グループウェア 36−18
（2000．5．26）

参照情報を利用した文書特徴量抽出方式

野口 進祐　木下 哲男　白鳥 則郎
東北大学電気通信研究所 / 情報科学研究科
e-mail: {noguchi,kino,norio}@shiratori.riec.tohoku.ac.jp

本稿では，文書にあらかじめ付与されている他文書への参照情報を利用することで文書ベクトルを拡張する手法を提案し，文書分類・検索の精度改善を試みる．提案手法では参照情報を利用して意味的にまとまりのある文書の集合を形成し，その解析を行うことでベクトルを構成する単語の重みを修正する．提案手法の評価実験として，学術論文を利用した文書分類実験と，HTML 文書を利用した文書検索実験を行い，既存手法と比較して分類精度・検索精度ともに改善されることを確認した．

A Method to Extract Features Based on Reference Information

Shinsuke Noguchi, Tetsuo Kinoshita and Norio Shiratori
Research Institute of Electrical Communication /
Graduate School of Information Sciences, Tohoku University
e-mail: {noguchi,kino,norio}@shiratori.riec.tohoku.ac.jp

In this paper, we propose a method to expand the document bectors based on reference information and improve the accuracy of document retreival and classification. This method analyzes the document sets which consist of reference information and modifies the term weight. Through the experiment of both the classification of scientific papers and the retrieval of HTML documents, we confirm that the performance of the classification and retrieval can be improved by the proposed method comparing with the existing method.

1 はじめに

近年，コンピュータに搭載される記憶装置の大容量化や，ネットワークの発達により，膨大な数の電子化された文書が流通するようになっている．大量の電子化文書を効率良く分類・検索するために，あらかじめ各文書の特徴を抽出し，それを効果的に活用する手法が工夫されている．

現在最も広く使われている特徴量抽出手法としては，文書中に出現する単語とその重みを用いて文書の情報を表現する文書ベクトル化手法が知られている．しかし，既存のベクトル化手法では，文書を個別に解析して単語の重み付けを行うために，文書の主題を正確にベクトルに反映できないという問題がある．

本稿では，文書間の関連性を利用して同一主題に基づいた文書の集合を形成し，その文書集合の解析を行う事で，文書ベクトルを拡張する手法を提案する．また，提案したベクトル拡張法に基づく文書分類・検索実験により，本手法の有効性を検証する．

2 参照情報を利用した特徴量抽出手法

2.1 既存手法とその問題点

文書のベクトル化手法としては TF/IDF 法が最も一般的である [1]．今コーパス M に出現する $word_1, word_2, \ldots, word_n$ の n 個の単語に注目して文書ベクトルを生成する．このとき，文献 x の文書ベクトル V_x を，

$$V_x = (v_{x1}, v_{x2}, \ldots, v_{xn})$$

で表すと，ベクトルの第 i 要素は文書 x に単語 $word_i$ が出現する回数 tf_{xi} と，単語 $word_i$ が一回の出現当りに持つ情報量 idf_i の積により求まる．

$$v_{xi} = tf_{xi} \cdot idf_i$$
$$idf_i = \log \frac{|M|}{m_i}$$
$$m_i : word_i \text{の出現する文書数}$$

以上の式からもわかるように，単純な TF/IDF 法

では文書中に出現する各単語に対して，表層的な統計処理を行ってベクトルを算出するため，文書ベクトルは文書を記述する著者の語彙体系に依存しやすく，的確に文書の特徴を反映できないという問題点を持つ (語彙曖昧性の問題).

また，各単語の情報量はコーパス全体から見た単語の出現確率（の逆数）にのみ依存しており，コーパス内で単語の持つ情報量が一意に定まってしまうという問題点を持つ．この場合コーパスに収められる文書全てが共通の分野の話題を扱うということが前提となるが，複数の分野に関する文書が同時に収められている場合には，単語の使用される状況や分野により単語の持つ重要度は異なるはずであるので，文書の主題が的確にベクトルに反映されないという問題が指摘されている (単語情報量の問題).

これらの問題点を解決するために，コーパス内で関連した内容を持つ文書の部分集合を形成し，その解析から文書ベクトルを拡張する方法がある．関連する文書の集合を形成し，その集合内で出現する単語をベクトル生成時に考慮できれば，単独文書の解析では陥りがちな語彙の曖昧性を解消できる．

また，類似した分野の文書を選別して文書集合を形成すれば，集合内で頻出する単語や希少な単語を特定することが可能であり，これらの特異な語をベクトルに反映させれば，単語情報量の問題を軽減できるはずである．

文書間の関連性を考慮した文書ベクトルの拡張については，すでに金沢らが学術論文にあらかじめ付与されているキーワードを用いる手法を提案している [2]．この手法では同一のキーワードを持つ論文から文書部分集合を形成し，この部分集合内の重要単語を集合内の各文書に補うことによって，検索の精度を高めることを試みている．

しかし，一般に異なる主題を扱う論文でもキーワードの表層的なマッチングのみで局所集合として結びつく可能性がありうること，また1つの論文には複数のキーワードが付与されることが知られており，その結果，部分集合にはノイズとなる文書が多数存在し，これを用いて文書ベクトルを拡張しても，文書の主題が一層埋没してしまうという可能性がある．

2.2 参照情報を利用した関連文書集合の形成

文書間の関連性を特定する方法の一つとして，文書にあらかじめ付与されている参照情報の利用が考えられる．参照情報とは，HTML 文書におけるハイパーリンク構造や，学術論文における引用関係などのように，文書記述者によって文書内に明示的に記述される他文書へのリンク情報である．

一般に，被参照文書には著者の論説の根拠や，比較対照として適切だと判断されたものだけが選別されることから，これを利用して文書集合が形成できれば，集合内の文書は共通の話題を扱っている可能性が極めて高いものになる．

図 1は参照関係によって形成される文書集合の概念図である．ここでは，ベクトル化の対象となる文書 x と，x が直接参照する被参照文書群によって集合 (参照集合) を形成することにする．

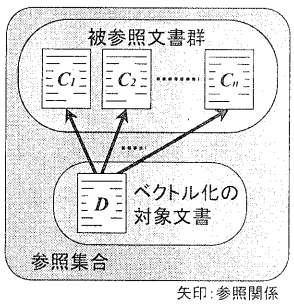

図 1: 参照集合の概念図

2.3 関連文書集合を利用したベクトルの拡張

関連文書集合を利用して文書ベクトルを拡張する手法を示す．まず，文書 x と x が直接参照する被参照文書から，関連文書集合 M_x を形成する．このとき文書 x に付与される文書ベクトル V'_x は以下のように拡張される．ただし関連文書集合内に含まれる各文書のベクトル $V_{yi} = (v_{y1}, v_{y2}, \ldots, v_{yn})$ はあらかじめ TF/IDF 法で計算しておくものとする．

$$V'_x = (v'_{x1}, v'_{x2}, \ldots, v'_{xn}) \quad (1)$$

$$v'_{xi} = max(v_{xi}, \frac{1}{|M_x|} \sum_{y \in M_x} v_{yi}) \quad (2)$$

式 2は関連文書集合 M_x 内の TF/IDF ベクトルを要素ごとに平均し，x の TF/IDF ベクトル要素 v_{xi} と比較して値の大きい方を拡張後のベクトル要素 v'_{xi} とすることを意味している．式 (2) のベクトル拡張により,「文書 x 中には出現しないが文書 x と関連の

深い語」が補われることが予想されるので，語彙の曖昧性に起因する分類·検索精度の低下を防ぐことが可能である．また関連文書集合内 M_x 内のベクトル平均をとり共通主題部分を強調させることで「集合内で重要な語」の重みが増えるはずである．TF/IDF 法では均一な情報量を用いるために分野毎の単語の重要性を考慮することが出来なかったが，共通主題を補うことにより，分野に重要な単語の重みを強調することが出来る．

3 評価実験

提案手法の有効性を検証するために，提案手法と既存手法の比較実験を行った．行った実験は，学術論文を用いた分類実験と，HTML 文書を用いた検索実験の二種類である．ここでは，各々の実験の概要と実験結果を示し，提案手法の性質について考察する．

3.1 学術論文を用いた文書分類実験

学術論文に提案手法を適用する際には，参照情報として引用関係を利用する．学術論文として発表される文書は，その内容がある程度審査されている事からノイズとなる情報が比較的少なく，参照集合を形成する過程で，参照情報を選別する必要がなくなる．

3.1.1 実験データ

実験データとして，物理分野の学術論文アーカイブである e-print アーカイブを用いた [3]．e-print アーカイブでは，あらかじめ文献が分野毎に分類管理されており，ここから表 1に示した 12 分野の論文を対象に，99 年 5 月時点で最新のものから 50 文献ずつを抽出して実験対象とした．ただし，提案手法は論文の引用関係に基づくため，対象文献のうち e-print アーカイブ内の論文を全く引用していないものは分類対象から外している．

実験対象 600 文献のうち，e-print アーカイブ内の文献を引用している文献は 422 文献であった．この 422 文献に関して実際に分類処理を行う．

e-print アーカイブに登録される文献では，引用関係が分野 ID+発表年月+通し番号の形式で与えられ，hep-th9905001 のように表現される．この情報を利用した単純なパターンマッチングにより引用関係が抽出できる．

分野 ID	分野
alg-geom	代数幾何
astro-ph	宇宙物理
gr-qc	一般相対論 & 量子宇宙論
hep-ex	高エネルギー物理（実験）
hep-lat	高エネルギー物理（格子）
hep-ph	高エネルギー物理（現象）
hep-th	高エネルギー物理（理論）
math	数理物理
nucl-th	原子核理論
physics	物理学
q-alg	量子代数
quant-ph	量子物理

表 1: 分類実験の対象となる分野

3.1.2 実験手順

実験対象となる文献データに対し，提案手法と単純な TF/IDF 法の二手法によって，それぞれ文書ベクトルを生成する．提案手法のベクトル生成手順を以下に示す．

◦ ベクトル生成手順

(1) 分類の対象となる文献を用意する．

(2) このデータから引用関係を抽出し，被引用文献となるものを用意する．抽出された被引用文献数はのべ 2135 文献であった．

(3) 対象文献からベクトル生成に使用する単語を抽出．ベクトルの単語数 (= ベクトルの次元数) はそのまま分類の処理速度に影響するので，少ない単語数でも高い分類精度を得られることが望ましい．実験では 128～4096 個まで 128 個刻みで，コーパス内で出現頻度の高い順に単語を選び出し，ベクトル生成に使用した．

(4) 分類の対象文献と被引用文献群の TF/IDF ベクトルを算出する．

(5) 前出の (2) 式にしたがって，分類対象文献と被引用文献の TF/IDF ベクトルの平均を取り，これを利用して文書ベクトルを拡張する．

次に，生成された文書ベクトルをもとに，各文書の分類を以下の手順で行う．

∘ 分類手順

(1) あらかじめ各分野の基準となるベクトル（分野基準ベクトル）を用意しておく．これはe-printアーカイブの各分野から，分類の対象以外の文献を任意に20個づつ選び出し，これらの文献のTF/IDFベクトルを平均する事により算出する．

(2) 分類対象文献の文書ベクトルと，分野基準ベクトルの類似度を計算し，これに基づいて文書分類を行う．文書xと文書yの類似度は，文書ベクトルV_x,V_yのそれぞれの絶対値を1に正規化してから両者の内積を求める事により得られる．分野iの基準ベクトルをC_iで表すと，文献xがどの分野に属するかを判定するには，全C_iについて類似度

$$S_{xi} = \frac{V_x \cdot C_i}{|V_x| \cdot |C_i|}$$

を計算し，最大類似度$S_{xc} = \max S_{xi}$となるcを求める．すると文献xは分野cに分類される．

3.1.3 評価項目

評価は，以下の二項目について，既存手法と提案手法による実験結果を比較する事により行う．

∘ **分類精度**　分類精度は類似度計算から求めた分類が，実際にe-printアーカイブの分類と適合した割合である．

∘ **類似度平均**　全文献それぞれについて正解分野の分野基準ベクトルとの類似度を計算し平均を取ったもので，分類の明確さを表す尺度といえる．分野iに属する文献xについて，分野基準ベクトルC_iとの類似度S_{xi}がわかっている時に，類似度平均は，

$$S_{avg} = \frac{1}{|M|} \sum_{x \in M} S_{xi}$$

で算出される．なお，ベクトルはそれぞれ1で正規化しているので，類似度平均の最大値も1である．

3.1.4 実験結果

実験は基準ベクトルを20回取り直して，その平均をとることで結果とする．実験の結果得られた分類精度を図2に，類似度平均を図3示す．

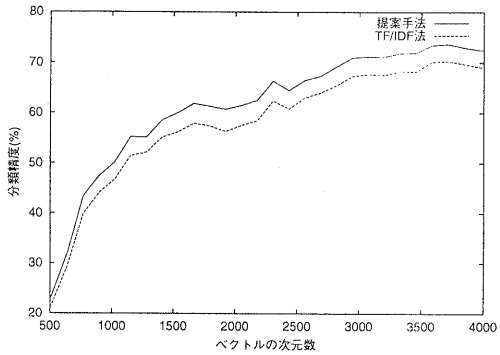

図 2: 分類精度

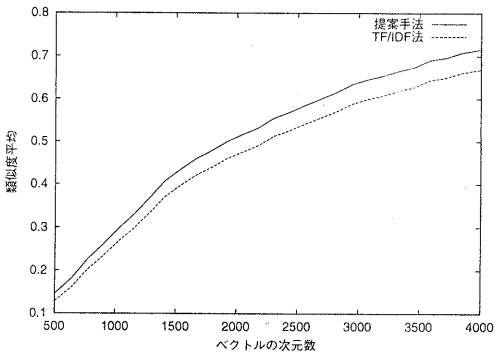

図 3: 類似度平均

3.2 HTML文書を用いた文書検索実験

HTML文書においては，他文書への参照情報としてハイパーリンク構造が利用できる．しかし，ハイパーリンクの種類には，意味的な関連性を表すものから，閲覧者の利便性のためだけに用意されているものまで，その用途が多岐に及ぶ．そのため学術論文のように全ての参照情報を利用する方法では集合形成に適さない．すなわち，本手法では，ハイパーリンクをある程度選別して集合を形成する．具体的には，ハイパーリンクに埋め込まれている参照先のURL情報から以下のリンクを取り除く事で選別を行う．

1. 非HTML文書へのリンク (*.jpg等)

2. ページ内移動 (*.html#???等)

3. トップページ等へのリンク (../index.html 等)

HTML 文書では，これらのページへのリンクを取り除くことにより，参照集合中の不要な文書を削減することができる．

3.2.1 実験データ

検索に用いるクエリーは表 2の 10 種類である．

query1	伊達正宗
query2	松坂大輔
query3	バナナ
query4	大相撲
query5	かきフライ
query6	WWW 情報検索の研究
query7	つくし料理
query8	子供遊びの歴史
query9	猫の飼い方
query10	新潟のアパート情報

表 2: 検索実験に用いたクエリー

これらのクエリーを検索エンジンである goo に入力し，その結果得られる HTML 文書上位 100 文書を実験対象文書とする．

3.2.2 実験手順

(1) 実験対象の HTML 文書に対して，提案手法と，単純な TF/IDF 法に基づく手法を適用して文書ベクトルを生成する．詳細なベクトル生成手順については分類実験と同一であるためここでは省略する．ただし，ベクトル生成に用いる単語は，文書に出現する全単語を用いるものとする．

(2) 生成されたベクトルをもとに，クエリーとの類似度 (スコア) を計算する．ここではスコアを検索クエリーに対応する文書ベクトル中の単語の重みの和とする．検索キーワードが $keyword_1, keyword_2, \ldots, keyword_m$ のとき，文書ベクトル $V = (v_1, v_2, \ldots, v_n)$ を持つ HTML 文書のスコアは，

$$Score = \sum_{i=1}^{m} v_{keyword_i} \qquad (3)$$

によって求められる．

3.2.3 評価項目

評価は，提案手法と既存手法のそれぞれについて平均適合率を算出し，それらを比較することで行う．

3.2.4 実験結果

実験の結果算出された平均適合率を表 3に示す．

クエリー	TF・IDF 法	本手法
query1	0.681185208	0.796585082
query2	0.962138776	0.965700029
query3	0.564365576	0.565532534
query4	0.97704853	0.984230056
query5	0.524597763	0.519478299
query6	0.623558618	0.745826949
query7	0.232081272	0.319088319
query8	0.6592945	0.621205389
query9	0.634505286	0.572670577
query10	0.719111444	0.700225464
平均	0.664541959	0.685016826

表 3: 平均適合率

3.3 実験の考察

分類実験では，提案手法によって，分類精度と類似度平均の両項目が改善されることが確認された．分類精度は平均で 5.9 ポイント程度，類似度平均は平均で 8.4 ポイント程度改善された．一方，検索実験における平均適合率は，全体としてわずかな改善にとどまった．これは対象が HTML 文書であることから，ベクトル拡張によって必要以上にベクトルを構成する単語数が増加し，文書の主題が埋没したことに依るものと考えられる．適合率を改善するための方法としては，例えば，文書の構造解析を行って，より詳細に文書を選別する方法や，辞書を用いて語彙空間の拡張を押える方法などが考えられる．

4 おわりに

本稿では，文書にあらかじめ付与されている他文書への参照情報を利用して参照集合を形成し，その解析から文書ベクトルを拡張する手法を提案した．本手法により，参照集合内に登場する単語群を考慮

することで，語彙曖昧性の問題を，また，参照集合内の単語頻度を考慮することで，単語情報量の問題を軽減することができる．

提案手法の検証を行うために，提案手法と既存手法の双方を用いた文書分類・検索実験を行った．その結果，分類実験では分類精度と類似度平均の双方の評価項目において処理性能の改善が見られた．一方，検索実験では，参照集合内の不要情報の選別が不十分であったこともあり，その改善は僅かなレベルに留まった．これらの実験を通して，ドキュメント相互間の参照情報を利用する本手法は，学術文献やHTML文書（ウェッブページ）といった性質の異なるドキュメントに対して十分適用可能であることが確認された．

参考文献

[1] Salton, G. and McGrill, M.J. : Introduction to Modern Information Retrieval, McGraw-Hill Inc., 1983

[2] 金沢 輝一，高須 淳宏，安達 淳: 文書関連性を考慮した検索方式， 情報処理学会 DBS 研究会,DBワークショップ'98,(48)

[3] **e-print** アーカイブ: http://xxx.yukawa.kyoto-u.ac.jp/