データベースシステム　127－21
情　報　学　基　礎　67－21
（２００２．５．２２）

文ベクトル集合モデルに基づく文書類似尺度の評価

城塚 音也　北内 啓

株式会社 NTT データ 技術開発本部

{sirotuka,kitauchi}@rd.nttdata.co.jp

類似文書検索、分類、クラスタリング等の近年の計算機による大量文書処理において、文書間の類似度計算には、文書を文書に含まれる単語を次元とするベクトルとして扱うベクトル空間モデルを用いることが主流である。しかしながらベクトル空間モデルでは、文、段落といった文書の構造情報を扱うことが難しいため、文書の構造情報を反映した文書モデルおよび類似尺度が望まれる。本稿では、近年提案された文ベクトル集合モデルに基づく、新しい文書類似尺度を提案する。BMIR-J2 の新聞記事データおよび特許データを用いて文書類似尺度の比較実験を行った結果、従来のベクトル空間モデルと比較して、提案する文書類似尺度が、より文書の構造的類似性を反映していることを確認した。

Evaluation of Document Similarity Measure based on Sentence Vector Set Model

Otoya SHIROTSUKA and Akira KITAUCHI

Research and Development Headquarters
NTT DATA CORPORATION

{sirotuka,kitauchi}@rd.nttdata.co.jp

Vector Space Model (VSM) is popular in machine-based text processing of a large amount of documents such as similar document retrieval, automatic document classification and document clustering. However, VSM has difficulty in utilizing structural information of document like sentences or paragraphs. Therefore, a novel model for documents and document similarity measure is expected which can express the structural information of documents and calculate its structural similarity. In this paper, we propose a method of document similarity measure based on sentence vector set model, which is recently proposed. According to the experimental results with newspaper articles from BMIR-J2 collection and Japanese patent data, we confirmed better expression capability of document similarity of the proposed method compared to existing VSM.

1 はじめに

インターネット、イントラネットの普及により、電子情報としてのテキストデータが大量にネットワーク上を流通し、企業等の組織内に蓄積されるようになってきた。検索、分類、クラスタリング等のテキスト処理は、このようなテキストデータのハンドリングのために欠かせない。一方、現在用いられているテキスト処理は、大量データの高速処理に重きを置いている側面があり、計算量や精度、汎用的な処理への適用の困難性といった問題から、文書中の単語の頻度頻度以外の情報である、テキストの統語、意味レベルの情報の利用が避けられる傾向にある。

現在のテキスト処理で用いられる文書類似尺度はベクトル空間モデル（Vector Space Model: VSM）に基づくものがほとんどであり、基本的に文書を構成する単語の出現頻度分布を表すベクトルで文書を表現する。この方法は文書の類似度計算をベクトル演算として行うことができるため、汎用の圧縮、照合アルゴリズムによる高速な計算が可能となる。しかしながらベクトル空間モデルでは、文、段落といった文書の構造情報を扱うことが難しいため、文書中で使用される単語分布が類似していても、単語が使用される文脈が異なるような文書同士の類似度を、文書の意味内容が異なるのにもかかわらず高く推定してしまうという問題がある。

これに対して、川谷[1][2]が提案している文ベクトル集合モデル（Sentence Vector Set Model: SVSM）に基づいた文書類似尺度は、文書を構成する要素（文）のベクトル集合として表現することから、ベクトル演算レベルの処理でありながら、文単位の単語頻度分布を考慮することで、文書の構造情報の反映が可能であり、VSM と比較してより詳細な文書のモデル化が期待できる。

本稿は、SVSM に基づいた文書類似尺度と、文ベクトルが構成する部分空間の正准角を利用した新しい類似度計算方法について述べ、本文書類似尺度と VSM との文書類似度比較評価実験について報告する。

まず、2 章で提案する文書類似尺度について説明し、3 章にて、新聞記事及び特許データを用いた提案する文書類似尺度を用いた文書類似度類比較実験について報告する。4 章では本論文のまとめと、今後の課題を述べる。

2 文ベクトル集合モデルに基づく文書類似尺度

2.1 文集合モデルの定義

SVSM について説明する。図 1 に示すように、文書には、段落の集合、段落は文の集合、文は単語の集合という多階層構造が存在する。文書ベクトル集合モデルでは、この階層関係を文書-文-単語という 3 階層としてとらえ、モデル化している。

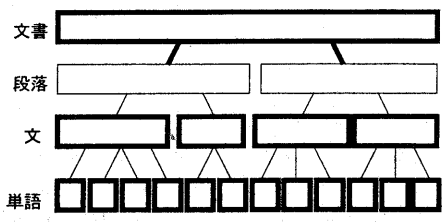


図 1: 文書の階層構造

当然、SVMS の中間層である文は、文に限ったものである必要は無く、部分文（フレーズ）、段落やそれらの組み合わせを使用することが可能である。3 章の実験では、文以外の中間層を使用した検討についても報告する。

2.2 類似尺度

本節では、文ベクトル集合からなる文書どうしの類似尺度の算出方法について説明する。

まず、Watanabe の部分空間法[3]と同様に、文ベクトル集合に対して KL 展開を施すことにより部分空間を求める。文書 d に n 個の文が含まれているとする。文ベクトル s の自己相関行列

$$R = \sum_{s \in d} s^t / n \qquad (1)$$

を計算する。R を対角化して得られた固有ベクトルから、累積寄与率などの基準により固有値が大きい順に m 個の固有ベクトル v_i $(i=1,\ldots,l)$ を取り出す。この固有ベクトルが、求める部分空間の基底である。求められた部分空間は、文ベクトル集合の分布をもっともよく近似するものとなっている。

次に、文書類似尺度を求める。文書類似尺度は相互部分空間法[4]により、文ベクトル集合の部分空間どうしの正準角として求められ

る。二つの文書の文ベクトル集合に対する部分空間の基底をそれぞれ v_i $(i=1,\ldots,l)$, w_j $(j=1,\ldots,m)$ とするとき、

$$X_{ij} = \sum_{k=1}^{m} (v_i \cdot w_k)(w_k \cdot v_j) \qquad (2)$$

によって計算される行列 X についての固有値問題

$$\lambda x = Xx \qquad (3)$$

を考える。得られる最大の固有値が部分空間どうしの角度、すなわち文書類似尺度となる。

川谷は文書間の類似度を、文ベクトルのすべての組み合わせについて求めた内積の 2 乗和をもとに定義している。これに対し相互部分空間法では文ベクトル集合の部分空間どうしの角度を求めており、より直観的で自然な方法で類似度を算出しているといえる。

2.3 他手法との比較

VSM の問題である、構造情報の脱落についての解決方法として今までに提案されている手法としては、例えば、共起情報を利用するもの[5]、係り受け情報を利用するもの[6]がある。

前者は、VSM の改良方法として有効であるが、単語の出現分布とともにすべての共起情報を次元として管理する必要があるため、大規模な文書データを扱う際必要なメモリ空間が非常に大きくなるという問題をクリアする必要がある。

後者に関しては、頻度情報に比べてより直接意味内容を表現している係り受け情報を扱っている点が優れているが、係り受け情報の取得のための計算コストが高く、また、係り受け解析自体もある程度の誤った解析が発生することが避けられないという問題をクリアする必要がある。SVSM では、単語頻度ベースの表現能力ではありながらも、[5], [6]と比較して汎用的な文書構造情報反映の枠組みを目指している。

3 文書類似度比較実験

前述した文ベクトル集合モデルに基づく文書類似尺度と従来のベクトル空間モデルに基づく文書類似尺度の比較を目的に類似文書検索実験を行った。

実験に用いた文書データはあらかじめ人手により、以下の 3 種類の文書集合に分けられているものである。

I. ある話題が文書の主題となっている文書集合
II. ある話題が文書中に見られるが主題ではない文書集合
III. 話題に関係した単語が文書に含まれるが、内容としては関係ない文書集合

実験では、提案手法および従来方法により各文書間の類似度を求め、文書集合内および文書集合間の文書類似度の分布を観察することにより両手法の比較を行った。

3.1 実験データ

実際に実験に使用したデータについて述べる。実験には 2 種類のデータ、BMIR-J2 テストコレクション[7]及び、特許データを使用した。

BMIR-J2 は、94 年の毎日新聞 5080 記事に対して、60 の検索要求が用意されており、それぞれの要求に対して、適合した話題が主題となっている記事に A、検索要求に適合した話題がわずかに見られる記事に B、検索要求を構成する単語を含むが内容としては関係ない記事に C の 3 種類の記号が付与されている。本実験では、検索要求 n に対応する文書集合 A_n、B_n、C_n を文書集合Ⅰ、Ⅱ、Ⅲとして実験に用いた。

特許データは 1995 年の公開公報データ中から同じ国際特許分類（IPC）コード[8]を持つものを抽出することにより文書集合を作成した。使用した内容は特許文書中の特許請求の範囲の部分である。

IPC コードは 5 階層（上位階層からセクション、クラス、サブクラス、グループ、サブグループ）、6 万余の分類体系であり、同一の IPC コードを持つものは、ある程度類似した内容をもつ。特に、下位階層の IPC コードで

はその傾向が強い。第5階層の同一サブグループから抽出した文書集合を文書集合Ⅰ、第4階層の同一グループから抽出した文書集合を文書集合Ⅱ、第3階層の同一サブクラスから抽出した文書集合を文書集合Ⅲとして実験に用いた。用いたデータセットの数は30セット、総文書数は約450である。

一文書あたりの平均文数はBMIR-J2が11.6文、特許が25.2文であり、特許データのほうが2倍以上文数が多い。

3.2 実験パラメータ

実験パラメータとして、各単語の出現頻度の重み付け方法、文ベクトルの作成方法に注目した。出現頻度の重み付け方法は、TF·IDF、TF、出現の有無（0 or 1）の3通りである。

2.1節において述べたように、SVSMでは、通常の文以外の文書構成単位を使用することが可能である。そこで我々は、文ベクトルの生成の際に対象文の前後の文内容を取り込むことにより対象文のコンテキスト情報を反映させることを考えた。図2に示すように、具体的には、TextTiling[9]で使用される窓掛け手法を利用して、対象文を中心とした$2n+1$（n＝考慮する対象文前後の文数）の幅のウィンドウを用意し、文書を構成する文それぞれについてウィンドウ内の文から文ベクトルを生成する。

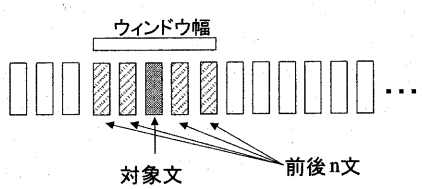


図2: 窓掛けによる文ベクトル生成

コンテキスト情報を反映した対象文のベクトルVの計算には以下の式を使用する。

$$V = \alpha\sum_{i=0}^{n-1} V_{k-i-1} + \alpha\sum_{i=0}^{n-1} V_{k+i+1} + (1-2\alpha)V_k \tag{4}$$

但し、V_kは対象文Kのベクトル、nは考慮する対象文前後の文数、αは前後の文ベクトルに対する重みである。

3.3 評価方法

検索、分類といった利用方法を念頭に置いた評価指標として、Leave One Out法を用いた交差検定による分類性能評価法を使用した。文書分類アルゴリズムには1-NN法を用いた。具体的には、データセットⅠに属するデータを一つづつ取り出し、取り出したデータに最も類似度の高いデータが所属するカテゴリを、取り出したデータに対する推定カテゴリとした時の正しくカテゴリを推定できた比率を比較することにより類似尺度の性能比較を行う。

3.4 実験結果

表1に実験結果を示す。実験結果を見ると、SVSMに基づく文書類似尺度はVSMと比較して、特許に関してはある程度性能の優位性が見られるが、新聞記事に関しては逆に劣っている。この原因は、両データの平均文数の違いと推測する。提案手法では、文書表現に使用できる特徴次元数は最大で文書に含まれる文数と比例するため、文書に含まれる文数が少ないと、文書特徴を充分に表現できなくなる。そのため平均文数の少ない新聞記事において提案手法による分類正解率が低かったと見られる。

文書ベクトルの要素に対する単語の重み付けに関しては、TF·IDFが最も良く、次いでTF、重みなしの順であった。

表1: 分類正解率

（新聞記事）

類似尺度	重み付け	分類精度(%)
SVSM	TF·IDF	62.3
	TF	61.9
	なし	70.6
VSM	TF·IDF	76.8
	TF	72.2
	なし	73.8

（特許）

類似尺度	重み付け	分類精度(%)
SVSM	TF·IDF	58.0
	TF	54.0
	なし	52.9
VSM	TF·IDF	50.3
	TF	50.7
	なし	47.1

表 2 に窓掛けによる文ベクトルの重み付けを使用した場合の結果を示す。文ベクトルへの文脈情報の反映に関して、その実験結果を見てみると、文単独のベクトルを使用するSVSM と比較して特に良好な結果は得られなかった。これは文間の構造（文脈情報）が文内構造に比べて、文書類似性に対する影響力が少なかったため、文脈を考慮することによって逆に文内構造の表現が鈍ってしまったためと考える。

表 2: 分類正解率
（特許、文脈重み付けあり）

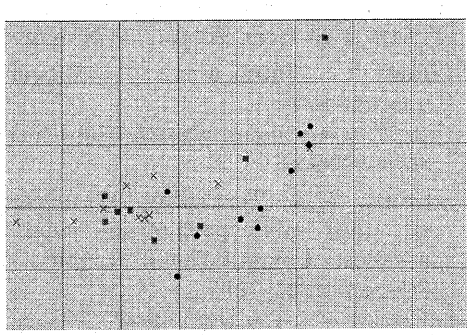
類似尺度	重み付け	分類精度(%)
SVSM	TF･IDF	57.3
	TF	55.0
	なし	53.3

4 分析

4.1 文書の視覚化

図 3 にⅠ、Ⅱ、Ⅲの各カテゴリに属するデータ同士の類似度の関係を二次元に視覚化したものを示す。視覚化手法には、クラスタ構造に着目した判別分析に基づく二次元視覚化手法[10]を使用した。この視覚化では多次元尺度法と同様に、データ間の関係の強さをデータ間の距離となるように低次元の空間上に配置する。視覚化するデータセットⅠ内の類似度を●、ⅠとⅡの間の類似度を■、ⅠとⅢの間の類似度を×で表している。

VSM による結果



SVSM による結果

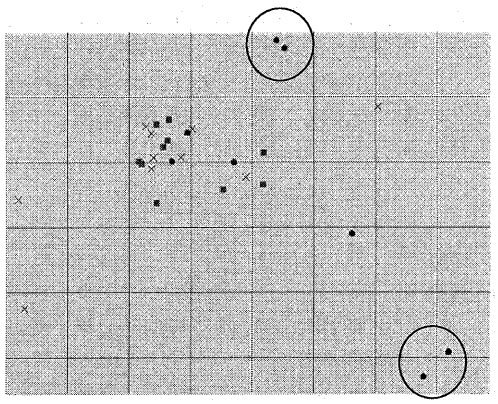


図 3: 各カテゴリ間の類似度視覚化結果

VSM による結果は、●、■、×の 3 種類のデータにおけるクラスター構造が、かなり重複しながらも観察できることが分かる。それに対して、SVSM では、3 種類のデータが集中しておらず、画面中央上部や右下の●のように、他のデータから離れた状態で、少数の●が固まっており、SVSM では、特定のタイプの文書に対して、高い類似度を示す類似尺度であることが分かる。

しかしながら SVSM では、画面中央のデータが集中している部分が示すように、VSM と比較してクラスター構造が判然としない。これは、3.4 節で考察したように、提案する手法が、文ベクトルを複数のベクトルで表現される部分空間に変換する際、文書に含まれる文数によって使用できるベクトルの数の最大値が決まってしまい、オリジナルの文の内容に含まれる情報が失われてしまうという問題が原因と推測している。

4.2 文書データ

BMIR-J2 のデータセットⅠに所属するデータを例に SVSM により高い文書類似度を示した文書の特徴について分析する。表 4 に分析した 9 つの類似記事を示す。これらの記事は日本の航空会社関係の記事であるが、特に記事番号 1、2、3、5 は日本航空の経営不振による人的コスト削減策が話題となっており、これらについては VSM、SVSM 共に高い類似度

を出している。

SVSM の類似度に注目してみると特に、1と5、3と5の記事ペアの類似度が高い。両者の文に共通して含む単語セットを調べてみると、表5に示すとおり、複数の共通単語を含む文が多く存在し、高い類似度の原因と予測される。

表3 データセットIの例

記事番号	内容
1	日航赤字のため人削減。コスト削減の取り組み
2	日航赤字のため人削減。赤字の原因の説明
3	日航業績不振のためスチュワーデスの配置転換
4	日航の新規導入中型機 J Bird について
5	日航従業員削減策の説明
6	スチュワーデス就職難について。航空各社赤字が原因。
7	スチュワーデス・整備士単身赴任による家庭崩壊、育児問題
8	日航時給制スチュワーデス採用。人件費節約のため。
9	時給雇用スチュワーデスに運輸大臣待った

表4 共通する単語の例

記事ペア	文内に共起する単語セット
1-5	三月期　赤字　見込まれる
	希望　退職者　募集
	スチュワーデス　地上職
	日本航空　従業員　削減　計画
3-5	当社　赤字　見込まれる
	スチュワーデス　乗客　乗務員
	人　削減　計画

5 おわりに

本論文では、文の構造情報を反映した文書モデルである SVSM による文書類似尺度を提案し、新聞記事データ及び特許データを使用してその評価を行った。

実験結果から、提案する文書類似尺度は、文書の構成する文単位での単語分布の類似性に鋭敏な類似尺度であることが分かった。

しかしながら、提案する文書類似尺度は文書分類タスクに対して、一部のデータセットにおいては優れた性能を示したが、全体的には VSM を凌ぐ性能を達成できていない。今後は、SVSM の特徴を生かしつつ、構成単位の多様化、複数の構成単位の併用化による本文書類似尺度の性能向上に取り組む予定である。

参考文献

[1] 川谷: 文ベクトル集合モデルによるテキスト処理, 情報処理学会自然言語処理研究報告, 2000-NL-140, pp.31-38, 2000.

[2] 川谷: 文ベクトル集合モデルによるテキスト処理（Ⅱ）, 情報処理学会自然言語処理研究報告, 2001-NL-143 pp.1-8, 2000.

[3] S. Watanabe and N. Pakvasa, Subspace method of pattern recognition, Proc. 1st IJCPR, pp.25-32, 1973.

[4] 前田賢一, 渡辺貞一, 局所構造を導入したパターン・マッチング手法, 信学論(D), vol.J68-D, no.3, pp.345-352, 1985.

[5] 高木他: 単語共起関係を用いた文書重要度付与の検討, 情報学基礎研究会報告 96-FI-41-8, 情報処理学会, 1996.

[6] 村松他: 単語間の係り受け関係を用いた情報検索手法の評価, 情報処理学会論文誌, pp.22-31, Vol.41 No.SIG1(TOD5), Feb,2000.

[7] T. Sakai, et al., BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems, SIGIR Forum, Fall 1999, Volume 33 Number 1

[8] World Intellectual Property Organization, http://www.wipo.int/classifications/fulltext/ipc/ipc6en/.

[9] Hearst, M.A. TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1), pp.33-64, 1997.

[10] 末永他: クラスタ構造に着目した特徴空間の可視化, 電子情報通信学会論文誌 Vol. J-85 D-II, (2002.5 掲載予定).