cap\vec{y}\|}{min(\|\vec{x}\|,\|\vec{y}\|)}$	T 検定	$\frac{p(w\cap w')-p(w')p(w)}{\sqrt{p(w')p(w)}}$
Lin[1]	$\frac{\sum(x_i+y_i)}{\|\vec{x}\|+\|\vec{y}\|}$	Lin98A[2]	$\log\left(\frac{f(w,r,w')f(*,r,*)}{f(*,r,w')f(w,r,*)}\right)$

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 [5, 36, 30, 8]、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 [18] 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 [3, 1, 33]。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 [28, 17, 10]。

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。

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に Kilgarriff らは検索エンジンを用いた手法を紹介している [16]。「語 w_a」をクエリー

[2] [18] で提案されている手法
[3] [18] で提案されている手法

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 [13]。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 [2, 35]。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 [40]。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 [34]。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、[2] のものと同一である[3]。関連度を計りたい語を、例えば「インク」「インターレーザー」「プリンタ」「印刷」「液晶」「Aquos」「TV」「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、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[4]。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、

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

[3]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。

[4]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	TV	Aquos	Sharp
17000000	103000000	215	18900000	69100000	1760000000	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	TV	Aquos	Sharp	合計
プリンタ	0	4780000	179	4720000	4820000	4530000	201000	990000	20041273
印刷	4780000	0	183	4800000	6520000	8390000	86400	1390000	25966583
インターレーザー	179	183	0	116	176	65	0	0	813
インク	4720000	4800000	116	0	3230000	10600000	144000	656000	24150116
液晶	4820000	6520000	176	3230000	0	13900000	903000	4880000	34253176
TV	4530000	8390000	65	10600000	13900000	0	1660000	42300000	81380065
Aquos	201000	86400	0	144000	903000	1660000	0	1790000	4784400
Sharp	990000	1390000	0	656000	4880000	42300000	1790000	0	52006000
合計	20041273	25966583	813	24150116	34253176	81380065	4784400	52006000	242582426

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、2005 年末現在では、Google は約 100 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「TV」「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、「TV」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「出現確率に影響を受ける」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しまう。例えば、「TV」のように出現確率に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「TV」の列に注目すると、いずれの語においても「TV」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は -0.65 となり、相互情報量と語の出現確率に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 -0.15 となり、あまり相関がないことが分かる。

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。

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	TV	Aquos	Sharp
プリンタ	0	4.195	7.504	5.878	4.602	1.303	4.740	2.029
印刷	4.195	0	5.302	4.093	3.103	0.117	2.094	0.567
インターレーザー	7.504	5.302	0	6.542	5.663	1.429	0.000	0.000
インク	5.878	4.093	6.542	0	4.096	2.047	4.301	1.512
液晶	4.602	3.103	5.663	4.096	0	1.021	4.840	2.222
TV	1.303	0.117	1.429	2.047	1.021	0	2.212	1.144
Aquos	4.740	2.094	0.000	4.301	4.840	2.212	0	4.534
Sharp	2.029	0.567	0.000	1.512	2.222	1.144	4.534	0

3.2 χ^2 値を用いた関連度の指標

本研究では、χ^2 値を使った関連度の指標を用いる。χ^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 [8]。

χ^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られる。χ^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 [39]。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、χ^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 χ^2 値は次式で表される。

$$\begin{aligned} \chi^2(w_i, w_j) &= \frac{n(w_i, w_j) - E(w_i, w_j)}{E(w_i, w_j)} \\ E(w_i, w_j) &= S_{w_i} \times \frac{S_{w_j}}{S_G} \end{aligned} \tag{3}$$

$E(w_i, w_j)$ 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 179、$S_{w_i} = 20041273$、$S_{w_j}/S_G = 813/242582426$ となる。表 5 は、表 3 から計算された χ^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 χ^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: χ^2 行列

語/語	プリンタ	印刷	インターレーザー	インク	液晶	TV	Aquos	Sharp
プリンタ	0.000	3235887	630.8	3721225	1399572	0.000	0.000	0.000
印刷	3235887	0.000	105.8	1897753	2220688	0.000	0.000	0.000
インターレーザー	630.8	105.8	0.000	15.19	32.63	0.000	0.000	0.000
インク	3721225	1897753	15.19	0.000	0.000	770371	0.000	0.000
液晶	1399572	2220688	32.63	0.000	0.000	505007	76566	0.000
TV	0.000	0.000	0.000	770371	505007	0.000	1882	35404428
Aquos	0.000	0.000	0.000	0.000	76566	1882	0.000	569512
Sharp	0.000	0.000	0.000	0.000	0.000	35404428	569512	0.000

表 6: χ^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着目を集めている。

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。ここでは、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。

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

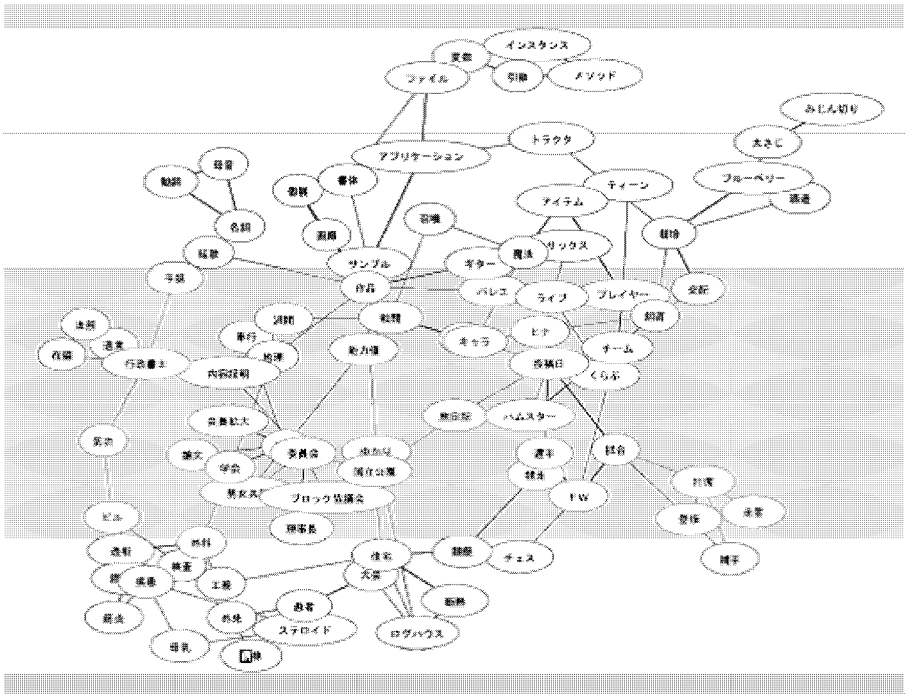

図 1: 関連語ネットワーク

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。

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。[28, 10]。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 [9, 27]。

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 [37, 22, 26]。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 [32]。

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[5]の betweenness というエッジの媒介性を表す

[5]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

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
$D(c_i, c_j)$	$\max_{w_k \in c_i, w_l \in c_j} Sim(w_k, w_l)$	$\min_{w_k \in c_i, w_l \in c_j} Sim(w_k, w_l)$	$\frac{1}{n_i n_j} \sum_{w_k \in c_i} \sum_{w_l \in c_j} Sim(w_k, w_l)$

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 [11]。

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。

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 [24]。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。

$$Q = \frac{1}{2m} \left[\left(\sum_{v,w} A_{vw} \delta(c_v, c_w) \right) - \left(\sum_{x,w} \frac{k_v k_w}{2m} \delta(c_v, c_w) \right) \right] \tag{5}$$

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 δ である。式 (5) 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。すなわち、Q の変化量 ΔQ が最大になるようなクラスタのマージを繰り返していくことで、Q の極大値を求める。そして、Q が極大値となった時点でクラスタリングを終了する。

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。

い、接続関係だけを表すものである。

表 8: 評価実験の例 (ヴァイオリン)

	関連語	適合率	再現率
正解データ	ビオラ, チェロ, 笛, ギター		
手法 1	ビオラ, チェロ, ビール	0.67	0.5
手法 2	ビオラ, チェロ, ピック, 笛, ギター	0.8	1.0

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 χ^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 ΔQ(式 6) 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. ΔQ 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。

$$\begin{aligned} \Delta Q_{ij} &= 2(e_{ij} - a_i a_j) \\ e_{ij} &= \text{クラスタ } i, j \text{ 間のエッジの本数 (割合)} \\ a_i &= \sum_i e_{ii} \end{aligned} \tag{6}$$

5 評価

5.1 評価実験の概要と正解データの作成

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 [15, 7]、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 [30, 14] 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、OpenDirectory[6]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した。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」の 9 つのカテゴリを用いた[7]。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。

[6] http://dmoz.org/World/Japanese/

[7] 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。

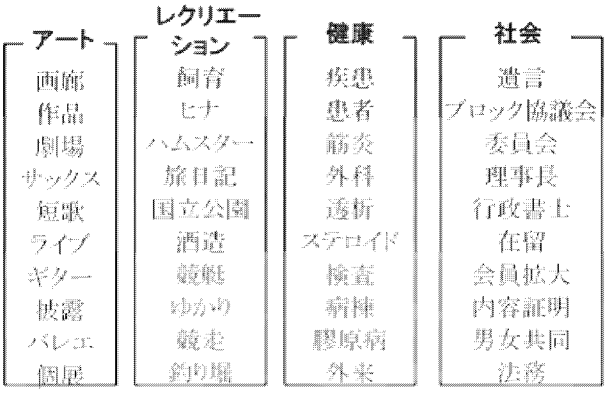

図 2: 獲得された単語リスト

1. 各カテゴリ $C_i(i = 1...9)$ ごとに登録順に 1000 ページを取得する。
2. 全ての文書に形態素解析[8]を行う。そして 5-gram までを単語として取り出す [20]。
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。

$$score^i_{w_a} = f^i_{w_a} \times \log(N/f^{all}_{w_a}) \tag{7}$$

   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。

得られた語の一部を図 2 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、[23] や [38] 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 [4] 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、ここでの正解データは完全ではなく、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 3 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

[8]茶筌。http://chasen.aist-nara.ac.jp/.

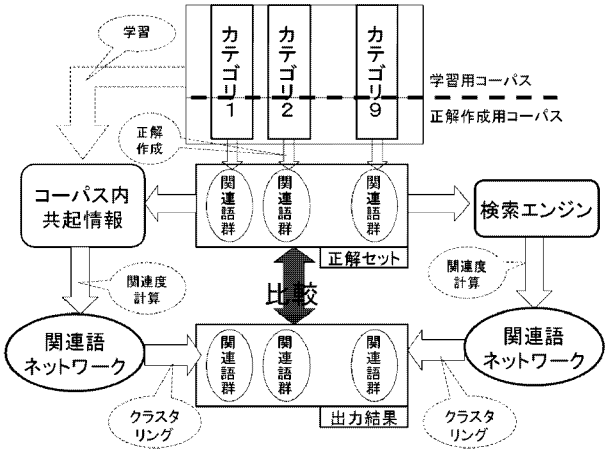

図 3: 評価実験の概略図

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 3 に示すとおり、本評価実験では正解データ作成用コーパス、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 千ページをコーパスとして用意し、1/5 を正解データ作成用に、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[9]。正解データ作成用のコーパスを変えたそれぞれの正解セットを $A_i(i = 1, 2, 3, 4, 5)$ とする。

シソーラスの評価は、適合率と再現率によって測る。簡単な算出例を表 8 に示す。この場合、手法 1 による出力は 3 語中 2 語が正解であるので適合率は 0.667、正解データ 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、再現率は 0.5 となる。関連語の正解データを用いたこのような評価方法はいくつかの論文中でも用いられている [37, 7]。

5.2 関連度の指標の評価

まず、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 χ^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。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
2. 各指標ごとに語 w 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。G_w と正解データを比較し、適合率を計算する。
3. (2) を語 $w \in A_i$ 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 $A_i(i = 1\sim5)$ について行う。

正解セットごとの適合率とそれらの平均値を表 9 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 χ^2 値が他の 2 つの計算指標よりもよい値を示している。これより、χ^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 よりも低い適合率である。cosine と χ^2 値を比較すると正解セットによって適合率の優劣が変化している。正解セット A_1, A_3, A_4 では χ^2 の方が高い適合率を示し

[9]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。

表 9: 指標による適合率の違い

正解セット/指標	cosine	相互情報量	Jaccard 係数	χ^2 値
セット A_1	0.557	0.447	0.424	0.567
セット A_2	0.513	0.406	0.389	0.493
セット A_3	0.519	0.396	0.376	0.539
セット A_4	0.561	0.404	0.417	0.569
セット A_5	0.529	0.421	0.404	0.519
平均	0.535	0.415	0.402	0.538

表 10: 関連語抽出実験結果（上段：適合率 中段：再現率 下段：F 値）

クラスタ		cosine	相互情報量	Jaccard 係数	χ^2 値
群平均法	適合率	0.449	0.633	0.092	0.486
	再現率	0.096	0.102	0.101	0.100
	F 値	0.156	0.179	0.173	0.164
Newman 法	適合率	0.620	0.751	0.739	0.546
	再現率	0.335	0.103	0.103	0.431
	F 値	0.431	0.182	0.181	0.480

ているが、A_2, A_5 では cosine の方が高い適合率を示している。しかし、いずれのセットにおいても値の差は小さく、平均でもほとんど差がないことから、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、χ^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。

これより、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な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 [10]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 [31]。
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。
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。この結果を正解データと比較し、適合率・再現率・F 値を求める。
5. (1) から (4) を正解セット $A_i(i = 1\sim5)$ について行う。

[10] χ^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。

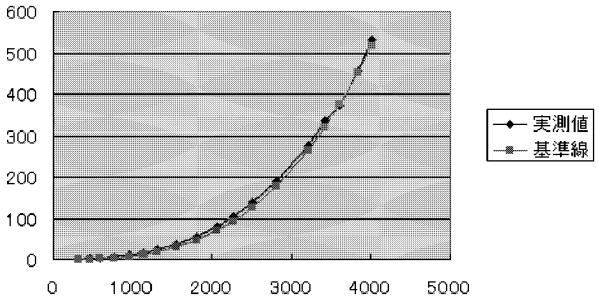

図 4: ノード数と実行時間　横軸：ノード数　縦軸：実行時間 (秒)

結果を表 10 に示す。示されている値は 5 つの正解セット $A_i(i = 1\sim5)$ 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、これは小さいクラスタが多数でき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 であると、これが含まれていれば適合率が 1.0 になり、小さいクラスタができる手法が精度が上がりやすい。しかし、再現率や F 値で見ると、適切な大きさのクラスタを生成する手法が評価が高くなる。

表 10 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、χ^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 χ^2 値による関連度を用いることが適切であると言える。

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[11]。基準線は、x をノード数、z をエッジ数とするとき、式 $y = 1.8\times10^{-8}x(z+x)$ のあらわす直線である (1.8×10^{-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考えられる。

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度を得るには、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語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 [25]。

[11]実行環境　CPU:Pentium4 3.0Ghz　メモリ:1GB

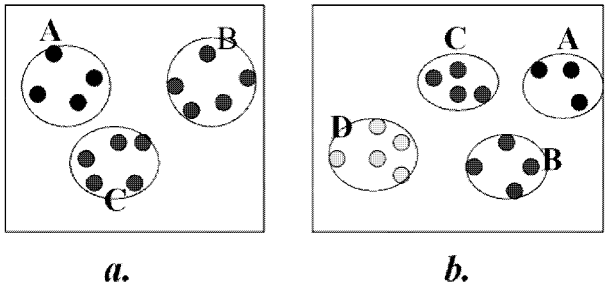

図 5: クラスタリングによる関連語の同定

本研究では、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、式 (6) を次式のように修正した Newman 法を用いることで、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、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。

$$\begin{aligned} \Delta Q_{ij} &= 2(e_{ij} - a_i a_j) \\ e_{ij} &= \frac{\text{クラスタ } i, j \text{ 間のエッジの重みの和}}{\text{全エッジの重みの和}} \\ a_i &= \frac{\text{クラスタ } i \text{ 内の頂点と結び付いたエッジの重みの和}}{\text{全エッジの重みの和}} \end{aligned} \tag{8}$$

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 [29]、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 [19] 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 むすび

本研究では、関連語の自動的に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。

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。χ^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。

参考文献

[1] Baker, D. McCallum, A. : Distributional Clustering for Text Classification Proc. SIGIR-98, 21st ACM Int'l Conf. on Research and Development in Information Retrieval, 96–103, 1998.

[2] Baroni, M. Bisi, S. : Using cooccurrence statistics and the web to discover synonyms in a technical language Proc. of LREC2004, 26–28, 2004.

[3] Brown, P., Pietra, V., deSouza, P., Lai, J., Mercer, R. : Class-based n-gram model of natural language Comput. Linguist., 18 (4), 467–479, 1992.

[4] Chang, J. : Domain Specific Word Extraction from Hierarchical Web Documents: A First Step Toward Building Lexicon Trees from Web Corpora In Proc. 4th SIGHAN Workshop on Chinese Language Processing, 64–71, 2005.

[5] Church, W. Hanks, P. 1990. Word association norms, mutual information, and lexicography Comput. Linguist., 16 (1), 1990.

[6] Crouch, C. J. Yang, B. 1992. Experiments in automatic statistical thesaurus construction In SIGIR '92: Proc. 15th annual international ACM SIGIR conference on Research and development in information retrieval, 77–88, 1992.

[7] Curran, J. : Ensemble Methods for Automatic Thesaurus Extraction Proc. 2002 Conf. on Empirical Methods in NLP, 222–229, 2002.

[8] Curran, J. Moens, M. : Improvements in Automatic Thesaurus Extraction Proc. Workshop of the ACL SIGLEX, 59–66, 2002.

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

[10] Dhillon, S. : Enhanced Word Clustering for Hierarchical Text Classification Proc. 8th ACM SIGKDD, 191–200, 2002.

[11] Girvan, M. Newman, M. : Community structure in social and biological networks Proc. of National Academic Science, 7821–7826, 2002.

[12] Grefenstette, G. : Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, 1994.

[13] Heylighen, F. : Mining Associative Meanings from the Web: from word disambiguation to the global brain Proc. Int'l Colloquium: Trends in Special Language Language Technology, R. Temmerman M. Lutjeharms, 15–44, 2001.

[14] Hodge, V. Austin, J. : Hierarchical word clustering – automatic thesaurus generation Neurocomputing, 48, 819–846, 2002.

[15] Jarmasz, M. Szpakowicz, S. : Roget's Thesaurus and Semantic Similarity Proc. Conf. Recnet Advances in NLP, 212–219, 2003.

[16] Kilgarriff, A. Grefenstette, G. : Web as Corpus In Proc. ACL Workshop on Intellignent Scalable Text Summarization, 2003.

[17] Li, H. Abe, N. : Word clustering and disambiguation based on co-occurrence data Proc. 17th int'l Conf. on Computational linguistics, 749–755. Association for Computational Linguistics, 1998.

[18] Lin, D. : Automatic retrieval and clustering of similar words In Proc. 17th Int'l Conf. on Computational linguistics, 768–774 Morristown, NJ, USA. Association for Computational Linguistics, 1998.

[19] Litkowski, C. : Digraph Analysis of Dictionary Preposition definition Proc. SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, 9–16, 2002.

[20] Manning, C. Schutze, H. : Foundations of statistical natural language processing MITPress, 1999.

[21] Miller, G. : WordNet:an on-line lexical database. In Int'l Booktitle of Lexicography, 1990.

[22] Motter, A., Moura, A., Lai, Y., Dasgupta, P. : Topology of the conceptual network of language Physical Review E, 65 (065102), 2002.

[23] 長尾真, 水谷幹男, 池田浩之 : 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, 1976.

[24] Newman, M. : Fast algorithm for detecting community structure in networks Phys. Rev. E 69, 2004.

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

[26] Palla, G., Derényi, I., Farkas, I., Vicsek, T. : Uncovering the overlapping community structure of complex networks in nature and society Nature, 435, 814–818, 2005.

[27] Papadimitriou, C., Tamaki, H., Raghavan, P., Vempala, S. : Latent Semantic Indexing: A Probabilistic Analysis, 1998.

[28] Pereira, F. Tishby, N. Lee, L. : Distributional Clustering of English words Proc. 31st Annual Meeting on Association for Computational Linguistics, 183–190, 1993.

[29] Reichardt, J. Bornholdt, S. : Detecting Fuzzy Community Structures in Complex Networks with a Potts Model Physical Review Letters, 93, 2004.

[30] Sanderson, M. Croft, B. : Deriving concept hierarchies from text SIGIR'99: Proc. 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval, 206–213. ACM Press, 1999.

[31] Scott, J. : Social Network Analysis: A Handbook. SAGE Publications, 2000.

[32] Sigman, M. Cecchi, G. : Global Organization of the Wordnet Lexicon PNAS, 99 (3), 1742–1747, 2002.

[33] Slonim, N. Tishby, N. : Document Clustring using Word Cluster via the Information Bottle neck Method In Research and Development Information Retrieval, 208–215, 2000.

[34] Szpektor, I., Tanev, H., Dagan, I., Coppola, B. : Scaling Web-based Acquisition of Entailment Relations Proc. EMNLP 2004, 41–48. Association for Computational Linguistics, 2004.

[35] Turney, P. : Mining the web for synonyms: PMI-IR versus LSA on TOEFL EMCL'01: Proc. 12th European Conference on Machine Learning, 491–502, 2001.

[36] Wettler, M. Rapp, R. : Computation of word associations based on the co-occurrences of words in large corpora. In Proc. 1st Workshop on Very Large Corpora:Academic and Industrial Perspectives, 84–93, 1993.

[37] Widdows, D. Dorow, B. : A Graph Model for Unsupervised Lexical Acquisition. COLING 2002, 19th Int'l Conf. on Computational Linguistics, 2002.

[38] Xu, F., Kurz, D., Piskorski, J., Schmeier, S. : A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping Proc. 3rd Int'l Conf. on Language Resources an Evaluation (LREC'02), 2002.

[39] Yang, Y. Pedersen, J. : A Comparative Study on Feature Selection in Text Categorization ICML'97: Proc. Fourteenth Int'l Conf. on Machine Learning, 412–420. Morgan Kaufmann Publishers Inc., 1997.

[40] 佐々木靖弘, 佐藤理史, 宇津呂武仁 : 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,, 895–898, 2005.

[41] 日本電子化辞書研究所 : EDR 電子化辞書 仕様説明書. 日本電子化辞書研究所, 1996.