自然言語処理 133−16
(1999. 9. 10)

タームのrepresentativeness を測る

久光 徹† 丹羽 芳樹† 辻井潤一‡

†日立製作所 中央研究所
‡東京大学理学部情報科学科

要旨

文書検索において，検索の結果得られた文書数が大きい場合，その内容を把握し，意図した方向へと検索を進めることは容易ではない．これを補助するためには，文書集合の内容を俯瞰するために，文書集合中の特徴語を提示することが有効であるとわかってきた．本研究は，特徴語を選ぶための，単語の話題性もしくは分野代表性(representativeness)を測る新しい指標を提案する．日経新聞を用いた指標の選別能力の評価実験，学術情報センターのAI論文アブストラクトを用いた用語抽出実験を通して，新指標の有効性を示す．

キーワード 語彙抽出，重み付け，representativeness，stop-word list，情報検索

Measuring Representativeness of Terms

Toru HISAMITSU†, Yoshiki NIWA†, and Jun'ichi TSUJII‡

† Central Research Laboratory, Hitachi, Ltd.
‡ Department of Information Science, The University of Tokyo

Abstract

This paper presents a novel method of measuring the representativeness of a term, i.e., informativeness or domain-specificity of a term. The value is defined by normalizing the distance between the word distribution in the documents containing the term and the background word distribution. The measure can compare the representativeness of two terms with highly different frequencies, and has a naturally defined threshold of being representative. Experiments showed that the measure clearly outperforms existing measures in evaluating terms in different domains; newspaper articles and abstracts of AI papers.

keywords: term extraction, term weighting, representativeness, stop-word list, information retrieval

1. はじめに

大量の文書が電子的に利用可能となった現在，文書検索の結果得られる文書数はしばしば膨大なものとなる．このようなときに，検索を意図した方向へ進める，もしくは，検索目的自体をリファインしてゆくためには，検索された文書集合の内容を俯瞰できることが必要である．

しかし，文書名の一覧だけでこれを実現することはきわめて困難であるため，文書集合の内容を俯瞰するための手助けとなると思われる語を自動的に選択し，「特徴単語グラフ」として提示する方法が提案され(Niwa 1997)，文書検索システムにおいて実際に利用されている(Nishioka et al. 1997)．その結果，特徴単語グラフの有効性があきらかになってきたが，インターフェイスの一層の改善のためには，解決すべき課題がある．

最も大きな問題は，語の品質に関するものである．現時点では，人手で，stop-word list を作らないかぎり，「する」や「ない」のような明らかに話題を担わない高頻度語の出現を避けることはできない．しかし，stop-word list の要素を選定するための方法は確立されておらず，恣意的になりがちであった．

図1は，キーワードを「電子マネー」とした場合の特徴単語グラフである．ここでは stop-word listを用いているために「する」などは現れていないが，「上」のような，stop-word list から漏れた単語は現れている．また，単語は上から下へ頻度順に5階層に分けられ，各階層から*tf-idf*(2.1参照)を用いて適宜選択されているが，語間の話題性の違いは必ずしも反映されておらず，例えば，「暗号化」と「読みとる」は，ほぼ同じ位置に表示され，二つの話題性の違いは，特に区別されてはいない．

そこで，「話題性，もしくは分野代表性(representativeness)の高い語」を選ぶための何らかの指標が必要となる．本報告の目的は，上記目的のための新しい指標を提案し，その能力を予備的に検証することである．

2. 従来用いられてきた諸指標

2.1 概観

情報検索やターム抽出の分野でも，語の「話題性」や「分野代表性」(すなわちrepresentativeness)を測る為の指標が数多く提案されてきた．これらについては，優れたサーベイが存在するので，これに沿って概観を述べる(Kageura et al. 1998).

上記文献は，専門用語(ターム)の自動抽出を目的とした研究のサーベイであり，まずタームのタームたる特徴を整理して述べている．すなわち，タームとは語の列であって，*unithood* と*termhood*を併せ持つものである．ここに，unithoodとは，"the degree of strength or stability of syntagmatic combinations or collocations"であり，termhoodとは，"the degree that a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts"である．Kageuraのいうtermhoodは，上で述べたrepresentativeness とほぼ同等の概念と言って良い．

語の重要度に関する指標は，歴史的には情報検索の分野で語の重み付けのために導入されてきた．最も単純なものは，文書内頻度に注目するもの(Luhn 1957)や全文書における出現頻度に注目するもの(Sparck-Jones et al. 1973)である．

より工夫されたものとして，注目する単語の分布の偏りをとらえようとするものであり，最も有名な例は，*tf-idf* (Salton et al., 1973)である．*idf* (Sparck-Jones, 1972)は，全文書数N_{total}をある単語wが現れる文書数$N(w)$で割ったものの対数であり，*tf*は単語wの文書 d 内での出現頻度 $f(w, d)$ である．*tf-idf*は，これらの積として，

$$f(w, d)\times \log(\frac{N_{total}}{N(w)}),$$

と現され，現在でもよく用いられている．他にもさまざまな変形があるが，*tf-idf*の基本的な性質として，「単語がより多く，より少ない文書に偏って出現するほど大きくなる」ように設定される．上記文献には記述されていないが，この指標を特定の文書中での単語の重要度でなく，文書集合全体での単語の重要度を測る指標に拡張する自然な方法は，$f(w,d)$を，wの全文書中での頻度$f(w)$に置き換えることであり，4.1.2で用いられる．

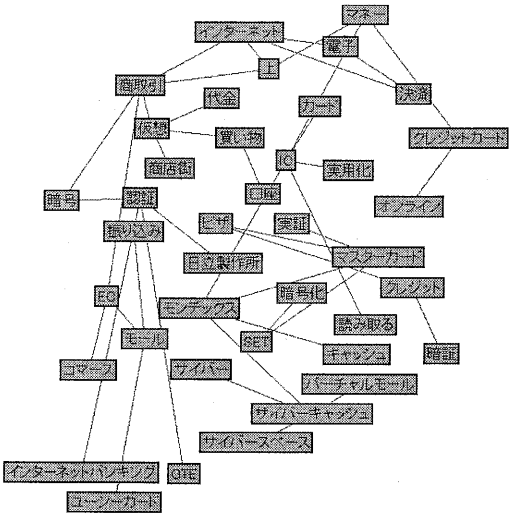


図1
「電子マネー」をキーワードとして用いた時の特徴単語グラフ

全文書中から重要語を抽出するための方法の一つとして，単語の出現の偏りをより精密に捕えるために，注目する単語の，与えられた文書カテゴリごとの出現頻度の差異の偶然性を測り，偶然でない度合いが高いものを重要語としようという方法があり，尺度としてχ^2検定等が利用されているが(長尾 他, 1976)，この場合，文書集合はあらかじめカテゴリに分類されている必要がある．

これらと別系統の研究として，自然言語処理の立場から，タームとしてふさわしい語のまとまりを捕えようとする一連の研究があり，隣り合う単語の共起の強さを相互情報量(Church et al. 1990)で測るもの，対数尤度比で測るもの(Cohen 1995)が提案されている．さらに，一般の連語まで扱うために，仕事量基準(Kita et al. 1994)，C-value (Franzi et al. 1996), Nested collocation (Nakagawa et al. 1997)等が導入されている．

2.2 問題点

従来提案されてきた指標には，以下のような問題があった：

(1) *tf-idf*(もしくはその類似手法)の精度は不充分である．経験上語の頻度の寄与が大きすぎる傾向があり，例えば「する」のような一般的すぎる不用語の排除ができにくい．
(2) 特定の語のカテゴリ間での分布の違いを比較する方法では，あらかじめ文書が分類されている必要があるが，この条件は一般に満たされない．
(3) 隣り合う単語の共起の強さを利用する手法では，1単語のみの場合重要度が評価できない．
(4) 重要／非重要を分ける閾値の設定が困難かつ恣意的になりがちであった．

本報の目的は，このような問題の無い指標を提案することである．

3. "representativeness"を測るための新指標

3.1 基本方針

指標の定義をする前に，今一度我々の目的に沿った「タームの好ましさ」の指標，すなわち"representativeness"の定義に戻ると，あるタームが"representative"であるとは，そのタームがある話題(もしくはいくつかの話題群)を想起させてくれることを指す．これは検索の結果得られた文書集合を俯瞰し，新たにキーとなるタームを示唆する際に重要である．

このような性質を測る際の基本的な考え方は，John Rupert Firthの次の言葉，*"You shall know a word by the company it keeps"*.(Firth 1957) に集約される．本論文では，これを数学的に解釈することにより，タームのrepresentativenessを測る指標を導入する．すなわち，任意のタームW(単語または単語列)について，Wを含む文書すべての集合における単語分布と，全文書の単語分布の異なりの度合いに着目する。具体的には，

W: ターム(任意の個数の単語からなる)
$D(W)$: Wを含む文書すべての集合
D_0: 全文書の集合
$P_{D(W)}$: $D(W)$における単語分布
P_0: D_0 における単語分布

とするとき，Wのrepresentativeness $Rep(W)$を，２つの分布$\{P_{D(W)},\ P_0\}$の距離 $Dist\{P_{D(W)},\ P_0\}$に基づいて定義する．

単語分布間の距離の計測の方法としては，主要なものだけでも，

(1) 対数尤度比(log-likelihood ratio),
(2) Kullback-Leibler divergence,
(3) transition probability,
(4) vector-space model (cosign 法)

等が考えられるが，ここでは，(1)を用いた．すなわち，全単語を$\{W_1,...,W_n\}$，k_i とK_i を，単語w_i が$D(W)$，D_0 に出現する頻度とするとき，$P_{D(W)}$ と P_0 の距離 $Dist\{P_{D(W)},\ P_0\}$を，以下で定義する：

$$Dist(P_{D(W)}, P_0) = \sum_{i=1}^{n} k_i \log \frac{k_i}{\#D(W)} - \sum_{i=1}^{n} k_i \log \frac{K_i}{\#D_0}.$$

図２は，日経新聞1996年版の記事を用い，そこにあらわれるいくつかの語Wに対し、各語Wについて，$D(W)$の含む単語数 $\#D(W)$を横軸に，$Dist\{P_{D(W)},\ P_0\}$を縦軸にプロットしたものである．図から見られるとおり，$\#D(W)$が近いターム同士で比較すれば，たとえば「米国」は「する」，「オウム」は「結び付ける」より$Dist\{P_{D(W)},\ P_0\}$の値が高く直感と合致する．しかし，このままでは$\#D(W)$が離れたターム(これは概ね，二つのタームの頻度が大きく異なることと等価である)同士のrepresentativenessを適切に比較することができない．なぜならば，一般に$Dist\{P_{D(W)},\ P_0\}$は，$\#D(W)$が大きくなるにつれて増加するからである．実際，「オウム」は「する」と$Dist\{P_{D(W)},\ P_0\}$の値が同程度となる。

3.2 距離の正規化

そこで特定のタームから離れて$Dist\{\cdot,\ P_0\}$の振る舞いを調べるため，さまざまな数の文書をランダムサンプリングし，その結果得られたさまざまな数の単語を含む文書集合Dに対して$(\#D,\ Dist\{P_D,\ P_0\})$を計算し，図２に「×」を用いてプロットした。これらの点は，(0, 0)に始まり$(\#D_0,\ 0)$に終わる一つのなめらかな曲線により良く近似できる．以下，この曲線をベースライン曲線と呼ぶことにする．

$D = \phi$ のときと、$D = D_0$のときに$Dist\{P_D,\ P_0\}$が0となるのは定義から明らかであるが，#D = 0 付近の挙動は，比較的全文書数が少ないとき(2,000文書程度)から，新聞1年分(300,000文書程度)まで，全文書集合が様々な大きさの場合にかなり安定して近似できることが確認できた．

そこで，上記のさまざまな大きさの全文書集合において，ベースライン曲線が指数関数を用いた近似関数を用いて精度良く求められる区間($1000 \leqq \#D < 20000$)上で近似関数$B(\cdot)$を求め，$1000 \leqq \#D(W) < 20000$を満たすWのrepresenentativeness を，$Dist\{P_{D(W)},\ P_0\}$に，$B(\cdot)$による正規化を施した値：

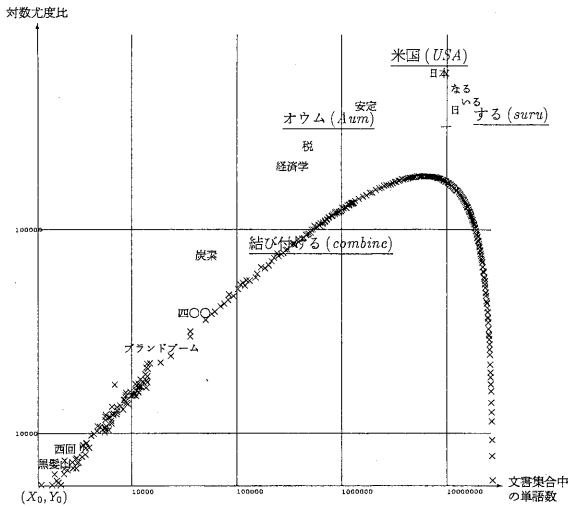
$$Rep(W) = Dist\{P_{D(W)},\ P_0\}/B(\#D(W))$$

により定義する（ただし，ここでいう単語は，記号や助詞，格助詞などの情報検索の検索語として確実に不要とみなされたものはすでに除いたものを指す(これらを含めた場合でも同様の手法が実現できるが，その場合は上記の数字は若干異なってくる).

ここで，「する」のように著しく$\#D(W)$が大きい場合でも，上記のベースライン関数の有効域を用いることを可能にすることと，計算量を低減することを意図して，$20{,}000 < \#D(W)$となるようなWに対しては，$D(W)$として150文書程度をランダム抽出し，$1000 \leqq \#D(W) < 20000$を満たすようにしてから$Rep(W)$を計算する．

一方，上記の区間で求めたベースライン曲線の近似関数は，$\{x \mid 0 \leq x < 1000\}$で，値を大きめに見積もる傾向があるため，$\#D(W) \leqq 1000$となる$W$については，正規化の結果$Rep(W)$は低めに出る．しかし，1000単語はほぼ新聞の２，３記事に相当するが，出現文書数がその程度のタームは我々の目的からの重要度は低いため，そのまま適用した．

ランダムサンプリングした文書集合Dにおける$Dist\{P_D,\ P_0\}/B(\#D)$は，さまざまなコーパスにおいて，安定して平均Avrがほぼ1(± 0.01)，標準偏差σが0.05程度であった．また，最大値が $Avr + 4\sigma$ を越えることはなかったので，あるタームWの$Rep(W)$の値が，「意味のある値である」と判断するための閾値として，$Avr + 4\sigma = 1.20$を設ける．



$X_0 = 1289, Y_0 = 5492.277272\, X_{max} = 26108800, Y_{max} = 901755.795689$

図2 $Dist(P_{D(W)}, P_0)$ と $Dist(P_D, P_0)$

3.3 指標の性質

上記指標$Rep(\cdot)$は，

(1) 数学的な意味付けが明瞭である．
(2) 高頻度タームと低頻度タームの比較ができる．
(3) 閾値の設定が自然にできる．
(4) 任意の長さのタームに対して適用できる．

等の好ましい性質を持つ．*tf-idf* や χ^2 等との最も大きな違いは，それらが「重要度を測りたい単語自体の分布」に着目していたのに対し，提案指標はその語と共に現れる語群に注目する点である．

なお，距離計測の基本的な指標として対数尤度比を用いた理由は，予備実験の結果から，対数尤度比を用いた場合が最も安定した結果が得られたからである．Kullback-Leibler情報量は対数尤度比と同じ序列を与え(偶然でなく，定義より従う)，transition probability も類似した効果を与えるが，vector-space model を用いた場合，$Rep(\cdot)$ の分散が大きく，その振る舞いが少し不安定である．詳しい検討は別の機会にゆずる．

4. 実験結果

4.1 新聞記事中のモノグラムに関する実験

特徴単語グラフにおける単語(モノグラム)選択を念頭に，単語が「検索内容の概観に現われることについてのふさわしさ」を提案指標がどの程度測ることができるかを調べる実験を行った．

4.1.1 実験方法

日経新聞1996年分の，document frequency が3以上の単語(約86000語)から20,000語を無作為抽出し，そのうちの2,000個を，検索内容の概観に現われることが「好ましい」「どちらでもよい」「好ましくない」の3種類に人手で分類した．表1は20000個の単語の一部，表2は2000個の単語の分類の一部である．ここで，"a"は，「好ましい」に対応し，"d"は「好ましくない」に対応する．"p"は，固有名詞，"n"は無意味数字(「2000年問題」の"2000"や「55体制」の"55"のような特殊なものでなく，複数の記事中にたまたま偶然に共通して現れる数字)を示す．

「好ましい」という分類にあたっては，あくまで，ナヴィゲーションのための好ましさを想定し，「ある程度の大きさを持った，いくつかの話題の固まりへとつながる」ような単語を優先する．また，「好ましくない」を選ぶにあたっては，できるだけ慎重な立場を取った．これらはあくまで主観評価であり，現時点で複数人の主観評価を厳密に相対比較したわけではなく，特に「好ましい」と「どちらでも良い」の差は微妙なものではあるが，「好ましくない」ものについてはかなり高い一致度が得られると考えられる．また，「好ましくない」語を取り除く能力は，stop-word list 作成への応用に関しては重要である．

上記20,000語をランダムに並べたときと，何等かの指標でソートしたときで，あるクラスの分類された語が先頭からN位までにいくつ出現するかという累積出現頻度グラフを比較する．

4.1.2 比較の対象

比較の対象として，ランダムソートだけでなく，最も単純な指標としての頻度，及び，2.1で述べた全文書を対象とした*tf-idf*の変形版を用いた．すなわち，

$$tf\text{-}idf = f(W)^{1/2} \times \log\left(\frac{N_{total}}{N(W)}\right),$$

ここで，$f(W)$のかわりに$f(W)^{1/2}$を用いたのは，頻度の寄与を若干押さえるためである．

表1
無作為抽出した20,000個の単語の一部

一億八千六百万	用金	保持	四億九千万
神野	グリーンフィールド	如才ない	片倉
石神	真智子	カルソニック	惑い
アルバータ	音別	外食	雪
聞きつけ	一司	江戸前	ラグ
七万五千	一・三七	伊三夫	木材業
登記所	ファイアウオール	フレームリレー	執拗だ
豊臣	働き続け	危ぶむ	ウィンラック
独身	神事	六万四千八百	ビジネスマネジャー
下関市	八千三百万	マウンテンバイク	積極的
消沈	疑心	嵯峨野	クローネ
助走	引地	排気	シクラメン
脅迫状	筋違いだ	私人	

表2
2,000個の分類された単語の一部

「好ましい」		「どちらでも良い」		「好ましくない」	
アミューズメントパーク	a	ひんやり		八千三百万	dn
登記所	a	消沈		どっと	d
豊臣	ap	助走		千百四十六	dn
リテラシー	a	グリーンフィールド	p	七万五千	dn
マウンテンバイク	a	伊三夫	p	すべて	d
第三セクター方式	a	一司	p	二百五億	dn
ファイアウオール	a	引地	p	六万四千八百	dn
骨董品	a	筋違いだ		四二・八	dn
アトランタ	ap	ネクタイメーカー		多大	d
：		：		：	

4.1.3 結果

図3は，分類が"a"となったものの累積頻度を，ランダム，頻度，*tf-idf*，新指標のそれぞれを用いた場合で比較したものである．人手でしらべた2000個のうち，分類が"a"であったものは876個であったので，図の折れ線は最も理想的な場合の累積頻度である．

グラフから明らかに，ランダム＜頻度＜*tf-idf*＜新指標の順で「好ましい」と分類される語の優先順位を上げる力が強く，頻度，*tf-idf*に比べてあきらかに

有為な改善を示している．この改善の大きさをどう評価するかにはいろいろな立場がありうるが，重要語抽出においては，実は頻度を上回ることはかなり困難であると報告されていることを勘案すれば (Daille et al. 1994, Caraballo et al. 1999)，この結果は充分肯定的なものであると考えられる．

図4は，分類が"d"となったものの累積頻度を，ランダム，頻度，*tf-idf*，新指標のそれぞれを用いた場合で比較したものである．2000個のうち，分類が"d"であったものは452個であったので，折れ線は理想的な場合の累積頻度である．この比較では，新指標の選別能力の優位性がより際だつ．頻度，*tf-idf*はランダムな場合とさして変わらず，これらの指標による高頻度不要語排除の困難さを示している．

図5は，分類が"a"となったものの中で，特に固有名詞について累積頻度を比較したものである．2000個のうち，分類が"ap"であったものは221個であり，この場合には，*tf-idf*とランダムな場合との差はほとんど無く，新指標が突出している．

図6は，分類が"d"となったものの中で，特に無意味数字について累積頻度を比較したものである．2000個のうち，分類が"dn"であったものは160個である．この場合，頻度と*tf-idf*は全くランダムな場合に比べて優位性が見られず，新指標は著しい性能を示している．

4.1.4 分析

以上の結果を細かく調べると，固有名詞のうち，競走馬や力士の名前など特定のジャンルのものの新指標の値が非常に高い等の現象が見られた．これは，日経新聞において，比較的均質な大多数の一般記事と，「相撲の勝敗」，「競馬の勝敗」，「人事記事」，「コラム」などのような，形式，出現単語，文体等に強い特徴のある記事が混在しており，後者のような記事に偏って出現する属する単語に，強くバイアスがかかるためだと思われる．このようなものを除いた場合にどうなるかを調べるのは興味深い．

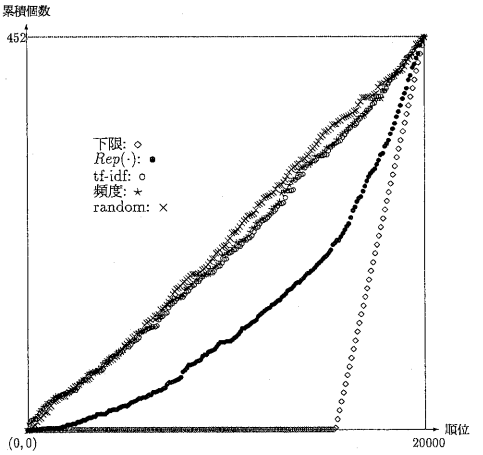


図4 Class-*d* に属する単語のソーティング

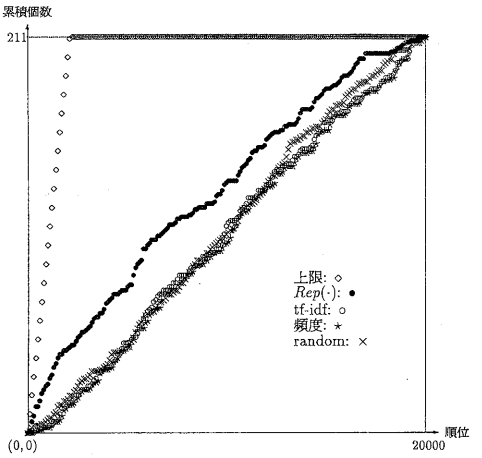


図5 Class-*ap* に属する単語のソーティング

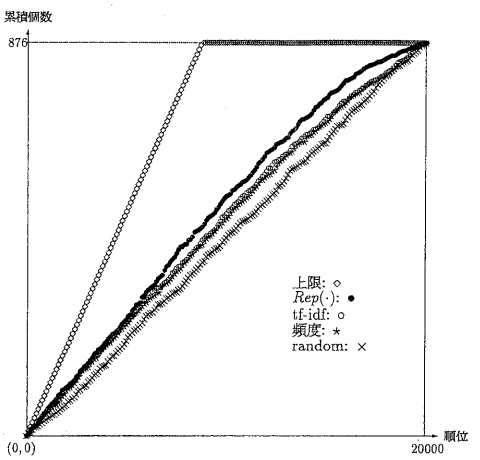


図3 Class-*a* に属する単語のソーティング

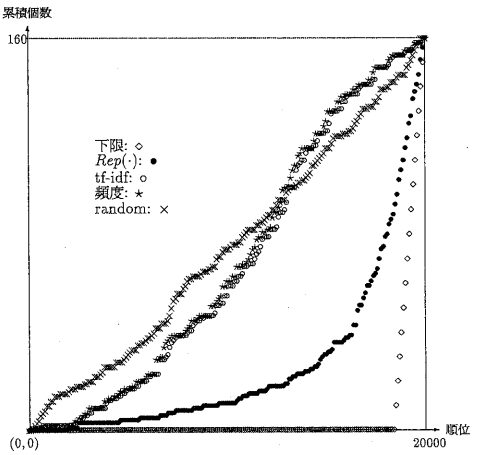


図6 Class-*dn* に属する単語のソーティング

4.2 NACSISコーパスからのターム抽出

4.2.1 実験内容

学術情報センターにより提供して頂いた，人工知能関係の論文のアブストラクト1870件のなかから用語を抽出する実験を行った．まず，コーパスの形態素解析を行い，それにより得られる内容語のモノグラム，バイグラムの集合に対して，さまざまな指標でこれらをソートし，新指標による閾値を使って足切りを行った．閾値以上の指標をもつバイグラムについては，それを含むトライグラムを追加し，同様に足切りを行った．最後に，簡単な文字種・品詞フィルターであきらにタームとして不適切なものをとりのぞいた．

詳しい定量評価は，NACSIS NTCIR Workshop の予稿集にて発表されるのでここでは触れず，以下では従来指標及び新指標によりバイグラムをソートした場合の先頭100位を掲載し，定性的な比較を行う．従来の指標としては頻度，*tf-idf*，相互情報量，対数尤度比(これらについては，Hisamitsu 1998を参照)を選ぶため，以下ではバイグラムに限定して比較する．

4.2.2 抽出結果

表3は，頻度によりバイグラムをソートしたものである．ただし，単語境界は表示しておらず，形態素解析誤りにより誤って単語バイグラムとして捉えられているものもある。コーパスの規模がさほど大きくなく，かつバイグラムに限定しているため，比較的きれいな結果である．実際，頻度は，通常考えられているより「重要語」を抽出するための有効な指標であるが(Daille et al. 1994, Caraballo et al. 1999)，重要な語の頻度が高いのは自然であるので，ある意味でこれは当然の結果であり，頻度より良い結果を得るために本質的なことは，高頻度であっても重要でない語の弁別である．例えば，「本稿」，「本論文」，「本研究」などは，典型的な高頻度不要タームである．

表4は，*tf-idf* によりバイグラムをソートしたものである．取り除くべき，「一般的すぎる言葉」の優先度を下げることに関しては，*tf-idf* はある程度有効であることが分かる．

表5は，相互情報量によりバイグラムをソートしたものである．この例で分かる通り，相互情報量を頻度のそろわない比較対象に直接適用すると，すでに指摘されているように(Dunning 1993)，低頻度バイグラムの過剰評価が生じ，頻度に比べても不適切な結果となる．

表6は，対数尤度によりバイグラムをソートしたものである．これもすでに指摘されていることであるが(Dunning 1993)，相互情報量と異なり，低頻度バイグラムの過剰評価が抑制され，結果は頻度に近づく．「一般的すぎる言葉」の優先度を下げることに関しては，*tf-idf*より若干有効と思われる．

表7は，新指標によりバイグラムをソートしたものである．この例からわかることは，新指標はおおむね粒度のそろった，しかもあまり特殊過ぎないバイグラムを抽出できることである．しかし，人工知能分野においては珍しい，経済関係の内容を扱った比較的長いアブストラクトから抽出されたタームがいくつか出現しており，これらは必ずしも重要度が低いとは限らないが，他のバイグラムとは異質である．本格的な用語抽出を行う場合は，新指標単独で無く，頻度情報を反映した何らかの指標と組み合わせる必要を示唆している．

表8は，上記の考察に基づき，対数尤度比によりバイグラムをソートし，さらに新指標による閾値を下回るものをとりのぞいたものである．対数尤度比自体，基本的には頻度を反映しつつ，不要語除去の観点からは比較的好ましい振る舞いをするが，新指標による不要語除去と組み合わせることにより，表6よりかなり好ましい結果となっている．

4.3 考察

新指標は，representativeness の高いタームを集めるという観点からは，フォーマットや長さのそろった文書群を対象とするならば，かなり粒度のそろった適切なタームを選択する力があると期待される．

逆に，representativeness の低いタームを排除するという観点からは，高頻度不要語の排除という点で，きわめて効果的であり，stop-word list の作成支援に寄与するところが大きいと考えられる．

表3
頻度を用いたときの先頭100位

本稿 594	動的 73	自然言語 55	解決過程 39
学習者 496	対象モデル 72	学習アルゴリズム 55	機械翻訳 37
問題解決 445	相互作用 72	ベース推論 55	支援環境 36
本論文 420	ＣＡＩシステム 71	定性推論 54	思考過程 36
本研究 390	音声認識 70	故障診断 54	最適解 36
知的 243	論理プログラム 69	因果関係 54	一階 36
知識ベース 229	類似度 69	強化学習 50	知識処理 35
支援システム 213	定式化 68	構造化 49	機能モデル 35
有効性 166	自動的に 68	設計過程 47	目的 34
本システム 142	推論システム 63	教材知識 47	対象世界 34
知識表現 133	時間 63	自動生成 46	多項式時間 34
知識獲得 127	決定木 62	学習環境 46	述語論理 34
再利用 100	設計対象 61	曖昧性 44	本方式 33
ＧＡ 99	教育システム 61	利用者 44	知識ベースシステム 33
本手法 97	学習システム 60	背景知識 44	設計問題 32
事例ベース 95	本報告 59	制約充足 44	制約条件 32
遺伝的 90	人工知能 59	実験結果 44	熟練者 32
仮説推論 89	エージェント間 59	高速化 44	自動化 32
対話システム 87	設計支援 58	概念設計 44	構成要素 32
類似性 85	言語処理 58	構文解析 43	処理システム 31
音声対話 83	帰納的 58	機械学習 43	有用性 30
設計者 78	オブジェクト指向 58	設計知識 42	充足問題 30
最適化 77	定性的 57	法的 41	協調問題 30
意思決定 76	論理式 55	学習支援 41	理解状態 29
モデル化 76		情報処理 39	帰納推論 29
帰納学習 75			

表4

*tf-idf*を用いたときの先頭100位

学習者 496	音声認識 70	オブジェクト指	統合関係 20
問題解決 445	類似度 69	向 58	思考過程 36
知識ベース 229	対象モデル 72	定式化 68	本報告 59
知的 243	設計者 78	概念設計 44	機械翻訳 37
ＧＡ 99	論理式 55	故障仮説 23	一階 36
仮説推論 89	ＣＡＩシステム	設計知識 42	熟練者 32
支援システム	71	機能モデル 35	自動生成 46
213	三面図 28	学習アルゴリズ	利用者 44
決定木 62	法的 41	ム 55	知識コミュニテ
知識獲得 127	本手法 97	構造化 49	ィ 24
意思決定 76	相互作用 72	ベース推論 55	物理現象 28
事例ベース 95	制約充足 44	時間 63	制御知識 25
知識表現 133	強化学習 50	言語処理 58	言語モデル 26
再利用 100	エージェント間	背景知識 44	意味素 20
類似性 85	59	グループ学習 28	解決過程 39
遺伝的 90	教材知識 47	構文解析 43	学習支援 41
本システム 142	本稿 594	設計過程 47	統語 28
本研究 390	動的 73	人工知能 59	情報処理 39
帰納学習 75	モデル化 76	学習システム 60	文字列 27
対話システム 87	教育システム 61	自然言語 55	単一化 26
設計対象 61	定性推論 54	戦略知識 25	学習効果 26
音声対話 83	故障診断 54	曖昧性 44	多項式時間 34
本論文 420	因果関係 54	高速化 44	エージェント組
最適化 77	推論システム 63	自動的に 68	織 23
論理プログラム	定性的 57	充足問題 30	空間的 24
69	設計支援 58	学習環境 46	最適解 36
有効性 166	帰納的 58	機械学習 43	協調問題 30

表5

相互情報量を用いたときの先頭100位

連成 1	照合点 1	荷役ヤード 1	背腹 1
履修科目 1	小破断 1	科目届 1	背景色 1
落射 1	従属文 1	演劇経験者 1	東大工学部 1
有人観測所 1	秋葉三尺 1	渦巻ポンプ羽根	鉄道台車 2
免疫ワードスポ	射照明 1	車 1	提灯チョウチン
ッティング 1	似顔絵師 1	印加 1	2
魔法陣 1	資金使途 1	意見 1	鳥類図鑑 2
付属テープ 1	残り体力 1	伊勢神宮 1	地区住民 1
瀬出し 1	三和銀行 1	ツルカメ算 1	大阪府高専 1
姫高原 1	三尺坊 1	タイル取り 1	大阪大学溝口 1
費補助金 1	埼玉県立 1	サービス業務 1	相似異同 1
売土 1	才児 1	オーストラリア	接続し実感 1
電動ウインチ 1	黒姫 1	国立 1	製薬業 1
通謀虚偽 1	高級幹部 1	お札降り 1	数箇市町村 1
長野県黒 1	公認会計士 1	あき缶分別 1	条通謀 1
中和滴定 1	現実感 14	利用法 9	証券取引所 1
地中ライフライ	原油安 1	有価証券 1	小売業 1
ン 1	県立久喜 1	輸出立国 1	受容器 2
断冷却材 1	空気ダンパ 1	野外テント 1	手書き帳 2
大阪府立 1	京都大学西田 1	模範文例 1	取り 1
耐荷 1	久喜北 1	鳴き真似 2	主査溝口 1
川崎製鉄 1	喜田二郎 1	未定乗数 1	自走 1
川喜田 1	缶コーヒー 1	北陽 1	指守 1
水圧鉄管 1	冠婚葬祭 1	府立高専 1	索引付け 2
人称語尾 1	改良策 1	筆耕テクスト 1	座標点 2
深部圧覚 1		被災地 2	混合音 1
蒸留塔 1		発着信 1	

表6

対数尤度を用いたときの先頭100位

本稿 594	動的 73	帰納的 58	熟練者 32
学習者 496	音声認識 70	モデル化 76	有用性 30
問題解決 445	論理プログラム	言語処理 58	文脈自由 24
本論文 420	69	高速化 44	学習アルゴリズ
本研究 390	論理式 55	機械翻訳 37	ム 55
知的 243	故障診断 54	定性的 57	自己組織化 24
知識ベース 229	時間 63	統語 28	話し言葉 19
有効性 166	本システム 142	本報告 59	試行錯誤 18
支援システム	帰納学習 75	最適解 36	設計過程 47
213	因果関係 54	ベース推論 55	設計支援 58
再利用 100	自然言語 55	教育システム 61	構成要素 32
相互作用 72	制約充足 44	項目 27	対象世界 34
意思決定 76	本手法 97	設計対象 61	評価値 17
決定木 62	曖昧性 44	ＧＡ 99	知識ベースシス
事例ベース 95	対話システム 87	文字列 27	テム 33
知識獲得 127	設計者 78	自動生成 46	終端 16
人工知能 59	構文解析 43	入出力 25	可視化 28
遺伝的 90	一階 36	背景知識 44	構造化 49
オブジェクト指	エージェント間	思考過程 36	極小限定 20
向 58	59	教材知識 47	解決過程 39
仮説推論 89	定性推論 54	物理現象 28	制約条件 32
類似度 69	自動的に 68	実験結果 44	定理証明 22
最適化 77	対象モデル 72	不適格 20	非線形 22
類似性 85	強化学習 50	運転員 23	巡回セールスマ
音声対話 83	三面図 28	述語論理 34	ン 15
知識表現 133	創発 26	現実感 14	単一化 26
定式化 68	多項式時間 34		機械学習 43

表7

新指標を用いたときの先頭100位

学習者 496	仮説推論 89	方針決定 1	技術事業 1
ＧＡ 99	事例ベース 95	独創的 1	技術開発 3
遺伝的 90	支援システム	調査システム 1	技術シーズ 1
先進工業 3	213	第４報 1	基本業務 3
先行開発 2	意思決定 76	総合明確 1	基本基調 1
製品化 3	機械翻訳 37	先駆製品 1	基調的な 1
新製品 7	生産システム 3	生存基盤 1	開発途上国指導
新技術 7	多項式時間 34	水準高 1	1
事業化 3	信念管理 7	資源無 1	開発目標 1
工業国 3	地球環境 15	仕様明確 1	開発市場 1
空洞化 2	管理構造 7	昨年末 1	開発基本 3
故障診断 54	文脈自由 24	根本的 1	回復感 1
試行研究 4	製造業界 2	国民多 1	マスコミ的 1
論理プログラム	問題解決 445	国化 1	ニーズ創造 1
69	知的 243	高国民 1	シーズ創造 1
音声対話 83	一次変電所 16	構造改革 1	グローバル化 1
連想検査 4	再利用 100	顧客ニーズ 1	対象モデル 72
遠隔性 8	運転員 23	原油安 1	参加者 23
音声認識 70	統合関係 20	経済復興 1	相対位置 3
エージェント組	学習アルゴリズ	景気回復 1	自由文法 20
織 23	ム 55	金融業界 1	産業界 5
分散信念 6	ＣＡＩシステム	業務明確 2	最適解 36
エージェント間	71	業務総合 1	充足問題 30
59	因果関係 54	教育水準 1	設計過程 47
本枠組み 9	類似度 69	技術標準 1	平均的 7
最適化 77	ベース推論 55	技術創造 1	ファジィ理論 12
対話システム 87	輸出立国 1	技術先行 1	

表8
尤度比＋新指標の先頭100位

学習者 496	制約充足 44	現実感 14	見え方 11
問題解決 445	曖昧性 44	熟練者 32	実時間 21
知的 243	対話システム 87	文脈自由 24	直観主義 11
知識ベース 229	設計者 78	学習アルゴリズム 55	地球環境 15
有効性 166	構文解析 43	設計過程 47	人工生命 12
支援システム 213	一階 36	設計支援 58	ブール代数 11
再利用 100	エージェント間 59	評価値 17	適格文 14
相互作用 72	定性推論 54	制約条件 32	隣接 9
意思決定 76	対象モデル 72	巡回セールスマン 15	自由発話 16
決定木 62	強化学習 50	正例 24	決定支援 26
事例ベース 95	三面図 28	一次変電所 16	変電所事故 11
人工知能 59	創発 26	推論システム 63	ストリーム分離 10
遺伝的 90	多項式時間 34	自由文法 20	学習支援 41
仮説推論 89	帰納的 58	充足問題 30	マルコフ連鎖 9
類似度 69	言語処理 58	参加者 23	セールスマン問題 15
最適化 77	機械翻訳 37	異常診断 22	ゲーム木 12
音声対話 83	統語 28	学習環境 46	ベース構築 23
動的 73	最適解 36	線形関数 20	刺激語 7
音声認識 70	ベース推論 55	定性的な 36	解析法 22
論理プログラム 69	項目 27	各エージェント 27	近似解法 11
論理式 55	設計対象 61	エージェント組織 23	統合関係 20
故障診断 54	ＧＡ 99	ＣＡＩシステム 71	空間内 15
時間 63	文字列 27	多重線形 15	認識率 14
因果関係 54	背景知識 44		古典論理 11
自然言語 55	不適格 20		所属性 12
	運転員 23		
	述語論理 34		

5. おわりに

本報では，タームの"representativeness"をはかるための指標を導入した．基本的な考え方は，「その単語と文書内共起する単語の集合」の偏りを測ることでその語の特徴度をはかろうとするものである．この方法は単純なだけに数学的な意味付けが明瞭であり，閾値が自然に定まるだけでなく，高頻度語と低頻度語を無理なく比較できる等の特徴がある．

予備的な実験によれば，新指標は不要語の同定にとりわけ有効であり，不要語リストの自動生成，文献類似検索精度の向上に役立つと思われる．用語抽出に用いる場合，一様性の高いコーパスからは，粒度のそろったコアタームを抽出する能力があると考えられる．

謝辞

representativenessの計算にあたって，単語と文書の対応付けの高速化のため，日立製作所中央研究所の西岡真吾研究員の開発されたプログラムを利用させて頂いた．ここに感謝致します．

参考文献

Caraballo, S. A., Charniak, E. (1999). Determining the specificity of nouns from text. *Proc. of WVLC'99*

Church, K. W., and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics* 6(1), pp.22-29.

Daille, B. and Gaussier, E., and Lange, J. (1994). Towards automatic extraction of monolingual and bilingual terminology. *Proc. of COLING'94*, pp.515-521.

Dunning, T. (1993). Accurate Method for the Statistics of Surprise and Coincidence, *Computational Linguistics* 19(1), pp.61-74.

Firth, J. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, Philological Society, Oxford. (1957).

Frantzi, K. T., and Ananiadou, S., and Tsujii, J. (1996). Extracting Terminological Expressions, *IPSJ Technical Report of SIGNL*, NL112-12, pp.83-88.

Kageura, K. and Umeno, B. (1998). Methods of automatic term recognition: A review. *Terminology* 3(2), pp.259-289.

Kita, Y. Kato, Y., Otomo, T., and Yano, Y. (1994). A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria, *Journal of Natural Language Processing* 1(1), pp.21-33.

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development 2(2), pp.159-165.

長尾真，水谷幹男，池田浩之 (1976). 日本語文献における専門用語の自動抽出，情報処理学会論文誌 17(2), pp.110-117.

Nakagawa, H. and Mori, T. (1998). Nested Collocation and Compound Noun For Term Extraction, *Proc.of Computerm'98*, pp.64-70.

Nishioka, S., Niwa, Y., Iwayama, M., and Takano, A. (1997). *DualNAVI*: An information retrieval interface. *Proc. of WISS'97*, pp.43-48, 1997. (in Japanese)

Niwa, Y., Nishioka, S., Iwayama, M., and Takano, A. (1997). Topic graph generation for query navigation: Use of frequency classes for topic extraction. *Proc. of NLPRS'97*, pp.95-100.

Hisamitsu, T. and Niwa, Y. (1998). Extraction of Useful Terms from Parenthetical Expressions by Using Simple Rules and Statistical Measures - A Comparative Evaluation of Bigram Statistics -Proceedings of COMPUTERM'98, pp. 36-42.

Salton, G. and Yang, C. G. (1973). On the Specification of Term Values in Automatic Indexing, Journal of Documentation 29(4), pp.351-372.

Sparck-Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval, Journal of Documentation 28(1), pp.11-21.