1G3-4

Earth Mover's Distanceを用いたテキスト分類

Text Categorization Using Earth Mover's Distance

柳本 豪一[*1] 大松 繁[*1]
Hidekazu Yanagimoto Sigeru Omatu

[*1]大阪府立大学
Osaka Prefecture University

We propose a text categorization system using Earth Mover's Distance (EMD) as similarity measure between documents. Many text categorization systems adopt the Vector Space Model and us cosine similarity as similarity measure between documents. There is an assumption that each of words included in documents is uncorrelated because of orthogonal vector space. However, the assumption is not desirable when a document includes a lot of synonyms and polysemic words. The EMD does not demand the assumption because it is computed as a solution of a transportation problem. To compute the EMD in consideration of dependency among words, we define the distance between words, which needs to compute the EMD, using a co-occurrence frequency between the words. We evaluate our proposing method with ModApte split of Reuters-21578 text categorization test collection and confirm that the proposing method improves a precision rate for text categorization.

1. はじめに

電子化情報の増加にともない，情報を内容に応じてクラスタに分類するテキスト分類の研究が盛んになっている．テキスト分類は情報検索，パターン認識，機械学習などのさまざまな研究成果を用いて実現されている [岸田 03, Setabstiani 02]．分類される情報は自然言語で記述された文書であり，ベクトル空間法 [Salton 75] により索引語を要素とするベクトルとして表現されることが多い．そして，文書間の類似度としてコサイン類似度がよく用いられる．コサイン類似度を用いる際，文書が表現されるベクトル空間は直交空間であると仮定されている．この仮定は索引語が互いに無相関であるというと言うことを表わしている．しかし，同義語や多義語などの単語間の関連性や特定の単語同士が頻繁に共起するという状況を考えると，文書で用いられている単語が互いに無相関であると言う仮定は妥当とは言えない．そこで，索引語間の相関性を考慮した文書類似度を考えることが必要である．

このような観点から新しい文書類似度に取り組んだ研究として Earth Mover's Distance[Rubner 00] を文書類似度として用いた Wan らの研究 [Wan 05] があげられる．この論文では，文書間の類似度として類似画像検索でよく用いられる Earth Mover's Distance (EMD) を文書類似度として用いることで，同義語や多義語の問題に対応している．具体的には，WordNet を用いて索引語間の距離を定義し，索引語間の関連性に配慮した文書類似度を求めている．しかし，この手法は WordNet を用いているため，新語への対応，対象文書特有の情報を十分利用できないというシソーラスを用いた手法一般が有する問題を抱えている．また，言語学的な観点から単語を分類しているシソーラスから単語間の関連性を距離という数値で表現するという問題も含んでいる．他に EMD を用いた研究として，画像検索においてキーワードと画像特徴量を組み合わせた研究 [竹内 07] がある．この論文では，本来異質な画像特徴量とキーワードを統一的に扱って類似度を計算することが実現され，EMD が有効であると報告されている．キーワード間および画像特徴量とキーワード間の距離は共起頻度を用いて定義されており，上記の問題点を回避することができている．この手法が対象とする単語は画像に割り当てられた単語群であり，一般的な文書が対象とはなっていない．したがって，EMD が一般的なテキスト分類において有効であるかどうかは不明である．

本論文では，Earth Mover's Distance を文書類似度として用いたテキスト分類手法の提案を行う．とくに上記で述べた索引語間の距離を定義するという問題に対応するため，単語の共起情報を用いた距離の定義方法を提案する．これによって，対象文書に対する索引語の網羅性の問題，索引語間の距離の数値化に関する困難さを回避する．よって，従来類似度計算時になされていた索引語間の無相関の仮定をゆるめることを実現し，テキスト分類の性能改善を実現する．

以下では，2 章で提案手法である EMD および EMD を文書に応用する方法について説明する．3 章では Reuter-21578 を用いたテキスト分類実験を行い本手法と従来のコサイン類似度，DICE 係数，JACCARD 係数を用いた手法と比較し，本手法の有効性を確認する．さらにコサイン類似度で定義される近傍と EMD で定義される近傍について議論し，本手法の検討を行なう．最後に 4 章において本論文をまとめる．

2. 提案手法

提案手法である Earth Mover's Distance を用いたテキスト分類手法について説明する．まず，分布間の距離を求める Earth Mover's Distance について説明を行い，つぎに Earth Mover's Distance の文書への応用について述べる．

2.1 Earth Mover's Distance

EMD は分布間の距離を表すものであり，類似画像検索の分野でよく用いられている．EMD は，分布間の距離の計算を輸送問題としてとらえ，最適な輸送コストを用いて定義される．EMD を求める際，2 つの分布はシグニチャとして表現される．一方の分布 P をシグニチャとして表現すると，$P = \{(\mathbf{p}_1, w_{\mathbf{p}_1}), \cdots, (\mathbf{p}_m, w_{\mathbf{p}_m})\}$ となる．今，分布 P は m 個の特徴量で表現されており，$\mathbf{p}_i$ は特徴量ベクトル，$w_{\mathbf{p}_i}$ はその特徴量に対する重みである．同様に，もう一方の分布 Q もシグニチャで表すと，$Q = \{(\mathbf{q}_1, w_{\mathbf{q}_1}), \cdots, (\mathbf{q}_n, w_{\mathbf{q}_n})\}$ となる．EMD の計算は，2 つの分布において特徴量の数が異なっている場合でも計算が可能であるという特徴を持っている．今 $\mathbf{p}_i$ と $\mathbf{q}_j$ の

A: 柳本 豪一，大阪府立大学，堺市中区学園町 1-1，072-254-9275，072-254-9909，hidekazu@cs.osakafu-u.ac.jp

距離を d_{ij} とし，全特徴間の距離を $\mathbf{D} = [d_{ij}]$ とする．ここで，$\mathbf{p}_i$ から $\mathbf{q}_j$ への輸送量を f_{ij} とすると，全輸送量は $\mathbf{F} = [f_{ij}]$ となる．ここで，式 (1) に示すコスト関数を最小とする $\mathbf{F}^*$ を求め，EMD を計算する．

$$\mathrm{WORK}(P, Q, F) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} \tag{1}$$

ただし，上記のコスト関数を最小化する際，以下の制約条件を満たす必要がある．

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq n \tag{2}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{\mathbf{p}_i} \qquad 1 \leq i \leq m \tag{3}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{\mathbf{q}_j} \qquad 1 \leq j \leq n \tag{4}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_{\mathbf{p}_i}, \sum_{j=1}^{n} w_{\mathbf{q}_j}) \tag{5}$$

ここで，式 (2) は輸送量が正であることを表し，$\mathbf{p}_i$ から $\mathbf{q}_j$ に送られる一方通行であることを表している．式 (3) は輸送元である $\mathbf{p}_i$ の重み以上に輸送できないことを表す．式 (4) は輸送先である $\mathbf{q}_j$ の重み以上に受け入れることができないことを表す．最後に式 (5) は総輸送量の上限を表し，それは輸送先または輸送元の総和の小さい方に制限されることを表す．

以上の制約条件の下で求められた最適な全輸送量 $\mathbf{F}^*$ を用いて，分布 P，Q 間の EMD を以下のように求める．

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} {f_{ij}}^*}{\sum_{i=1}^{m}\sum_{j=1}^{n} {f_{ij}}^*} \tag{6}$$

ここで，最適なコスト関数 $\mathrm{WORK}(P, Q, F^*)$ を EMD としてそのまま用いないのは，コスト関数は輸送元もしくは輸送先の重みの総和に依存するので，正規化することによってその影響を取り除くためである．

2.2 Earth Mover's Distance を用いたテキスト分類

文書間の EMD を求めるためには，文書をシグニチャとして表現する必要がある．文書は自然言語で書かれているため，特徴量ベクトル $\mathbf{p}_j$ を索引語 t_j，特徴量の重み $w_{\mathbf{p}_j}$ を索引語 t_j の重み w_j^i とする．したがって，文書 Doc_i は $\{(\mathrm{t}_i, w_1^i), \cdots, (\mathrm{t}_m, w_m^i)\}$ としてシグニチャで表現される．ここで，重み w_j^i は式 (7) に示す tf*idf より求める．

$$w_j^i = \mathrm{tf}_j^i \log \frac{N}{\mathrm{df}_j} \tag{7}$$

ここで，tf_j^i は文書 Doc_i における索引語 t_j の出現頻度，df_j は索引語 t_j を含む文書数，N は全文書数を表す．

EMD の計算時に必要となる特徴量間の距離 d_{ij}，つまり索引語間の距離の定義方法について説明する．本論文では，索引語の共起を用いて索引語間の距離を定義する．これにより，シソーラスなどの辞書が不要となるとともに，テキスト分類を行う文書群に固有な特徴も捉えることができる．まず索引語 t_i，t_j の共起回数を $\mathrm{occur}(\mathrm{t}_i, \mathrm{t}_j)$ とし，式 (8) を用いて条件付き確率を求める．ここで，$\mathrm{occur}(\mathrm{t}_i, \mathrm{t}_j)$ は索引語 t_i と t_j が同時に出現している文章数を表す．

$$P(\mathrm{t}_j|\mathrm{t}_i) = \frac{\mathrm{occur}(\mathrm{t}_i, \mathrm{t}_j)}{\sum_{j:j\neq i} \mathrm{occur}(\mathrm{t}_i, \mathrm{t}_j)} \tag{8}$$

条件付き確率 $P(\mathrm{t}_j|\mathrm{t}_i)$ は索引語間の関連性を表しているが，関連がある索引語ほど大きな値となり，距離とは異なった特性を持っている．したがって，この条件付き確率を用いて距離を定義するため，以下の式により距離 d_{ij} を求める．

$$d_{ij} = 1 - P(\mathrm{t}_j|\mathrm{t}_i) \tag{9}$$

ここで，確率の基本的な性質である $P(\mathrm{t}_j|\mathrm{t}_i) \leq 1$ を用いている．

今シグニチャにおける索引語の重みの総和を 1 とすると，EMD は必ず 1 以下となるので，文書間の類似度 $\mathrm{sim}_{\mathrm{EMD}}(\mathrm{Doc}_i, \mathrm{Doc}_j)$ を以下のように定義する．

$$\mathrm{sim}_{\mathrm{EMD}}(\mathrm{Doc}_i, \mathrm{Doc}_j) = 1 - EMD(\mathrm{Doc}_i, \mathrm{Doc}_j) \tag{10}$$

カテゴリの推定には，本論文では Yang らの手法 [Yang 94] を採用する．この手法は文書に対するカテゴリを決定するために k-NN 法を用いており，ラベル付き文書とカテゴリに対する条件付き確率を用いてラベルなし文書とカテゴリの関連度を求め，その関連度により推定カテゴリを決定する．提案手法では，類似度 $\mathrm{sim}_{\mathrm{EMD}}(\mathrm{Doc}_i, \mathrm{Doc}_j)$ により K 個の文書を選択することとする．カテゴリ c_k と文書 Doc_i の条件付き確率 $P(c_k|\mathrm{Doc}_i)$ を以下のように求める．

$$P(c_k|\mathrm{Doc}_i) = \frac{\mathrm{Doc}_i \text{ における } c_k \text{ の出現回数}}{\mathrm{Doc}_i \text{ に割り当てられた全カテゴリ数}} \tag{11}$$

これはラベル付き文書には複数のカテゴリが割り当てられていると仮定しているためである．ラベルなし文書 Doc とカテゴリ c_k の関連度 $\mathrm{rel}(c_k|\mathrm{Doc})$ は式 (11) の条件付き確率を用いて，以下のように定義する．

$$\mathrm{rel}(c_k|\mathrm{Doc}) = \sum_{i=1}^{K} P(c_k|\mathrm{Doc}_i)\mathrm{sim}_{\mathrm{EMD}}(\mathrm{Doc}, \mathrm{Doc}_i) \tag{12}$$

$P(c_k|\mathrm{Doc}_i)$ と $\mathrm{sim}_{\mathrm{EMD}}(\mathrm{Doc}, \mathrm{Doc}_i)$ の特性より，$\mathrm{rel}(c_k|\mathrm{Doc})$ は関連のあると考えられるカテゴリに対して大きな値を持つこととなる．ラベルなし文書に対しても複数のカテゴリが割り当てられることが考えられるが，本論文では 1 つのカテゴリを割り当てることとする．したがって，文書 Doc への推定カテゴリ $\hat{c}$ は関連度 $\mathrm{rel}(c_k|\mathrm{Doc})$ が最大となるカテゴリとする．

$$\hat{c} = \arg\max_{c_k} \mathrm{rel}(c_k|\mathrm{Doc}) \tag{13}$$

3. 実験

提案手法の有効性を確認するため Reuters-21578[Reuters] を用いたテキスト分類実験を行った．まず，Reuters-21578 から訓練文書とテスト文書を抽出するため，ModApte-split を施した．つぎに，SMART stop-list[SMART] を用いて不要語を削除し，Porter アルゴリズムによりステミング [Porter 80] を行うことで，訓練文書とテスト文書から索引語を取り出した．ただし，数字のみで構成される文字列，2 文字以下の文字列は索引語から削除した．以上の処理により，7,733 の訓練文書，3,008 のテスト文書が得られた．表 1 に得られた訓練文書，テスト文書の構成を示す．表中の索引語数はユニークな索引語の数を表している．

テスト文書中の 3,407 の索引語は訓練文書に含まれなかったので，これらの索引語は類似度計算に用いなかった．また，テスト文書に割り当てられているカテゴリについても，訓練文書に含まれないカテゴリが 3 つ含まれていたが，その文書は削

表 1: ModApte-split により得られた訓練文書とテスト文書の構成

	文書数	索引語数	カテゴリ数
訓練文書	7,733	17,488	115
テスト文書	3,008	10,731	93

除せず実験を行った．以下では，訓練文書をラベル付き文書として実験を行った．

比較手法としては，コサイン類似度，DICE 係数，JACCARD 係数の 3 種類の類似度 [徳永 99] を用いてラベル付き文書を選択する手法を採用した．したがって，提案手法とは K 個のラベル付き文書を選択するための類似度計算にのみ違いがあることとなる．コサイン類似度は以下のようにして求めた．

$$\mathrm{sim}_{\cos}(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i {\mathbf{d}_j}^{\mathrm{T}}}{||\mathbf{d}_i||_2 ||\mathbf{d}_j||_2} \tag{14}$$

ここで，$\cdot^{\mathrm{T}}$ は行列の転置を表し，$\mathbf{d}_i$ はドキュメントベクトルを表し，式 (7) の w_j^i を用いて，$\mathbf{d}_i = (w_1^i, w_2^i, \cdots, w_V^i)$ となり，V は全語彙数 (17,488) を表す．そして，$||\mathbf{d}_i||_2$ は 2 次ノルムを表し以下のように計算される．

$$||\mathbf{d}_i||_2 = \sqrt{\sum_{j=1}^{V} {w_j^i}^2} \tag{15}$$

DICE 係数は以下のようにして求めた．

$$\mathrm{sim}_{\mathrm{DICE}}(\mathbf{d}_i, \mathbf{d}_j) = \frac{2\mathbf{d}_i \mathbf{d}_j^{\mathrm{T}}}{||\mathbf{d}_i||_2^2 + ||\mathbf{d}_j||_2^2} \tag{16}$$

JACCARD 係数は以下のようにして求めた．

$$\mathrm{sim}_{\mathrm{JACCARD}}(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \mathbf{d}_j^{\mathrm{T}}}{||\mathbf{d}_i||_2^2 + ||\mathbf{d}_j||_2^2 - \mathbf{d}_i \mathbf{d}_j^{\mathrm{T}}} \tag{17}$$

テキスト分類の評価には精度 (Precision) を用いることとし，以下の式で精度を求めた．

$$\mathrm{Precision} = \frac{\text{正しいカテゴリが推定できた文書数}}{\text{全文書数}} \tag{18}$$

本実験では 1 つのカテゴリのみを推定するので，複数のカテゴリが割り当てられたテスト文書もあるため，推定されたカテゴリが割り当てられたカテゴリに含まれる場合には正解とした．

以上の条件で提案手法 (EMD1)，同一の索引語の距離を 0 としてそれ以外はすべて 1 として EMD を求めた場合 (EMD2) と従来手法 (VSM, DICE, JACCARD) を用いてテキスト分類を行った結果を図 1 に示す．ここで，EMD2 は各索引語が無相関であると考えて距離を設定した場合に対応する．各手法とも k-NN 法を用いているため，テキスト分類の精度はカテゴリ推定に用いるラベル付き文書数 K に依存すると考えられるので，K の値を 1，10，20，30，40，50 と変化させて実験を行った．

この実験結果より，EMD1 はすべての K の値で他の手法より優れており，以下 EMD2，VSM，JACCARD，DICE という順番となった．DICE と JACCARD はほぼ同じような精度を示しており，他の 3 手法に比べて大きく精度が悪くなった．一方，EMD2 は EMD1 と VSM の中間的な振舞を示した．精度

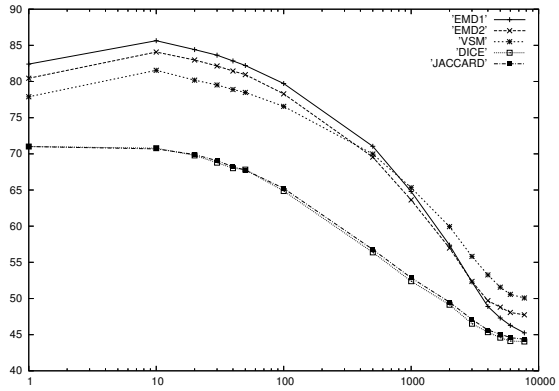

図 1: カテゴリ推定に用いるラベル付き文書数 K とテスト文書に対するカテゴリの精度の関係

表 2: 提案手法 (EMD1) と従来手法 (VSM) により分類されたテキストの詳細

		VSM	
		correct	error
EMD1	correct	2,359	217
	error	94	338

が最大となった K の値について見ると，$K = 10$ の時に EMD1 と EMD2 と VSM で，$K = 1$ の時に DICE と JACCARD で高い精度を示した．$K = 10$ での精度を比較すると，EMD1 では 85.6%，EMD2 では 84.1%，VSM では 81.6%，DICE で 70.7%，JACCARD で 70.7%であり，EMD1 が VSM に比べ 4.0%，EMD2 に比べ 1.5%精度を改善した．ただし，$K = 1$ の時の DICE と JACCARD の精度はともに 71.0%である．以下では EMD1 と EMD2 と VSM に注目して検討を行なう．

表 2 から表 4 に各手法によって分類された文書の内訳を示す．表 2 より，従来手法では誤ったカテゴリを推定した文書が 555 文書あったが，提案手法では 432 文書と減少した．しかし，もう少し詳しくこの結果を見ると，提案手法は従来手法で正しくカテゴリ推定できなかった 217 文書を正しく推定できている反面，従来手法で正しく推定できていた 94 文書に対して誤ったカテゴリを推定する結果となっていた．表 3 は索引語の無相関と仮定した手法は，従来手法では正しく推定できていた文書を誤ったカテゴリにしてしまう件数が多くなっている点に違いが見られたが，その他の値は表 2 と似た傾向を示していた．表 4 では，提案手法により索引語の無相関という仮定によって誤ったカテゴリ推定を行った文書を正しく推定できるようになった．

以上の実験結果より，精度の点からは提案手法が従来のコサイン類似度を用いた手法に比べテキスト分類の性能を改善することが分かった．これは文書類似度として EMD を用いた効果と索引語の相関性を考慮した効果の 2 点が考えられる．まず，VSM と EMD2 を比較すると，ともに索引語間の無相関性を仮定しているが，EMD2 の方が良い結果となっていた．これは，EMD を用いた効果と考えられる．例えば，$(\frac{\sqrt{3}}{2}, \frac{1}{2})$ と $(1, 0)$ および $(\frac{1}{2}, \frac{\sqrt{3}}{2})$ の類似度を考えると，コサイン類似度ではともに 0.866 と同じ値となるが，EMD では 0.866 および 0.732 と異なった値となる．これは，$(\frac{\sqrt{3}}{2}, \frac{1}{2})$ と $(1, 0)$ では，ともに 1 つ目の要素が 2 つ目の要素より大きくなっているため，輸送コストが小さくなるためであると考えられる．つまり，コ

表 3: 索引語の無相関性を仮定した手法 (EMD2) と従来手法 (VSM) により分類されたテキストの詳細

		VSM	
		correct	error
EMD2	correct	2,308	221
	error	145	334

表 4: 提案手法 (EMD1) と索引語の無相関性を考慮した手法 (EMD2) により分類されたテキストの詳細

		EMD2	
		correct	error
EMD1	correct	2,502	74
	error	27	405

サイン類似度で定義される近傍と EMD で定義される近傍とが異なっていることを表している．この影響が，EMD2 と VSM の精度の違いの原因の一つであると思われる．次に EMD1 と EMD2 を比べると，EMD2 がよい結果となっており，これは索引語間の相関性を考慮した効果であると考えられる．EMD1 と EMD2 ではともに文書類似度として EMD を用いているので，カテゴリ推定の精度の違いは距離の設定に起因するからである．よって，共起情報を用いて索引語間の距離を定義することは有効であると考えられる．しかし，表 2 を見ると，索引語の関連性を考慮することによりコサイン類似度では正しく分類できたものが分類できなくなってきていることが分かった．この理由としては，本論文では索引語間のすべての共起を考慮しているため，過剰に索引語同士の相関性を距離 **D** に反映させたためだと考えられる．したがって，相関性を考える索引語の範囲を限定するなどの距離の定義方法についてさらなる検討が必要である．

4. おわりに

本論文では，Earth Mover's Distance を用いることで索引語間の相関性を考慮したテキスト分類手法を提案した．対象文書での索引語の共起情報をもとに索引語間の距離を定義する方法を提案した．そして，Reuters-21578 を用いて評価実験を行うことにより，従来のコサイン類似度を用いた場合に比べ分類性能が改善することを確認した．

今後は索引語間の距離の決定方法についてさらなる検討を行う予定である．そして，本手法は一般的な文書類似度を求める手法であるため，テキストマイニングや情報検索分野へ応用し，手法の有効性を検討するつもりである．

参考文献

[岸田 03] 岸田 和明:文書クラスタリングの技法：文献レビュー, Library and Information Science, No.49, pp.33–75, (2003).

[竹内 07] 竹内 謹次, 黄瀬　浩一:Earth Mover's Distance に基づく Text-Based Image Retrieval, 情報処理学会研究報告, NL-177-5, pp.33–38, (2007).

[Porter 80] Poter, M: An Algorithm for Suffix Stripping, Program, 14(3), pp.130–137, (1980).

[Reuters] Reuters-21578 text categoriztion collection: http://kddi.cs.uci.edu/databases/reuters21578/reuters21578.html

[Rubner 00] Rubner, Y, Tomasi, C, and Guibas, I: The Earth Mover's Distance as a Metric for Image Retrieval, International Journal of Computer Vision, 14(3), pp.130–137, (2000).

[Salton 75] Salton, G, Wong, A, and Yang, C. S.: A Vector Space Model for Automatic Indexing, Communications of the ACM, 18(11), pp.613–620, (1975).

[Setabstiani 02] Sebastiani, F: Machine Learning in Automated Text Categorization, ACM Computing Surveys, 34(1), pp.1–47, (2002).

[SMART] SMART stop-list: ftp://ftp.cs.cornell.edu/pub/smart/english.stop

[徳永 99] 情報検索と言語処理, 東京大学出版会, (1999).

[Wan 05] Wan, X and Peng, Y: The Earth Mover's Distance as a Semantic Measure for Document Similarity, Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp.301–302, (2005).

[Yang 94] Yang, Y and Chute, C. G. : An Example-based Mapping Method for Text Categorization and Retrieval, ACM Transactions on Information Systems, 12(3), pp.252–277, (1994).