論 文

Technical Papers

語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム

Keyword Extraction from a Document using Word Co-occurrence Statistical Information

松尾 豊
Yutaka Matsuo
東京大学工学系研究科†1
Graduate School of Engineering, University of Tokyo
matsuo@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~matsuo/

石塚 満
Mitsuru Ishizuka
東京大学情報理工学系研究科
School of Information Science and Technology Engineering, University of Tokyo
ishizuka@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~ishizuka/

keywords: keyword extraction, word co-occurrence, $\chi^2 test$

Summary

We present a new keyword extraction algorithm that applies to a single document without using a large corpus. Frequent terms are extracted first, then a set of co-occurrence between each term and the frequent terms, i.e., occurrences in the same sentences, is generated. The distribution of co-occurrence shows the importance of a term in the document as follows. If the probability distribution of co-occurrence between term a and the frequent terms is biased to a particular subset of the frequent terms, then term a is likely to be a keyword. The degree of the biases of the distribution is measured by χ^2-measure. We show our algorithm performs well for indexing technical papers.

1. まえがき

キーワード抽出は，文書検索，Web ページ検索，文書クラスタリング，要約文抽出など，情報検索において重要な技術である．適切なキーワードを自動的に抽出することができれば，読むべき文書を選択しやすくなったり，文書間の関係を把握することが容易になるなどのメリットがある．大量のコーパスを背景とした情報検索を目的とするインデキシングには，tf·idf をはじめ様々な手法が用いられている．

一方で，近年ますます多くの電子的な文書が蓄えられるにしたがって，その文書の内容を大まかに把握するという目的でのキーワード抽出も重要になっている．例えば，あるひとつの文書がどういった内容であるかを知りたいときに，類似の文書を大量に必要とするようなキーワード抽出法を用いることはできない．また，Web ページはその多様性により適切なコーパスを集めることが難しく，文書単独でのキーワード抽出の手法が必要とされる．

コーパスを利用することなく，ひとつの文書だけからキーワードを抽出するには，語の出現頻度を用いる方法 [Luhn 57] や「要するに」などの手がかり語をもとにキーワードを抽出する方法 [Edmundson 69, 木本 91] などがある．しかし，前者は単純すぎて一般的な語も抽出してしまうし，後者は汎用性がない．本論文では，対象とする文書だけの情報から，語の共起をもとに統計的な指標を用いキーワードを抽出する一般的な手法を提案する．まず，対象とする文書の頻出語を取り出し，その頻出語と各語の共起頻度を求める．この共起頻度がどのくらい偏っているかを，その語が重要語であるかどうかの指標として用いる．単一の文書だけから手軽に，比較的高い精度でキーワードを取り出すことができるのが大きな特徴である．

2 章で手法の概要，3 章で詳細について述べ，4 章で評価を行う．5 章で関連研究と議論を記す．

2. 語の共起と重要語

文書中に出現する単語は，文毎に句点やピリオドによって区切られている．以下では，同文中に出現する 2 つの語は 1 回共起していると考える．すなわち，各文をひとつの「バスケット」として捉え，(n-gram 処理以外の）順序関係については考慮しない．

さて，ひとつの文書が与えられたとき，語の出現頻度を数えることで，頻出語を取出すことができる．ここで

†1 現在，産業技術総合研究所サイバーアシスト研究センター．

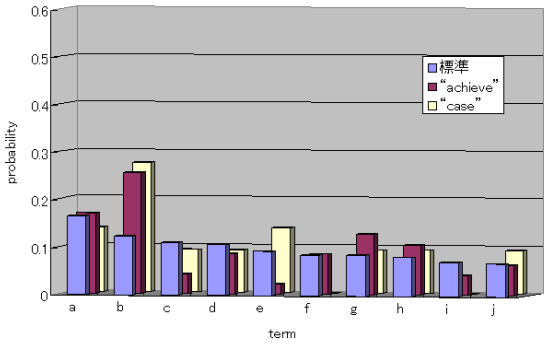

図 1 語 *“achieve”, “case”* の頻出語との共起の確率分布

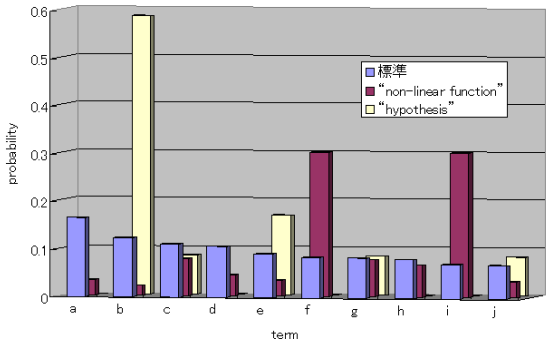

図 2 語 *“non-linear function”, “hypothesis”* の頻出語との共起の確率分布

語とは，単語もしくは複数の単語からなるフレーズである．例として，ある文書 [Ishizuka 98] を取り上げよう．表 1 に，頻出語上位 10 個の出現頻度と出現確率（全体が 1 になるように正規化したもの）を示す．語の出現頻度は，経験的に Zipf の法則 (順位 r と頻度 f の積が定数 C になる (第 1 法則))[徳永 99] に従うことが知られている．

次に，語の共起の頻度を集計することにより，表 2 のような共起行列を作ることができる（表形式で表している）．この行列は，例えば語 a と語 b は 22 文で共起していることを示している．共起行列は，文書中に出現する語の数を N とすると $N \times N$ の対称行列であるが，ここでは頻出語上位 10 語 (G とする）に対応する列だけを抜きだし，$N \times 10$ 行列としている．対角成分は，ここでは定義しない．

仮に，語 w が頻出語 $g \in G$ と全く独立に生起するなら，語 w と語 $g \in G$ が共起する確率は表 1 の確率と同様の分布になるはずである．一方，語 w と頻出語 $g \in G$ の間に何らかの意味的なつながりがあれば，この確率は偏ることになる．

図 1，図 2 に，いくつかの語と語 $g \in G$ との共起確率[*1]の分布を示す．図中に標準として，語 $g \in G$ の単独での出現頻度の分布（表 1）を示している．“achieve” や “case” などの語は，どの頻出語 $g \in G$ とも偏りなく用いられるのに対し，“non-linear function” や “hypothesis” などの語は特定の頻出語と選択的に多く共起している．このような偏りは，筆者が意味的なつながりを考慮し文書を書き進めていく上で生まれたものであり，分布が偏っている語は文書中において何らかの重要な意味を担っている語であると考えられる．実際，もとになった論文の主旨は「仮説推論を非線形関数に置き換え，探索を行うことにより解を得る手法．局所最適点に陥ると変数を真に固定することにより脱出を行う.」であり，“non-linear function” や “hypothesis” などの語は，論文中で重要な語である．

したがって，ある語 w の頻出語 $g \in G$ に対する共起確率が，頻出語単独での出現確率からどのくらい偏りがあるか測れば，その語の重要度を表す指標になると考えられる．しかしながら，語の出現頻度自体が少なければ確率分布の偏りは信頼できない．例えば，表 1 から語 a の出現確率は 0.167 であるが，出現回数 1 回の語 w_1 が語 a とだけ 1 回（つまり確率 1 で）共起していることよりも，出現回数 10 回の語 w_2 が語 a とだけ 10 回（つまり確率 1 で）共起している方が，より確実に偏っているといえるだろう．このように統計的に有意なずれを評価するために，分布の偏りを検定する方法として一般的である χ^2 検定（例えば [東京 91] を参照）を用いる[*2]．すなわち，ひとつひとつの語について，各頻出語との共起頻度を標本値とし,「$g \in G$ の出現する確率は語 w の出現いかんに関わらず等しい」を帰無仮説として検定を行えばよい．

頻出語単独での生起確率（表 1）を理論確率 $p_g(g \in G)$ とし，語 w と頻出語群 G の共起の総数を n_w，語 w と語 $g \in G$ の共起頻度を $freq(w,g)$ とすると，統計量 χ^2 は以下の式で与えられる．

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,g) - n_w p_g)^2}{n_w p_g} \qquad (1)$$

$\chi^2(w) > \chi^2_\alpha$ であれば，帰無仮説が有意水準 α で棄却される．(χ^2_α は通常 χ^2 分布表より得る.) $n_w p_g$ は，語 w と語 g の共起する期待頻度を表す．したがって，$\chi^2(w)$ の大きな語 w が理論確率分布からのずれが大きな語である．なお，本稿では χ^2 値を検定法としてではなく，単純に偏りの程度を示す指標，度合いとして用いている．

表 3 に，前述の例 [Ishizuka 98] に対する χ^2 値の高い語と低い語を示す．(出現頻度の上位 10 語を G としている．) 表から分かる通り，χ^2 値の高い語は論旨に関係の深い語が並んでおり，χ^2 値の低い語は一般的な語である傾向が強い．

すなわち，本手法はまず，頻出語を取り出すことによっ

*1 合計が 1 になるように正規化したもの

*2 2 つの分布のずれを検出するには，例えば 3･2 節の Kullback-Leibler divergence などを用いることもできるが，χ^2 検定が最もシンプルで一般的であるため，ここでは χ^2 検定を用いた．

表 1 頻度と確率分布

頻出語	a	b	c	d	e	f	g	h	i	j	計
頻度	79	59	53	51	44	41	41	39	34	33	474
出現確率	0.167	0.124	0.112	0.108	0.093	0.086	0.086	0.082	0.072	0.070	1.0

a: *method*, b: *solution*, c: *variable*, d: *problem*, e: *node*, f: *non-linear*, g: *search*, h: *point*, i: *function*, j: *local*

表 2 共起行列

	a	b	c	d	e	f	g	h	i	j	計
a	—	22	5	15	2	11	15	10	4	4	88
b	22	—	9	9	4	6	16	11	6	7	90
c	5	9	—	4	8	7	9	8	9	8	67
d	15	9	4	—	2	9	2	2	5	3	51
e	2	4	8	2	—	3	3	1	3	2	28
f	11	6	7	9	3	—	10	10	27	5	88
g	15	16	9	2	3	10	—	19	10	15	99
h	10	11	8	2	1	10	19	—	8	17	86
i	4	6	9	5	3	27	10	8	—	5	77
j	4	7	8	3	2	5	15	17	5	—	66
…	…	…	…	…	…	…	…	…	…	…	…
u	3	2	7	4	3	27	7	6	27	3	89
v	0	7	1	0	2	0	1	0	0	1	12
w	8	12	2	4	1	4	6	5	2	3	47
x	3	6	2	2	3	0	2	2	0	2	22

u: *non-linear function*, v: *hypothesis*, w:*achieve*, x:*case*

て文書自身の全体的な傾向を求め，この傾向から大きく逸脱する特徴を持つ語をキーワードとして取り出す．

3. アルゴリズムの詳細

本手法の大略は前節の通りであるが，単一の文書だけからキーワードを抽出するわけであるから，キーワードの精度を上げるために，さまざまな工夫が必要である．本節では，予備実験に基づき，いくつかのアルゴリズムの改良を示す．

3·1 χ^2 値の計算について

文書中の文の長さは様々であり，長い文に出現する語は他の語と共起しやすく，短い文に出現する語は他の語と共起しにくい．共起の範囲を一文としているので，文の長さが長ければそれだけ他の語と共起する確率は増えることになる．逆に，短い文にも関わらず共起しているときには，その関係はより強いと考える方が自然であろう．したがって，以下のような変更を行う．

- p_g を，(g が出現する文数)/(G 中の語が出現する延べ文数) ではなく，(g が出現する文の語数の合計)/(文書全体の語数の合計) とする．
- n_w を，語 w が出現する文の語数の合計とする．

式 (1) は，語 g の生起確率 p_g に，語 w と頻出語群との共起の総数 n_w を乗じて期待頻度を求めているのに対し，ここでは，文書中の任意の一語が g と共起している確率 p_g に，語 w と共起する語の総数 n_w を乗じて共起の期待頻度としている．この変更により，文の長さを考慮した，より正確な計算結果が得られる．なお，G に含まれる語についても χ^2 値の計算を行うが，その際には自分自身との共起は計算に含めない．

また，頻出語中の特定の一語 $g \in G$ とだけ共起する語は χ^2 値は高くなるが，重要な語であるというより，語 g に付随する語である場合がほとんどである．例えば，前述の例 [Ishizuka 98] では，“child” や “parent” は “node” とだけ選択的に共起するため，χ^2 値は高くなる．これは，論文中で “child node”，“parent node” という決まった形で使われるためであるが，“child” や “parent” が重要な語かというと，そうではない．こういった語は，仮に “node” が G に含まれないとすると，χ^2 値は急に低くなる．そこで，分布の偏りをロバストに算出する目的で，χ^2 値の最大の項を除いた値

$$\chi'^2(w) = \chi^2(w) - \max_{g \in G} \left\{ \frac{(freq(w,g) - n_w p_g)^2}{n_w p_g} \right\}$$

を重みづけの関数として用いる．

3·2 語のクラスタリング

本来，共起行列は $N \times N$ 行列であるが，本手法では頻出語の集合 G に対応する列を抜きだしている．これは，出現頻度の低い語は，正確な出現確率を得ることが難しく，標本確率を理論確率 p_g とすることのデメリットが大きいためである．

しかしながら，各語の χ^2 値は G との共起頻度をもとに計算されるので，G（つまり抜きだす列）を適切に定めることがキーワード抽出の性能をあげる上で極めて重要である．ここで，ある頻出語 $g1$ と $g2$ が互いによく共起するのであれば，語 w が $g1$ と共起していれば $g2$ とも共起するのは当然であろう．そこで，G 中の語をクラスタリングする，つまり共起行列の列をまとめる処理を行う．

文書中の語をクラスタリングする研究は数多く行われているが，大別すると次の 2 つに分けられる．

類似性 語 w_1 と語 w_2 の他の語との共起の分布が似ていれば，同じクラスタとする．

共起 語 w_1 と語 w_2 が頻繁に共起していれば，同じクラスタとする．

表 4 は共起行列から 2 つの列を抜き出したものであるが，前者は 2 つの列の太字部分に着目し，後者は斜体字部分に着目していることに相当する．

類似性によるクラスタリングでは，例えば”Sunday”，“Monday”，“Tuesday”，... や，“build”，“establish”，“found” など，同じような働きをする語が同じクラスタとなる．我々の予備実験では，言い替えを行っている語や，“shortest path” と “path” のように，フレーズとその要素語をひとつのクラスタとする傾向が多く見られた．2 つの分布の類似性は，Kullback-Leibler divergence や Jensen-Shanon divergence*3といった統計量により計ることができる [Dagan 99]

．

表 3 χ^2 値が高い語

順位	χ^2	ラベル	出現頻度
1	76.7	non-linear function	27
2	36.7	local optimal point	11
3	34.3	true	24
4	33.1	initial search point	9
5	30.6	false	19
6	24.2	hypothesis	32
7	23.1	fix	26
8	22.9	trap	7
9	22.8	element	15
10	22.4	goal	11
⋮	⋮	⋮	⋮
190	3.6	base	3
191	3.0	approach	3
192	3.0	resolve	4
193	3.0	reach	3
194	2.7	find	16
195	2.6	systematically	3
196	2.6	complex	3
197	2.5	human	3
198	2.2	represent	3
199	1.9	call	4
200	0.0	several	3
201	0.0	even	3
⋮	⋮	⋮	⋮
303	0.0	according	3

文書中に 3 回以上出現する語についての結果である．実際のアルゴリズムでは，出現頻度の極端に少ない語は χ^2 値も低くなるのでこのような閾値を設定する必要はない．

*3 語 w_1 と語 w_2 の Jensen-Shanon divergence は，以下で表される．

$$J(w1,w2) = \log 2 + \frac{1}{2}\sum_{w'\in C}\Big\{h(P(w'|w_1)+P(w'|w_2)) - h(P(w'|w_1)) - h(P(w'|w_2)\Big\}$$

ただし，$h(x) = -x\log x$，$P(w'|w_1) = freq(w', w_1)/freq(w_1)$ である．

表 4 2 つの列（を転置したもの）

	a	b	c	d	e	f	g	h	i	j	...
f	**11**	**6**	**7**	**9**	**3**	—	*10*	**10**	**27**	**5**	...
g	**15**	**16**	**9**	**2**	**3**	*10*	—	**19**	**10**	**15**	...

一方，共起によるクラスタリングは，関連のある語が同じクラスタとなる．ひとつひとつの語の意味は異なっても，クラスタ全体としてひとつの概念を表しやすい．例えば，“doctor”，“nurse”，“hospital” などがクラスタとなる [Tanaka-Ishii 96]．共起頻度 $freq(w_1, w_2)$ や相互情報量*4 を用い，関連の強さを計ることができる [Church 90, Dunning 93]．

本研究では，両方のクラスタリングを用いた．まず，類似性によるクラスタリング（Jensen-Shannon divergence を用いる）で同義語をまとめていき，さらに共起性によるクラスタリング（相互情報量を用いる）で関連のある語を同一クラスタとした．適切なクラスタリングにより，互いに（ある程度）独立なクラスタに対して χ^2 値が計算されることになる．

3·3 アルゴリズム

具体的なアルゴリズムを示す．閾値は予備実験により定め，次節の評価実験で用いた値を示している．

(1) 前処理：英語の場合には，stemming を行う．日本語の場合には，形態素解析を行い*5，分かち書きをする．さらに，フレーズを取り出す*6．*stop word* が与えられている場合には，これを取り除く．

(2) 頻出語の選択：文書中の語の延べ総数 N_{total} の 30%に達するまで頻出語の上位語を取り出す．

(3) 頻出語のクラスタリング：頻出語間の類似度の特徴量（Jensen-Shanon divergence）が閾値 $(0.95 \times \log 2)$ を越えるものは，クラスタとしてまとめる．共起の相互情報量が閾値（$\log(2.0)$）を越えるものも，クラスタとしてまとめる．得られたクラスタ群を C とする．以下，クラスタ $c \in C$ との共起とは，クラスタ中のいずれかの語との共起を指す．

(4) 理論確率の計算：クラスタ $c \in C$ と同文中で共起する語の延べ総数 n_c を調べ，理論確率 $p_c = n_c/N_{total}$ を求める．

*4 語 w_1 と語 w_2 の相互情報量は，以下で表される．ただし，N は語の総数である．

$$M(w_1,w_2) = \log\frac{P(w_1,w_2)}{P(w_1)P(w_2)} = \log\frac{N freq(w_1,w_2)}{freq(w_1)freq(w_2)}$$

*5 取り出す品詞は，基本的に，名詞，動詞，形容詞，副詞，未知語である．ただし，代名詞や非自立語は除くなどの設定を行っている．

*6 Apriori 的な手法により，出現回数が 3 回以上の 4-gram までのすべてのフレーズを取り出す [Fürnkranz 98]．

(5) χ^2 値の計算：すべての語 w について，w と $c \in C$ との共起頻度 $freq(w,c)$ と，w が出現する文の語の総数 n_w を求める．χ'^2 値を以下により求める．

$$\chi'^2(w) = \sum_{c \in G} \left\{ \frac{(freq(w,c) - n_w p_c)^2}{n_w p_c} \right\} - \max_{c \in G} \left\{ \frac{(freq(w,c) - n_w p_c)^2}{n_w p_c} \right\}$$

(6) キーワードの提示： χ'^2 値の上位語を一定数提示する．

4. 評　　価

本手法の実行例を示す．読者に分かりやすいように，対象とする文書はこの論文自身である．頻出語の上位は表 5 の通りである．頻出語の上位 18 語（累計で総語数の 30%に達する）をクラスタリングすると，表 6 となる．これらのクラスタに対して，χ^2 値を計算したものが表 7 である．表中の “+” はフレーズを表している．「キーワード抽出」や「χ^2 値」といった語が，上位にきていることが分かる．

さて，本手法のキーワード抽出の精度を評価するために，評価実験を行った．評価実験は，人工知能の分野の 7 著者 20 論文に対して行い，tf，tf·idf[*7]，KeyGraph[*8]と比較した．各手法でキーワード 15 個を出力し，各手法から得られたキーワードの上位語を混ぜてシャッフルし，著者に「論文を構成する重要な概念を表すと思う語にチェックをして下さい」という質問を行った．各手法による出力語中でキーワードであると判定された割合が precision である．さらに，「提示した全ての語（提示した以外の語でも覚えているものがあれば含めてよい）のうち，論文中で不可欠な概念を表す語 5 つ以上を選び A，B，C，D，E と印をつけ，それと同義の語にも同じ印をつけてください」という指示を行った．5 つ（以上）の概念のうち各手法で提示した語にいくつ含まれているかで coverage を測定した．

結果を表 8 に示す．どの手法も *precision* が 0.5 前後であるが，*coverage* は tf や KeyGraph よりも高く，大量のコーパスを必要とする tf·idf に匹敵する性能が得られている．また，tf や tf·idf は文書中でよく出てくる語の重みを大きくするので，出てくる語は当たり前の語が多い．それに対し，本手法では出現頻度が少なくても重要な語を取り出している．それを数値化したものが frequency index で，これは提示した語の出現頻度の平均を表している．tf は文書中に平均して 30 回近く出現する語を提示しているのに対し，本手法は平均 11.5 回出現する語を提示しており，それでいて同程度の precision を得ているところは評価できるだろう．また，上位 15 位までの語を対象とした場合に本手法の precision は 0.51 だが，これを上位 10 語までとすると precision は 0.52 に，上位 5 語では 0.60 に，上位 2 語では 0.72 となる．したがって，χ^2 値の値をキーワードの優先度とすることが可能である．

表 5 本論文に対しての頻出語

順位	頻度	ラベル
1	147	語
2	50	共起
3	36	文書
4	29	出現
5	25	キーワード
6	23	値
7	23	手法
8	22	頻出
9	21	中
10	21	χ^2

表 6 頻出語上位 18 個のクラスタリング

C1: 語，共起
C2: 文書
C3: 出現
C4: キーワード，抽出
C5: χ^2，値
C6: 手法
C7: 頻出，語+共起，頻出+語
C8: 中
C9: 確率
C10: 用いる
C11: 行う
C12: クラスタ
C13: 頻度

本手法では，フレーズもキーワードとして抽出するが，表 9 にフレーズの含まれる割合とフレーズを除いた場合の結果について示す．精度や再現率は下がるものの，tf·idf に準ずる結果となっている．

結果を定性的に評価すると，本手法は頻出語を基準とするが，tf による頻出語ですでに十分よいキーワードになっている場合には，本手法の提示する語は概念を特定しすぎた語になっているケースが多かった．逆に，頻出語が一般的な語で情報量が少ない場合には，本手法の提示する語が適切なキーワードとなっているケースが多かった．したがって，本手法は tf に置き換わるような手法ではないが，組み合わせて用いることで，より適切なキーワードの抽出ができると考えられる．

本手法の大きな特徴のひとつは，大規模なコーパスを必要としない手軽さにある．実験で用いた論文および JAIR の各論文についての時間と語の数のプロットを図 3 に示す．プログラムは C++で記述し，Celeron 333MHz の Linux OS 上に実装している．処理時間は，ほぼ語数に対して線形なオーダで増えており，10000 語程度なら数秒で処理が終了する．

*7 コーパスは JAIR(Journal of Artificial Intelligence Research) の 93 年 (Vol.1) から 2001 年 (Vol.14) までの論文全文 166 篇とした．また，語 v に対する idf の重みづけは $\log(D/df(w))+1$ とした．ただし D は全文書数，$df(w)$ は語 w が出現する文書数である．

*8 本手法と同様に構造的な特徴からキーワードを抽出するため，コーパスは不要である．

表 7 本論文に対しての χ^2 値上位の語

順位	χ^2 値	頻度	ラベル
1	126.1	147	語
2	81.2	14	キーワード+抽出
3	68.1	12	χ^2+値
4	45.6	20	確率
5	42.7	5	頻出+語
6	40.7	5	文書+キーワード+抽出
7	38.7	29	出現
8	35.0	5	分野
9	34.6	5	低い
10	34.3	17	語+共起

表 8 論文に対しての precision と coverage

	tf	KeyGraph	**本手法**	tf·idf
precision	0.53	0.42	**0.51**	0.55
coverage	0.48	0.44	**0.61**	0.61
frequency index	28.6	17.3	**11.5**	18.1

表 9 得られた結果におけるフレーズの内訳

	tf	KeyGraph	**本手法**	tf·idf
フレーズの割合	0.11	0.14	**0.33**	0.33
フレーズを除いた precision	0.42	0.36	**0.42**	0.45
フレーズを除いた recall	0.39	0.36	**0.46**	0.54

5. 関 連 研 究

本論文では，語の共起関係によりキーワードを抽出するが，語の共起に着目した研究は非常に多く行われている．[Pereira 93] では，ニュース記事 44 万語から，語を複数のクラスタに分割している．[Even-Zohar 99] は，複数のクラスタに属するような同義語を適切に処理する方法を示している．[Tanaka 96] では，2 言語の共起行列を用いて，コンテキストを考慮した訳語の割り当てを行っている．[Dagan 99] では，共起の確率的な視点からの分析が行われている．これらは大量のコーパスを用い，シソーラス作成や翻訳，音声認識など目的とする語の共起分析である．

文書からのキーワード抽出もしくは索引づけに関する研究は古く，1950 年代後半の Luhn の研究まで遡る [Kageura 96]．ひとつの文書からのキーワード抽出は，基本的には頻度を数えるものであり [Sparck-Jones 72, Noreault 77]，それ以外にはコーパスを用いた研究が主流である [相澤 00]．例えば，[長尾 76] では，全文献をいくつかの分野に分けて分野ごとの単語の頻度を数え，ある単語が各分野に偏りなく出現すれば一般語，少数の分野に偏って出現すれば重要語としており，この偏りを測るために χ^2 検定を用いている．また，文書をカテゴライズするための重要語抽出に χ^2 検定を用いる方法もいくつか提案されている [Schutze 95, Ng 97, 大平 99]．また,「特徴的な語は共起する語の種類が少ない」という本手法と類似の考え方を用いた語の重みづけも最近，提案されている [Hisamitsu 00]．しかし，いずれもコーパスを背景とした方法であり，単一の文書からキーワードを取り出すものではない．

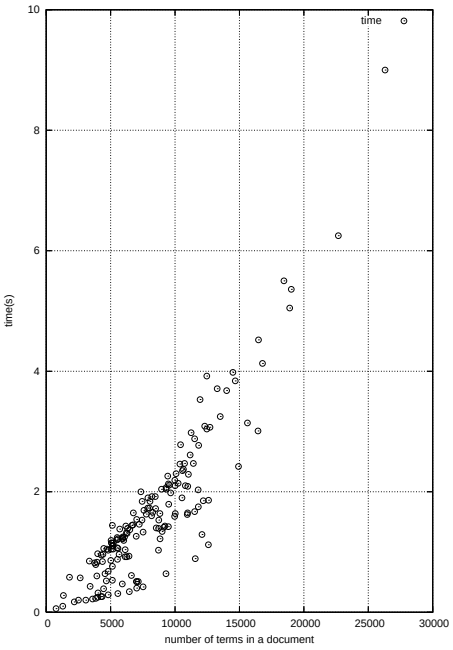

図 3 語数と処理時間

本手法は，KeyGraph[大澤 99] が基礎になっており関連が深い．KeyGraph では，以下で計算した key 値を用いて語 w の重みづけを行う．

$$key(w) = \left[1 - \prod_{g \in G}\left(1 - \frac{f(w,g)}{F(g)}\right)\right] \qquad (2)$$

ここで，G は土台（頻出語のクラスタ）の集合，$F(g)$ は土台 g 中の語の総出現回数 (w が g に含まれていた場合には除く)，$f(w,g)$ は語 w と土台 g 中の語の共起度である．key 値が高くなるためには，特定の土台 g と偏って共起することが必要であるので，上の式の意味するところは本手法のアイディアに近い．本手法は，KeyGraph を統計的に洗練した手法であるとも考えることができる．

6. ま と め

本論文では，単一の文書から語の共起情報を用いてキーワードを抽出する手法を提案した．コーパスを用意することなく，手元にあるテキストだけで処理できるという手軽さが大きな特徴である．今後，電子的な文書を作成／獲得するコストが下がるにしたがって，雑多な文書にも手軽に使える本手法は様々な応用の可能性があるだろう．また，テキストだけではなく，共起関係をもつ一般的なデータから特徴的なアイテムを発見するという用途にも用いることができると考えている．

謝 辞

本研究の内容について議論いただいた筑波大学 大澤幸生氏，東京大学 松村真宏氏に感謝いたします．また，実

験に御協力いたいただいた方々に感謝いたします．最後に，査読者の方には，非常に有益なコメントを頂きました．ありがとうございました．

◇ 参 考 文 献 ◇

[相澤 00] 相澤彰子：語と文書の共起に基づく特徴度の数量的表現について, 情報処理学会論文誌, Vol. 41, No. 12, pp. 3332–3343 (2000).

[Church 90] Church, K. W. and Hanks, P.: Word association norms, mutual information, and lexicography, *Computational Linguistics*, Vol. 16, No. 1, pp. 22–29 (1990).

[Dagan 99] Dagan, I., Lee, L., and Pereira, F.: Similarity-Based Models of Word Cooccurrence Probabilities, *Machine Learning*, Vol. 34, No. 1–3, pp. 43–69 (1999).

[Dunning 93] Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol. 19, No. 1, pp. 61–74 (1993).

[Edmundson 69] Edmundson, H.: New Methods in Automatic Abstracting, *Journal of ACM*, Vol. 16, No. 2, pp. 264–285 (1969).

[Even-Zohar 99] Even-Zohar, Y., Roth, D., and Zelenko, D.: Word Prediction and Clustering, in *Bar-Ilan Symposium on the foundations of artificial intelligence* (1999).

[Fürnkranz 98] Fürnkranz, J.: A Study Using N-grams Features for Text Categorization, Technical report, Austrian Research Institute for Artificial Intelligence (1998), OEFAI-TR-98-30.

[Hisamitsu 00] Hisamitsu, T., Niwa, Y., and Tsujii, J.: A Method of Measuring Term Representativeness — Baseline Method Using Co-occuences Distribution —, in *Proc. Coling 2000*, pp. 320–326 (2000).

[Ishizuka 98] Ishizuka, M. and Matsuo, Y.: SL Method for Computing a Near-optimal Solution using Linear and Non-linear Programming in Cost-based Hypothetical Reasoning, in *Proc. 5th Pacific Rim Conference on Artificial Intelligence (PRICAI'98)*, pp. 611–625 (1998).

[Kageura 96] Kageura, K. and Umino, B.: Methods of Automatic Term Recognition, *Terminology*, Vol. 3, No. 2, pp. 259–289 (1996).

[木本 91] 木本晴夫：日本語新聞記事からのキーワード自動抽出と重要度評価, 電子情報通信学会誌, Vol. 74-D-I, No. 8, pp. 556–266 (1991).

[Luhn 57] Luhn, H. P.: A statistical approach to mechanized encoding and searching of literary information, *IBM Journal of Research and Development*, Vol. 1, No. 4, pp. 390–317 (1957).

[長尾 76] 長尾, 水谷, 池田：日本語文献における重要語の自動抽出, 情報処理, Vol. 17, No. 2, pp. pp.110–117 (1976).

[Ng 97] Ng, H. T., Goh, W. B., and Low, K. L.: Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization, in *Proc. ACM SIGIR'97* (1997).

[Noreault 77] Noreault, T., McGill, M., and Koll, M. B.: *A Performance Evaluation of Similarity Measure, Document Term Weighting Schemes and Representations in a Boolean Environment*, Butterworths, London (1977).

[大平 99] 大平, 帆足, 松本, 橋本, 白井：AIC を用いた重要語抽出手法と重要語を用いたターム重みづけ手法の提案・評価, 知識発見のための自然言語処理シンポジウム (1999).

[大澤 99] 大澤, ネルス E., 石塚：KeyGraph: 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会誌, Vol. J82-D-I, No. 2, pp. 391–400 (1999).

[Pereira 93] Pereira, F., Tishby, N., and Lee, L.: Distributional Clustering of English Words, in *Proc. 31th Meeting of the Association for Computational Linguistics*, pp. 183–190 (1993).

[Schutze 95] Schutze, H., Hull, D. A., and Pedersen, J. O.: A comparison of classifiers and document representations for the routing problem, in *Proc. ACM SIGIR'95* (1995).

[Sparck-Jones 72] Sparck-Jones, K.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval, *Journal of Documentation*, Vol. 28, No. 5, pp. 111–121 (1972).

[Tanaka-Ishii 96] Tanaka-Ishii, K. and Iwasaki, H.: Clustering co-occurrence graph using transitivity, in *Proc. 16th International Conference on Computational Linguistics*, pp. 680–585 (1996).

[Tanaka 96] Tanaka, K. and Iwasaki, H.: Extraction of lexical translations from non-aligned copora, in *Proc. 16th International Conference on Computational Linguistics*, pp. 580–585 (1996).

[徳永 99] 徳永健伸：情報検索と言語処理, 東京大学出版会 (1999).

[東京 91] 東京大学教養学部統計学教室（編）：統計学入門, 東京大学出版会 (1991).

〔担当委員：山本秀樹〕

2001 年 8 月 10 日　受理

著 者 紹 介

松尾　豊(学生会員)

1997 年東京大学工学部電子情報工学科卒業．2002 年同大学院博士過程修了．工学博士．同年より，産業技術総合研究所サイバーアシスト研究所． 仮説推論，数理計画法，キーワード抽出，Web マイニング等に興味がある．ユーザにとって価値の高い情報の提示を目指している．情報処理学会，電気学会，AAAI の各会員．

石塚　満(正会員)

1971 年東京大学工学部電子卒業．1976 年同大学院博士課程修了．工学博士．同年 NTT 入社，横須賀研究所．1978 年東京大学生産技術研究所助教授，同教授を経て，1992 年工学部電子情報工学科教授．2001 年より情報理工学研究科電子情報学専攻．研究分野は人工知能，知識処理，マルチモーダル擬人化エージェント，ネットワーク化知的情報環境．IEEE，AAAI，情報処理学会，電子情報通信学会，映像情報メディア学会，画像電子学会等の会員．