マルチメディア通信と分散処理　94－12
分散システム/インターネット運用技術　14－12
（1999．7．15）

学術論文の引用関係に基づく特徴量の抽出手法

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，学術論文にあらかじめ付与されている引用関係を利用することで，文書ベクトルを拡張する手法を提案する．この手法では引用関係で形成される文献の部分集合を解析し，文献ごとの固有主題と，分野で共有する共通主題を特定し，そこから既存のベクトル化手法にはない主題ごとの単語の情報量を得ることが出来る．評価実験として，共通主題を強調するための文書ベクトル拡張法を示し，文献分類を行ったところ，既存手法と比較して分野基準ベクトルとの類似度平均におよそ20%の改善を確認することができた．

A Method to Extract Features of Scientific Papers Based on Cited Relations

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 vectors based on cited relations of scientific papers. This method analyzes the document sets which consist of cited papers, specify the proper subjects of documents and the common subjects shared among the similar documents. Then both subjects provide the measure of term information which can not be obtained by existing vector method. We propose the vector expansion method to emphasize the common subjects and evaluate our method through classification experiment. Compared the proposed method with the existing one, ratio of classification is improved 20% in the average of similarity.

1　はじめに

近年，コンピュータの普及やネットワークの発達により，電子化された文書が大量に流通するようになっている．

大量の電子化文書から効率良く文書を獲得するためには，あらかじめ各文書の主題概念を表すような特徴量を抽出しておく必要があり，その抽出手法は古くから研究の対象とされている．

統計的に特徴量を表現する方法としては，文書を単語とその重みで表現する文書ベクトルを用いる方法が一般的であるが，各単語の情報量は基本的にコーパス全体に対する単語の出現確率に依存するので，コーパス全体から見るとありふれた語でも，ある分野の文書に限れば特徴的な単語である場合など，単語の局所的な偏りを文書ベクトルに反映できない．

本稿では学術論文の引用関係に着目し，引用関係によって形成される論文の集合を解析することで，コーパス全体の解析からは得られない情報を文書ベクトルに反映する事を目的とする．

次章では，既存のベクトル化手法の問題点を指摘し，これを改善するために本稿の提案手法である引用関係に基づく特徴量抽出手法を提案する．3 章では提案手法に基づいた評価実験の結果を報告する．

2　引用関係を利用した特徴量抽出

2.1　既存手法

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，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）にのみ依存している．こうした手法では，コーパスに収められる文献全てが共通の主題を扱うということが前提となるが，複数の主題に関する文献が同時に収められている場合には，主題ごとに単語の持つ重要度は異なるため，文書の主題は的確にベクトルに反映されないという問題点を持つ．

これを解決するためには，文書の関連性からコーパス内に主題ごとの文書部分集合を形成し，その集合を解析することにより主題ごとの単語の情報量を算出し，文書ベクトルを拡張すればよい．

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問題点がある．

2.2 提案手法の基本概念

引用関係にある論文は共通分野の主題を扱う可能性が高いことから，主題ごとの文書部分集合形成に利用できる．

被引用文献は一般に著者自身の論説の根拠や比較対象として適切だと判断される文献だけが選定される．したがって引用関係によって得られる部分集合内の文書は共通の主題を持っている可能性が極めて高く，論文に付与されたキーワードを使う手法と比較して，文書集合内にノイズとなるような文献は少なくなることが予想される．また著者はあらかじめ引用する文献に目を通していることが想定されることから，引用集合内での著者間のオントロジーの差異も吸収できることが期待される．

ベクトル化の対象となる文献と，その文献が直接引用する被引用文献群から構成される文書集合（以下，引用集合と呼ぶ）は以下のような特性を持つと考えられる．

I. 引用集合内で，ある特定の文書にのみ出現する単語はその文書特有の主題（固有主題）を表す．
II. 引用集合内で多くの文書に共通して出現する単語は，引用集合内の文書で共通する分野の主題（共通主題）を表す．

共通主題と固有主題について概念図を図1に示す．今，文献a,b,cによって引用集合が構成されており，引用関係で結ばれたこれらの文献は共通した分野の問題を扱っている．

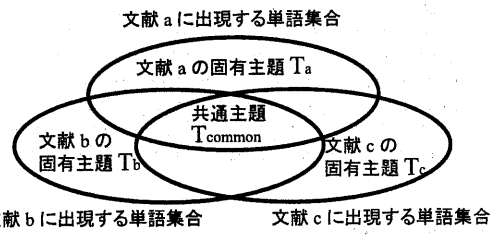

図1: 共通主題と固有主題の概念図

文献aの固有主題を構成する単語群T_aとは，文献aに出現する単語のうち，同じ分野の問題を扱っているにも関わらず，b,cにはそれほど出現しないような単語群である．これらは文献aの新規部分を表現するための単語であると考えられ，文献aの著者にとっては最も強く主張したい部分であるといえる．

文書検索は検索対象となる文書と検索者の意図との照合であるから，共通主題を表す単語群T_{common}よりT_aの重みを増加する事によって，文書ベクトルに固有主題をより強く反映することができ，文書検索の際の適合率向上が見込める．

一方，T_{common}は，その文書の属する分野を適切に表す単語である．文書分類などでは，文献ごとの固有主題よりも分野に共通する共通主題が重要であ

ることから，これを強調する事によって分類精度を向上することが出来る．

2.3 文書分類のための文書ベクトル拡張法

前節で述べた引用集合の特性IIを利用して文書分類のための文書ベクトルを生成する手法を提案する．文書分類ではベクトルが類似する文書を同分野とみなすことから，各文書に付与される文書ベクトルは固有主題よりも，共通主題が強調されている事が望ましい．そこで，文献xの引用集合M_xとした時，文献xに付与される文書ベクトルV'_xの第i要素を，

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

とする．これは文献xの引用関係から構築される引用集合M_x内のTF/IDFベクトルを平均化することを意味している．平均をとることで引用集合内の各文献に固有な単語群は重みを減らし，各文献に共通の単語群は相対的に重みを増す事になるため，共通主題が強調される事になる．

3 文献の分類実験

3.1 実験の目的

2.3で述べた文書分類のための文書ベクトル拡張法について，その有効性を確認するために，提案手法と既存手法の両手法により文書分類の精度を比較する実験を行った．

3.2 実験データ

分類対象の文献データベースとして物理分野の電子化文献アーカイブであるe-printアーカイブ[3]を利用した．このアーカイブは文献が分野毎に管理されており，ここから表1の6分野の論文を対象に，99年5月時点で最新のもの30文献ずつを対象として実験を行った．

提案手法は論文の引用関係に基づくため，分類の対象文献のうち，e-printアーカイブ内の論文を引用していないものは分類対象から外している．e-printアーカイブ内の文献を引用している文献は実験対象180文献中139文献であった．なお，この対象文献に対し抽出された被引用文献数はのべ1062文献である．

表1: 分類実験の対象となる分野

分野ID	分野
hep-th	高エネルギー物理（理論）
hep-lat	高エネルギー物理（格子）
hep-ph	高エネルギー物理（現象）
quant-ph	量子物理
math	数理物理
nucl-th	原子核理論

3.3 実験手順

まず実験対象の文献データに対し，以下の二手法によって文書ベクトルを生成する．

- 2.3で提案した手法
- 単純なTF/IDF法

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．文献中で引用関係は，分野ID+発表年月+通し番号の形式でhep-th9905001のように表現される．これを利用して単純なパターンマッチングで引用関係を抽出できる．
(3) 対象文献からベクトル生成に使用する単語を抽出．文書ベクトル生成時には，理想的にはコーパスに出現する全単語について調べるべきだが，記憶容量と処理速度の都合上，出現頻度の高い単語から選んでいき，累計出現回数が9割になるまでのもの2619単語を使用した．
(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) 2.3で提案した(1)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る事で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基にそれぞれ以下の手順で分類を行う．

○ 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10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で得られる．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．

3.4 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ものの数である．

○ 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．

3.5 実験結果

○ 分類精度

実験の結果得られた分類精度を表2に示す．表中で×→○は提案手法によって分類が改善された文献数を，○→×は逆に正しかった分類が誤った分野に分類されたものの数を表している．

表 2: 実験結果（分類精度）

	TF/IDF	提案手法	×→○	○→×
基準1	125/139	129/139	5	1
基準2	128/139	126/139	2	4
基準3	126/139	127/139	4	3
基準4	126/139	126/139	3	3
基準5	126/139	124/139	2	4

○ 類似度平均

実験の結果，類似度平均は表3のようになった．

表 3: 実験結果（類似度平均）

	TF/IDF 法	提案手法
基準1	0.3439	0.4146
基準2	0.3745	0.4505
基準3	0.3530	0.4232
基準4	0.3525	0.4170
基準5	0.3605	0.4345

○ 提案手法によって分類が改善された文献

5回中4回の実験において分類が改善された論文を例に，既存手法と提案手法の文書ベクトルの重み上位5単語を比較したものが表4である．この論文は，提案手法によって「高エネルギー物理（格子）」→「高エネルギー物理（理論）」と分類が改善された例である．「高エネルギー物理（格子）」分野の典型的な単語「fermion」の重みが下がり，「高エネルギー物理（理論）」分野の典型的な単語「supergrav」「supersymmetre」の重みが増しているのがわかる．また，相対的にTF/IDF法で上位にランクされた固有主題を表す単語は表中から姿を消している．

表 4: 分類が改善された文書ベクトルの具体例

TF/IDF 法	重み	提案手法	重み
fermion	0.5258	**fermion**	0.3399
twelve	0.4583	**supergrav**	0.2825
eight	0.3589	bosonic	0.2099
homogene	0.2228	**supersymmetre**	0.1963
casimir	0.1762	eight	0.1895

○ 次元数による分類精度，類似度平均の違い

ベクトル生成時の次元数による分類精度の違いを調べるために，次元数を 64 ～ 512 と変えて同様の実験を行った結果を表 5, 6 に示す．表中の値は基準ベクトルを変えながら 5 回ずつ実験を行った結果を平均したものであり，改善率は TF/IDF 法と比較した提案手法の改善率を示している．

表 5: 次元数による分類精度の違い

次元数	TF/IDF 法	提案手法	改善率 (%)
64	114.0/139	115.0/139	0.7
128	117.2/139	121.0/139	2.7
256	123.8/139	124.6/139	0.6
512	126.0/139	124.8/139	-0.8
2619	126.2/139	126.4/139	0.1

表 6: 次元数による類似度平均の違い

次元数	TF/IDF 法	提案手法	改善率 (%)
64	0.6754	0.7281	7.8
128	0.6274	0.6810	9.0
256	0.5652	0.6278	11.0
512	0.4909	0.5543	12.9
2619	0.3569	0.4280	19.9

3.6 考察

- 提案手法によって類似度平均はおおむね 20%程度改善された．分類精度の改善は最大で 3%程度にとどまり，逆に分類精度が下がる場合もあった．
- 提案手法によって分類が改善された文献の文書ベクトルでは，分野の共通主題部分が強調されていた．ところが，引用文献に別分野のものが多かったことから，対象文献自身とは別分野の主題が強調されてしまい，正しい分類が誤って分類されてしまったものがあった．
- TF/IDF 法では次元数が低くなると分類精度も下がるのに対し，提案手法では次元数が低い場合でも比較的良好な分類結果が得られている．ベクトルの次元数はそのまま処理時間に影響するので，本手法では低い次元数でも高い分類精度が得られる点で優れている．また，類似度平均は次元数によらず本手法の方が優れており，本手法の有効性が示された．

4 おわりに

本稿では，引用関係を利用してコーパス内に主題ごとの文書部分集合を形成し，文書ベクトルを拡張する手法を提案した．本手法は，引用集合内で特定の文書にのみ出現する単語はその文書特有の主題（固有主題）を表すことから文書検索に有効であり，逆に多くの文書に共通して出現する単語は分野に共通する主題（共通主題）を表すことから文書分類の性能向上に有用であることに着目して構成されたものである．

また提案手法を検証するために，共通主題を強調した文書ベクトルを用いて e-print アーカイブを対象とした文書分類実験を行った．その結果提案手法は TF/IDF 法と比較して，分類精度においては大きな改善効果は得られなかったものの，類似度平均で最大 20%程度の改善が確認された．またベクトルの次元数を変化させて実験を行った結果，比較的低い次元数のベクトルによる文書分類に効果を発揮する事がわかった．

現在，引用集合内の固有主題を利用した検索システムを構築中である．この検索システムでは，文書ごとに固有主題を強調した文書ベクトルを付与することにより，検索時の適合率向上を目指すものである．この結果については別途報告したい．

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/