2004－NL－161　(1)
2004／5／13

コーパスからのキーワード自動抽出

金田 有二　藤野 昭典　斉藤 和巳　上田 修功

日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

概要: 本報告では，コーパスからのキーワード自動抽出問題を扱う．従来手法では，主として，語の出現頻度や，語の文書中での出現位置などを土台としており，文書のグローバルな内容は慮されていなかった．しかし，実際の文書コーパス NIPS コーパス）を用いた我々の実験では，同じキーワード持つ文書群では，文書間類似度が高い傾向にあるという直観的に妥当な知見を得た．本稿では，この知見に基づき，キーワード抽出基準に，頻度だけでなく，文書間類似度も考慮した新たな手法を提案する．提案法を，NIPS コーパスに適用し，従来の tf-idf 法と比較した実験結果から，キーワード抽出における文書間類似度の利用の有効性を確認した．

Keyword extraction from corpus

Yuji KANEDA　Akinori FUJINO　Kazumi SAITO　Naonori UEDA

NTT Communication Science Laboratories, NTT Corporation

abstract: In this report, we address the problem of automatic keyword extraction. Conventional methods have been based on term frequencies and term position in a document, not considering the content of the document. In our experiment using NIPS corpus, we found that document similarities among documents with the same keyword are high, which is intuitively reasonable. Based on this finding, we propose a new method in which document similarity as well as the term frequency is incorporated into the criterion of keyword extraction. Comparing the proposed method with the conventional tf-idf method, we confirmed the effectiveness of the document similarity on keyword extraction.

1　はじめに

近年の電子的に蓄積された文書の増大に伴い，大量文書からの情報検索を容易にする手法の重要性が高まっている．文書を特徴づける語—キーワード—の文書への付与も，そのような手法の 1 つである．キーワードを付与する目的の 1 つは文書の要約である．文書の要約が与えられていれば，文書の有用性の判断が容易となり，有用な文書を効率よく検索できるようになる．実際，多くの論文誌や，国際会議の予稿集では，著者に，論文へのキーワードの付与を求めている．しかし，人手によるキーワードの付与はコストが高いため，計算機による自動キーワード抽出技術が広く研究されている．

本報告では，文書コーパスよりの自動キーワード抽出問題を扱う．コーパスよりのキーワード抽出を扱った既存研究の多くは，コーパス中の各文書のキーワード抽出問題を扱っている [2][3]．しかし，コーパス中に出現した出現頻度の高いキーワード (**コーパスのキーワード**と呼ぶ．) の抽出問題も，同様に重要である考えられる，そこで，本報告では，2 つのキーワード抽出問題を扱う．

表 1: NIPS コーパスのキーワード

”reinforcement learning”, ”generalization”, ”unsupervised learning”, ”EM algorithm”, ”recurrent networks”, ”speech recognition”, ”classification”, ”clustering”, ”vision”, ”hidden Markov models”, ”dynamic programming”, ”analog VLSI”, ”radial basis functions”, ”support vector machines”, ”statistical mechanics”, ”recurrent network”, ”object recognition”, ”density estimation”, ”mixture models”, ”on-line learning”, ”dynamical systems”, ”pattern recognition”, ”feature extraction”, ”Hebbian learning”, ”associative memory”

コーパスのキーワード抽出が重要である理由に

ついて述べる．コーパスのキーワードは，コーパスのある種の要約であり，情報検索の際に有用であると考えられる．例として，国際会議 Advances in Neural Information Processing Systems(NIPS) の論文集[1]で，出現頻度が上位の 25 個のキーワードを表 1 に示す．例えば，reinforcement learning (強化学習) は，最も多くの論文に付与されたキーワードである．表 1 より，NIPS が扱う分野をおおまかに把握することができる．また，この表より，関心のあるキーワードを選択し，そのキーワードが付与された文書を検索することで，目的の文書の検索を容易にすることも可能である．

精度の良いキーワード抽出を達成するには，文書間の類似度を考慮することが有効であると考えられる．なぜなら，キーワードが同じ文書では，文書内容が類似することが多いからである．例えば，キーワードが reinforcement learning である文書は，共通する内容は扱っているとみなせる．つまり，キーワードと文書間類似度の間に，なんらかの相関があると推定される．そこで，文書間の類似度を適切に利用すれば，精度の良いキーワード抽出性能が可能となることが期待される．

本報告では，まず，上記の仮説「キーワードが同じ文書の間の類似度が高い」を，実際の論文コーパスを用いて，実験的に確かめる．次に，その結果をふまえて，文書間類似度を用いた，新たなキーワード抽出手法を提案する．最後に，提案法を実コーパスに適用し，従来法である tf-idf と比較する．

1.1 関連研究

キーワード自動抽出手法については幅広く研究されている．大きく分けると，次の 2 つのアプローチ，1) 言語的な情報に基づくものと，2) 統計的な情報に基づくものがある．

言語的な情報に基づく手法は，複雑な計算処理や，シソーラス・概念辞書などの整備が必要な点から，適用分野が限定されるとの指摘がある [6].

統計的な情報に基づく手法としては，語の出現頻度に基づく tf-idf 法 [4] が広く用いられている．その他の特徴量としては，語の文書中での出現位置 [3] や，語の文内共起 [5][7] が用いられている．また，ユニークな手法としては，語を検索後としたときに，検索エンジンが返す文書数を用いる手法もある [1].

[1]NIPS コーパス．NIPS コーパスについては後述.

表 2: キーワードの単語数の分布

単語数	1	2	3	4	5	6
キーワード数	921	3373	941	170	25	1
割合 (%)	17.0	62.1	17.3	3.1	0.4	0.01

キーワード抽出に，文書間類似度を用いるのは，我々の知る限り，本報告が最初である．文書間類似度は，語の出現頻度より計算可能である．そのため，提案法で必要な特徴量は語の出現頻度のみである．

2 キーワードと文書間類似度

本節では，実際の文書コーパスを用いて，キーワードと文書間類似度の間の関係について調べる．まず，節 2.1 で，本報告で用いた文書コーパスについて述べる．次に，節 2.2 で，用いたコーパスにおける，キーワードに関するいくつかの統計量について調べた結果について述べる．本報告では，文書中に出現した語の出現頻度を基にキーワードを抽出する．そこで，節 2.3 で，語を選択する前処理について述べる．そして，節 2.4 で，用いた文書間類似度について述べる．最後に，節 2.5 で，この文書間類似度と，キーワードの間の関係について調べた結果について述べる．

2.1 NIPS コーパス

国際会議の論文コーパスを用いて，キーワードの統計的性質を調べる．コーパスとして，Sam Roweis[2]の収集した，国際会議 NIPS の論文コーパスから，NIPS-1 から NIPS-12 までの 12 年分の論文を用いた．なお，以後，この文書群を NIPS コーパスと呼ぶ．NIPS コーパスは，OCR により電子化された 1650 稿の論文よりなり，各論文には，その論文の主題 (“subject”) が複数個与えられている．“subject” は，論文を特徴づける単語，もしくは，フレーズであり，キーワードとみなすことができる．そこで，以後，NIPS コーパスにおける “subject” をキーワードとみなす．なお，論文本文と同様に，“subject” も OCR により電子化されている．

2.2 キーワードの統計量

キーワードの統計量について述べる前に，まず，記号を整理する．コーパスの文書群を $\mathcal{D} = \{d_1, \cdots, d_N\}$ とする．ただし，N は文書数である．また，文書

[2]http://www.cs.toronto.edu/ roweis/

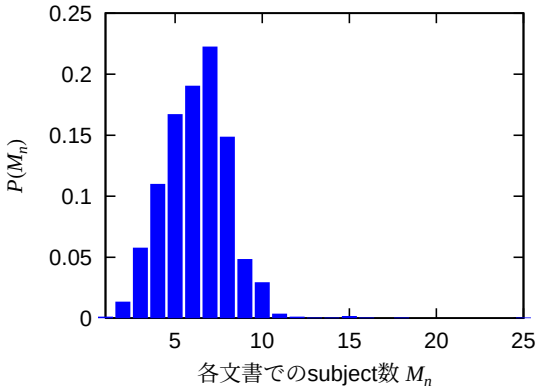

図 1: 各文書でのキーワード数の分布

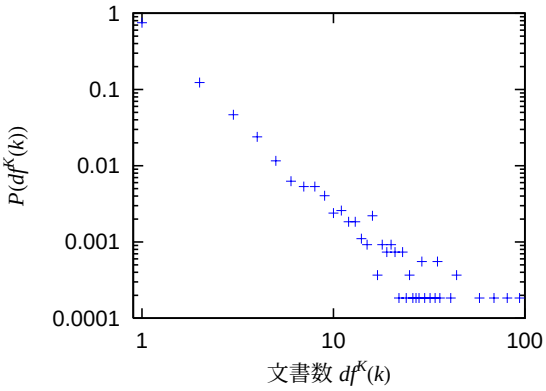

図 2: キーワードの出現書数の分布

d_n におけるキーワード数を M_n とし，文書 d_n のキーワードを $\mathcal{K}_n = \{k_{n,1}, \cdots, k_{n,M_n}\}$ とする．さらに，コーパス全体で出現したキーワードを $\mathcal{K}$ とする．すなわち，$\mathcal{K} = \bigcup_n \mathcal{K}_n$ である．また，ある語 $k \in \mathcal{K}$ がキーワードとなる文書全体を $\mathcal{D}_k^{\mathcal{K}}$ とする．すなわち，$\mathcal{D}_k^{\mathcal{K}} = \{d_n | d_n \in \mathcal{D}, \quad k \in \mathcal{K}_n\}$ とする．

NIPS コーパスにおける，キーワードに関するいくつかの統計量を以下に示す．

(i) キーワードの種類 は 5431 個であった．

(ii) 各キーワードの単語数の分布を表 2 に示す．表 2 が示すように，単語数が 4 以下のキーワードが 99 % 以上を占めていた．

(iii) 1 文書あたりのキーワード数 M_n の分布を，図 1 に示す．図 1 が示すように，多くの文書のキーワード数は 10 以下であった．

(iv) キーワード k が出現した文書数 $df^{\mathcal{K}}(k) = \left|\mathcal{D}_k^{\mathcal{K}}\right|$ の分布を，図 2 に示す．図 2 より， $df^{\mathcal{K}}(k)$ の分布は，おおむねべき分布に従うことが分かる．すなわち，ほとんどのキーワードは，少数の文書にしか出現しなかった一方，多数の文書に出現したキーワードも少数存在した．具体的には，1 つの文書にしか出現しなかったキーワードが 75% を占めたのに対し，94 個の文書に出現したキーワードも 1 つ存在した．

2.3 語選択の前処理

前処理により，考慮の対象とする語を選択する．なお，語が，単語を意味しないことに注意．表 2 が示すように，NIPS コーパスでは，単語数が 4 単語以下のキーワードが 99%以上 を占めていた．そこで，1, 2, 3, 4-gram を語として用いた．さらに，コーパス中に出現した，1, 2, 3, 4-gram のうち，次の前処理により，用いる語を選択した．

(i) ハイフンの除去，表記の統一．例えば，“back-propagation”, “backpropagation”, “back propagation” の表記を統一．

(ii) ストップワード，前置詞，区切り文字 (コンマ，カンマ，コロン，セミコロンなど) などをシンボルに置き換え．

(iii) 大文字を小文字に．各単語をステミングし，語末を除去．

(iv) キーワードとして無効であると推定される語を除去．(ストップワードを含む，前置詞で終わる，区切り文字を間に含む，文字数が少いなど．) なお，語の出現頻度による，語の削減は行わなかった[3]．

上記の処理によって得られた語の集合を $\mathcal{W}^{\mathcal{K}} = \{w_1^{\mathcal{K}}, \cdots w_{V^{\mathcal{K}}}^{\mathcal{K}}\}$ とする[4]ただし，$V^{\mathcal{K}}$ は語の総数である．前処理の結果，選択された語の種類 ($= V^{\mathcal{K}}$) は 1,484,618 個であった．

また，同様の処理をキーワードに対しても施す．簡単のため，このようにして得られたキーワードに対しても，前節と同様の表記を用いる．以後，前節の表記を，前処理を施した結果に対して用いることに注意．前処理により，キーワードの種類 ($=|\mathcal{K}|$) は 4,718 個に減少した (ステミング等により，複数のキーワードが同一視されたため)．これらのキーワードのうち，語集合 $\mathcal{W}^{\mathcal{K}}$ に含まれるもの個数 ($\left|\mathcal{K} \cap \mathcal{W}^{\mathcal{K}}\right|$) は 4,029 個 (85.4%) であった．

[3]少数頻度の語でもキーワードとなり得るため．例えば，キーワードと成り得る語の内，コーパス中に 1 回しか出現しない語が 6 %，2 回しか出現しない語が 5 % を占めていた．

[4]$\mathcal{W}$ に，$\mathcal{K}$ の添字を付けるのは，類似度計算用に用いる語集合 $\mathcal{W}^{\mathcal{S}} (\subset \mathcal{W}^{\mathcal{K}})$ と区別するためである．

得られた語集合 $\mathcal{W}^{\mathcal{K}}$ を用いて，語の出現頻度ベクトルを定める，文書 d_n における，語 $w_i^{\mathcal{K}}$ の出現頻度を $x_{n,i}^{\mathcal{K}}$ とし，また，文書 d_n の語出現頻度ベクトルを $\mathbf{x}_n^{\mathcal{K}} = (x_{n,1}^{\mathcal{K}}, \cdots, x_{n,V^{\mathcal{K}}}^{\mathcal{K}})$ とする．

2.4 文書間類似度の設定

文書間類似度を，tf-idf 変換された語出現頻度ベクトル間のコサイン類似度により定める．[4]．この類似度は，情報検索や文書クラスタリングにおいて広く用いられている．しかし，単純に， 語出現頻度ベクトル $\{\mathbf{x}_n^{\mathcal{K}}\}_{n=1}^N$ 間のコサイン類似度を用いるのは不適切であると考えられる．なぜなら，語集合 $\mathcal{W}^{\mathcal{K}}$ の次元が高次元の場合 (例えば， NIPS コーパスでは $\left|\mathcal{W}^{\mathcal{K}}\right| = 1,484,618$)，文書間類似度の信頼性に問題が生じ得るからである．そこで，語集合 $\mathcal{W}^{\mathcal{K}}$ から，類似度計算に用いる語を選択し，語数の少ない語集合 $\mathcal{W}^{\mathcal{S}} = \{w_1^{\mathcal{S}}, \cdots, w_{V^{\mathcal{S}}}^{\mathcal{S}}\}, \mathcal{W}^{\mathcal{S}} \subset \mathcal{W}^{\mathcal{K}}$ を構成する．そして，$\mathcal{W}^{\mathcal{S}}$ を用いて，より低次元の語出現頻度ベクトル $\{\mathbf{x}_n^{\mathcal{S}}\}_{n=1}^N$ を構成する．

本報告では，leave-one-out 法を用いて， $\mathcal{W}^{\mathcal{S}}$ を構成した．具体的には，コーパス中での出現頻度が 40 回以上の語を，文書間類似度計算に用いる語として選択した．このような語の種類は 7,907 個であり，元々の語数の約 5 % であった．選択基準については付録を参照のこと．

語集合 $\mathcal{W}^{\mathcal{S}}$ に関する，語出現頻度ベクトル $\{\mathbf{x}_n^{\mathcal{S}}\}_{n=1}^N$ を用いて，文書間類似度を定める．具体的には，この低次元の語出現頻度ベクトルを tf-idf 変換し，変換されたベクトル間のコサイン類似度を文書間類似度とする．すなわち，文書 d_n, d_m 間の文書間類似度 $s(d_n, d_m)$ を，次のように定める．

$$s(d_n, d_m) = s(\mathbf{x}_n^{\mathcal{S}}, \mathbf{x}_m^{\mathcal{S}}) = \frac{\sum_{i=1}^{V^{\mathcal{S}}} (idf_i\, x_{n,i}^{\mathcal{S}})(idf_i\, x_{m,i}^{\mathcal{S}})}{\sqrt{\sum_{i=1}^{V^{\mathcal{S}}} (idf_i\, x_{n,i}^{\mathcal{S}})^2}\sqrt{\sum_{i=1}^{V^{\mathcal{S}}} (idf_i\, x_{m,i}^{\mathcal{S}})^2}} \quad (1)$$

ただし， idf_i は，語 $w_i^{\mathcal{S}}$ の inversed document frequency であり，次のように定義する．$idf_i = \log \frac{N}{\left|\mathcal{D}_i^{\mathcal{S}}\right|}$, $\mathcal{D}_i^{\mathcal{S}} = \{d_n |\, d_n \in \mathcal{D},\, x_{n,i}^{\mathcal{S}} > 0\}$.

2.5 文書間類似度とキーワードの関係

キーワードが同じ文書の間では，文書間の類似度が高いと考えるのが自然である．そこで，この仮説を確かめるため，次の手順により，同一のキーワードを持つ文書間の類似度を，同一の語が出現した文書間の類似度と比較した:

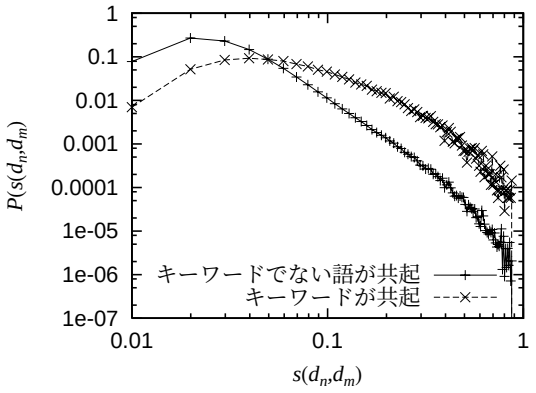

図 3: 文書間類似度の分布の比較

1) キーワードと成り得るような語 $w_i^{\mathcal{K}} \in \mathcal{K}$ を 1 つ選択する．2) この語がキーワードである文書を全て求め，これらの文書間の類似度を求める．3) この語が出現したが，キーワードではない文書を全て求め，これらの文書間の類似度を求める．

上記の手続きを，キーワードと成り得る全ての語 $w^{\mathcal{K}} \in \mathcal{K}$ に対して実行した．そして，2), 3) の手続きで得られた文書間類似度の分布を比較したものを，図 3 に示す．図 3 より，NIPS コーパスでは，共通のキーワードを持つ文書間の類似度が，共通の語を持つ文書間の類似度よりも高い傾向があることが分かる．これより，文書間類似度の適切な利用が，精度のよいキーワード抽出性能に寄与することが期待される．

3 提案法

前節で調べたように，NIPS コーパスにおいて，キーワードが共通の文書間の類似度は，共通の語が出現した文書間の類似度より高い傾向があった．そこで，このことを考慮した，新たなキーワード抽出法を提案する．

提案法では，従来法の多くと同様に，文書 d_n において，語 $w_i^{\mathcal{K}}$ がキーワードである基準値 (度合い) $r_{n,i}$ を求める．そして，この規準値 $r_{n,i}$ を基に，次の手順によりキーワードを抽出する．

文書毎のキーワード抽出: 文書毎に，基準値 $r_{n,i}$ が上位の語を，それぞれの文書のキーワードとして抽出する．

コーパスのキーワード抽出: 文書毎の基準値の和

表 3: キーワードの出現頻度を考慮した，コーパスのキーワード抽出の抽出性能 (%) の比較

	提案法						tf-idf (従来法)					
$\xi_{n,i}$	normal			log			normal			log		
正規化		L_1	L_2		L_1	L_2		L_1	L_2		L_1	L_2
Breaeven	30.10	30.27	**31.00**	29.31	30.37	30.95	24.58	24.85	24.75	24.98	25.29	25.28
Ave. Prec.	26.71	27.33	**28.03**	26.73	26.85	27.30	23.48	23.43	23.62	23.57	23.46	23.65

$\sum_{n=1}^{N} r_{n,i}$ が上位の語を，コーパスのキーワードとして抽出する．

次に，提案法における規準値 $r_{n,i}$ の求め方について説明する．前節で示したように，キーワードが同じ文書間の類似度は，そうでない文書間に比べて高いことが多かった．従って，文書間類似度が高い文書群では，共通のキーワードを持つ確率が高いと考えられる．しかし，文書間類似度が高い文書群であっても，必ずしも共通のキーワードを持つとは限らない．そこで，共通のキーワードを持つための条件を 1 つ加える．すなわち，文書間の類似度が高く，かつ，共通の重要語を持つ文書群では，その共通の重要語をキーワードとする文書が多いと仮定する．文書群に共通の重要語が，キーワードに成り易いと考えるのは自然である．この仮説を言い替えると，次の仮説となる．

仮説：**「ある文書におけるキーワードの多くは，その文書において重要な語，かつ，その文書に類似した文書においても重要な語である」**

ただし，本稿では「重要な語」については特に定義せず，なんらかの手法により，文書 d_n における語 $w_i^{\mathcal{K}}$ の重要度 $\xi_{n,i}$ が与えられたとする．そして，この仮説を基に，語の重要度と文書間類似度を用いて，キーワード抽出の規準値 $r_{n,i}$ を次のように定める．

$$r_{n,i} = \xi_{n,i} \sum_{m=1}^{N} \frac{s(d_n, d_m)}{\sum_{m'=1}^{N} s(d_n, d'_m)} \xi_{m,i} \qquad (2)$$

式 (2) の $\xi_{n,i}$ は，仮説の**「ある文書におけるキーワードの多くは，その文書において重要な語」**の部分を表す．また，式 (2) の $\sum_{m=1}^{N} \frac{s(d_n,d_m)}{\sum_{m'=1}^{N} s(d_n,d'_m)} \xi_{m,i}$ は，仮説の**「かつ，その文書に類似した文書においても重要な語である」**の部分を表す．

4 評価実験

提案法のキーワード抽出手法を，NIPS コーパスに適用し，コーパスのキーワードと，文書毎のキーワードを抽出した．そして，それらの抽出性能を，従来法である tf-idf 法の抽出性能と比較した．

4.1 語の重要度の選択

提案法では，語の重要度 $\xi_{n,i}$ の設定が重要である．tf-idf 法との比較を明確にするため，tf-idf 法を基に，語の重要度 $\xi_{n,i}$ を定めた．tf-idf にはいくつかのバリエーションがあるが，次の 2 つを用いた．(i) normal: $\xi_{n,i} = idf_i \times x^{\mathcal{K}}_{n,i}$, (ii) log: $\xi_{n,i} = idf_i \times \log(x^{\mathcal{K}}_{n,i})$. また，文書長の影響を低減化するために，tf-idf の値の正規化も考慮した．次の 3 種類の正規化を用いた．1) 正規化なし，2)L_1 ノルムによる正規化，3)L_2 ノルムによる正規化．従って，計 6 種類の語の重要度を用いた．例えば log を L_2 ノルムで正規した場合，語の重用度 $\xi_{n,i}$ は，$\xi_{n,i} = \frac{idf_i \ \log(x^{\mathcal{K}}_{n,i})}{\sqrt{\sum_{i=1}^{\mathcal{V}^{\mathcal{K}}} \left\{idf_i \ \log(x^{\mathcal{K}}_{n,i})\right\}^2}}$ となる．

tf-idf 法は，基準値 $r_{n,i}$ として，上記の語の重要度 $\xi_{n,i}$ を用いたものである．そこで，tf-idf 法でも，提案法と同様に， 6 種類の重要度を用いた．

4.2 評価尺度

コーパスのキーワード抽出性能を，**精度**(抽出した語が，キーワードである割合) と，キーワードの出現頻度を考慮した**再現率**(抽出すべきキーワードを抽出した割合) により評価した．再現率では，複数の文書に付与されたキーワードは，付与された文書数だけの重みをもつとした．例えば，キーワードとして，2 つの語を抽出したとき，1 つの語が，90 個の文書に付与されてたキーワードであり，もう 1 つの語はキーワードでないとする．このときの，精度を $1/2 = 50(\%)$ に，再現率を $90/10098 = 0.9(\%)$ とした．なお，10098 は，文書に付与されたキーワードの総数 $(= \bigcup_n |\mathcal{K}_n|)$ である．

また，文書毎のキーワード抽出問題に対しては，文書毎に，キーワード抽出の精度と再現率を求めた．そして，それらの値の平均値により，文書毎のキーワード抽出性能を評価した．

4.3 実験結果

まず，コーパスのキーワード抽出問題における，各手法の Breaeven point, 11-point averaged precision

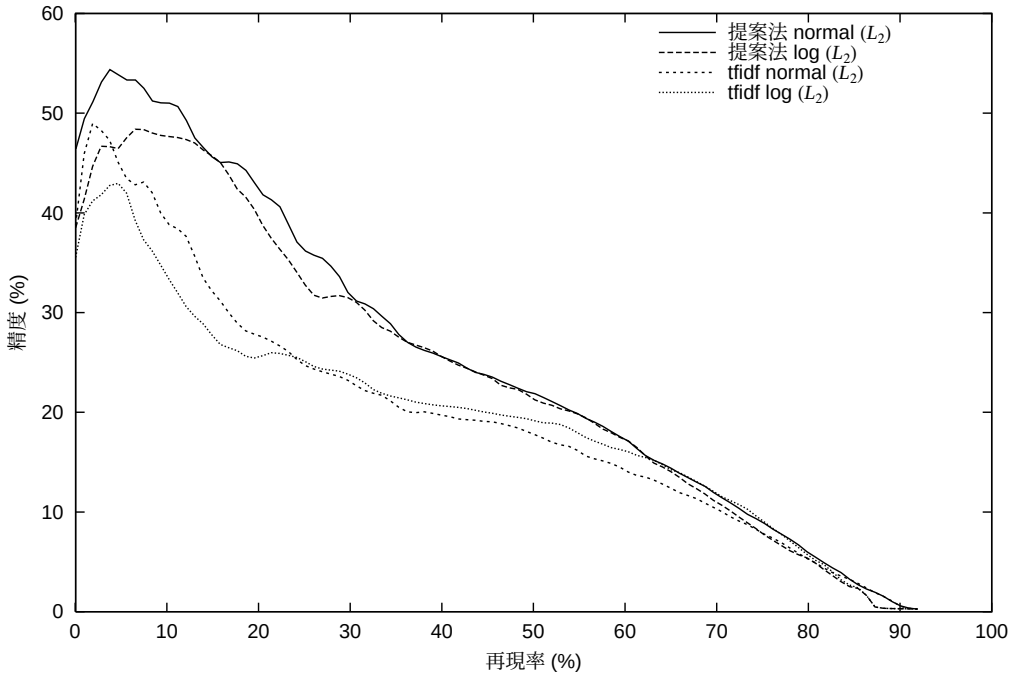

図 4: キーワードの出現頻度を考慮した，コーパスのキーワード抽出性能の比較

表 4: 文書毎のキーワード抽出精度 (%) の比較

	提案法						**tf-idf (従来法)**	
$\xi_{n,i}$	normal			log			normal	log
正規化		L_1	L_2		L_1	L_2		
$R=1$	15.03	14.73	15.03	18.06	17.58	17.88	**18.18**	14.48
$R=5$	12.58	12.79	12.69	**14.17**	13.84	14.01	13.10	9.87
$R=10$	10.30	10.40	10.41	11.16	**11.26**	11.25	10.12	7.83

[4] を表 3 を示す．表 3 に示すように，最良の結果を達成した語の重要度 $\xi_{n,i}$ は，提案法では，normal を L_2 ノルムで正規化したもの，一方，従来法では，log を L_2 ノルムで正規化したものであった．それぞれの最良の結果を比較すると，提案法が，従来法を， Breaeven point で 5.72 %，11-point averaged precision で 4.38 % それぞれ上回った．

次に，各手法の，文書毎のキーワード抽出の，$R=1,5,10$ のときの精度を表 4 に示す．なお，従来法では，正規化よる抽出性能の変化はないので，正規化については省略した．最良の重要度を用いたときの精度を比較する．表 4 に示すように，$R=1$ のおいて，提案法の精度は従来法と同等であった．また，$R=5,10$ のとき，提案法は，従来法を若干上回る精度を達成した．

最後に，各問題における，再現率・精度曲線を示す．ただし，見易さのため，抽出性能の比較的高かった，L_2 ノルムで正規化した場合のみを示す．コーパスのキーワード抽出の再現率・精度曲線を図 4 に，各文書毎のキーワード抽出の再現率･精度曲線を図 5 に示す．なお，図 4 のグラフは，スプラインにより平滑化している．図 4 で，再現率が 100 % に達しないのは，$\mathcal{W}^{\mathcal{K}}$ が，全てのキーワードを含まなかったためである．また，図 5 では $R=1,\cdots,100$ のときの値のみを示している．図 4, 5 より，重要度 $\xi_{n,i}$ として適切なものを選択するならば，いずれの問題においても，提案法が，従来法を上回る抽出性能を達成し得ることが分かる．

5 結び

実コーパスを用いて，キーワードと文書間類似度に関係があることを示し，そのことを考慮した，新たなキーワード抽出手法を提案した．また，提案法を，NIPS コーパスにおける 2 種類のキーワード抽出問題，1) 文書毎のキーワード抽出問題，2) コー

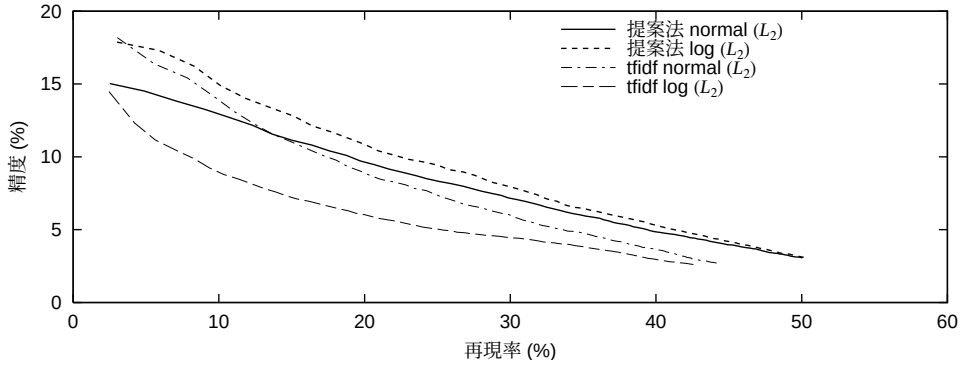

図 5: 文書毎のキーワード抽出性能の比較

パスのキーワード抽出問題に対して適用し，いずれに問題においても，従来法を上回る性能を性能を達成し得ることを示した．

参考文献

[1] Turney, P.: Coherent Keyphrase Extraction via Web Mining, *Proceedings Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico, pp. 434–439 (2003).

[2] Turney, P. D.: Learning Algorithms for Keyphrase Extraction, *Information Retrieval*, Vol. 2, No. 4, pp. 303–336 (2000).

[3] Witten, I., Paynter, G., Frank, E., Gutwin, C. and Nevill-Manning, C.: KEA: practical automatic keyphrase extraction, *Proc. Digital Libraries '99*, Berkeley, CA, pp. 254–255 (1999).

[4] 徳永健伸: 情報検索と言語処理，東京大学出版会 (1999).

[5] 松尾豊, 石塚満: 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, 人工知能学会誌, Vol. 17, No. 3, pp. 213–227 (2002).

[6] 原正巳, 中島浩之, 木谷強: テキストのフォーマットと単語の範囲内重要度を利用したキーワード抽出, 情報処理学会論文誌, Vol. 38, No. 2, pp. 299–309 (1997).

[7] 大澤幸生, Benson, N. E., 谷内田正彦: KeyGraph:単語共起グラフの分割統合によるキーワード抽出, 電子情報通信学会論文誌, Vol. J82-D1, No. 2, pp. 391–400 (1999).

A 文書間類似度計算に用いる語の選択規準

適切な類似度を定めるためには，適切な選択語集合 $\mathcal{W}^{\mathcal{S}}$ を定める必要がある．そこで，語の選択基準値を定義し，基準値を最大にする語集合を文書間類似度計算に用いる．

選択規準について説明する前に，いくつかの記号を導入する．語を選択する際に，削除された語を $\mathcal{W}^{\mathcal{D}} = \mathcal{W}^{\mathcal{K}} \setminus \mathcal{W}^{\mathcal{S}}$ とする．語を選択して，新たな語出現頻度ベクトルを構成することは，$\mathcal{W}^{\mathcal{D}}$ に属する語 $w_j^{\mathcal{K}}$ の出現頻度 $x_{n,j}^{\mathcal{K}}$ を 0 にすること，すなわち，$\mathbf{x}_n^{\mathcal{K}}$ を，次の式により，$\tilde{\mathbf{x}}_n^{\mathcal{K}} = (\tilde{x}_{n,1}^{\mathcal{K}}, \cdots, \tilde{x}_{n,V^{\mathcal{K}}}^{\mathcal{K}})$ に変換することに等しい．

$$\tilde{x}_{n,i}^{\mathcal{K}} = \begin{cases} x_{n,i}^{\mathcal{K}} & \text{if} \quad w_i \in \mathcal{W}^{\mathcal{S}} \\ 0 & \text{if} \quad w_i \in \mathcal{W}^{\mathcal{D}} \end{cases} \tag{3}$$

上記の $\mathbf{x}_n^{\mathcal{K}}, \tilde{\mathbf{x}}_n^{\mathcal{K}}$ を用いて，次の，語集合 $\mathcal{W}^{\mathcal{S}}$ の選択規準を提案する．

$$\frac{1}{N} \sum_{n=1}^{N} \max_{m \neq n} s(\mathbf{x}_n^{\mathcal{K}}, \tilde{\mathbf{x}}_m^{\mathcal{K}}) \tag{4}$$

この選択基準は， leave-one-out 統計量の 1 種である．まず，ある文書ベクトル $\mathbf{x}_n^{\mathcal{K}}$ を除く，全ての文書ベクトル $\mathbf{x}_m^{\mathcal{K}}, n \neq m$ を $\tilde{\mathbf{x}}_m^{\mathcal{K}}$ に変換したとする．そして，変換された $N-1$ 個のベクトル $\tilde{\mathbf{x}}_m^{\mathcal{K}}$ と，$\mathbf{x}_n^{\mathcal{K}}$ との間の類似度 $s(\mathbf{x}_n^{\mathcal{K}}, \tilde{\mathbf{x}}_m^{\mathcal{K}})$ を求め，それらの類似度の最大値の n に関する平均値が，式 (4) の基準量である．