社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (8)
2006－ＮＬ－175 (8)
2006／9／12

圧縮性に注目した文書の関係分析手法

松崎 大輔　渡辺 俊典　古賀 久志　張 諾

電気通信大学 大学院 情報システム学研究科
〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: [matsuzaki,watanabe,koga,zhang]@sd.is.uec.ac.jp

あらまし　文書間の関係を分析する手法として，辞書ベースの形態素解析を用いて，語の出現頻度による類似性や，キーワード抽出を用いる方法が幅広く利用されている．これらの伝統的手法は，日々新しい単語が生まれるインターネットなどの環境には万全とはいえない．その理由は，これらの伝統的解析手法の前提となる，辞書に登録されていない未知語が頻繁に出現するためである．本稿では文書の圧縮率に注目し，人手による解析辞書の事前整備が不要な，文書の関係分析手法を提案する．提案手法について実験を行いその有効性を検討する．
キーワード　文書分析，クラスタリング，データ圧縮

Document Relation Analysis by Data Compression

Daisuke MATSUZAKI, Toshinori WATANABE, Hisashi KOGA, and Nuo ZHANG

Graduate School of Information Systems, University of Electro-Communications
1–5–1,Chouhugaoka, Chouhu City, Tokyo, 182–8585 Japan
E-mail: [matsuzaki,watanabe,koga,zhang]@sd.is.uec.ac.jp

Abstract　Dictionary-based morphological analysis is one of the main techniques for document analysis. It is usually used for keyword extraction and classification of similar words. Dictionary-based methods are weak for such environment as the Internet where new words appear that are not contained in the dictionary. In this study, we propose a new document relation analysis method based on the document's compressibility, requiring no dictionary. The effectiveness of our method is examined through some experiments.
Key words　Document analysis, Clustering, Data compression

1　はじめに

膨大な数の文書を扱うとき，それぞれの内容を示すトピックが明示され，適切に整理されていれば便利に利用することができる．実際に図書館などでは，欲しい本が容易に見つかるように，標準的な分類体系に基づき整理されている．ただし，各文書に何か１つのトピックを適切に決めるのは難しい．それは，文書には著者達の多様な思惑や意図で内容が記載されており，１つの文書が同時に複数のトピックを持ち得るからである（多重トピック）．

多数のトピックを想定したとき，各文書が持つトピックの重なりは多様で膨大になる．実際に図書館では，多重トピックごとに分けて物理的な棚をつくることは難しい．しかし，インターネットなら，ハイパーリンクを駆使すれば，多重トピックによる Web ページの分類を簡単に実現することができる．多くのポータルサイトでは，実際にディレクトリ型サービスとして，多重トピックに基づく Web ページの分類・整理を人手で行っている．しかし，日々増大する Web ページを１つひとつ人間が読んで判断し，適切な多重トピックに分類することは膨大な作業になる．したがって，多重トピックの分類を自動的にできれば大変便利になる [1].

Web ページなどの文書に対し，その関係を分析する手法として，大規模な辞書を用いた形態素解析が広く利用

されている [2]．形態素解析を行うことで，キーワードを抽出したり，語の頻度から文書間の類似性の判定を行うことができる．大規模文書を扱う情報検索，テキストマイニング，クラスタリングに有効な手段であるが，辞書を用いる手法は，日々新しい単語が生まれるインターネットなどの情報処理としては問題がある．それは，辞書に登録されていない全くの未知の用語を処理できないことや，複合語の抽出単位に曖昧性があることなどが挙げられる．

一方，筆者らはこれまでに，音声・画像・文書といったマルチメディアデータを形態素解析を用いずに，統一的に処理することができる方式の研究を行ってきた．これまでの成果としてクラスタリング，類似検索，内容推定などが挙げられる．これらは全て PRDC (Pattern Representation Scheme Using Data Compression) [3] というフレームワークで実現されており，データの特徴を圧縮率ベクトルという形で表現することで，形態素解析等を不要化している．

本研究では，形態素解析などの自然言語処理を行わず，文書の圧縮率に注目することで，文書間のトピックの重なりの関係を分析する方法を提案する．また，提案手法の特性を検討するための，仮想文書の作成，及び，これらや，実文書を用いた実験について検証する．

2 原理

ここで提案手法の基礎となるフレームワーク，PRDC について説明する．PRDC は，データの特徴を圧縮率ベクトルという形で表現する．ベクトル間の距離を求めることで，データのクラスタリングや類似検索が行える．一般のデータ圧縮において，入力記号列に対してどのように符号化するかは，それをどのようにモデル化するかによって決まる．モデル化の一種に辞書に基づく符号化（図 1）があり，PRDC ではこの符号化を用いる．そして，この符号化によって生成される辞書で空間を構成し，新たな入力データをその空間に写像していく．以下に簡単な例を示す．入力ファイル A を「aaaaabbabbaaabaa」とし，A を圧縮する際に辞書が表 1 のように生成されたとする．この場合の出力記号は「01010010001000010100101」となる．1 記号が ASCII コードの 7 ビットとすれば，圧縮率は 23bit/112(7 × 16)bit=0.205 となる．同様に，入力ファイル B を「abbbbabbaabbabbb」とし，B を圧縮する際の辞書が表 2 のように生成されたとする．この場合の出力列記号は「000101001010000101000101」となり，圧縮率は 24bit/112(7 × 16)bit=0.214 となる．

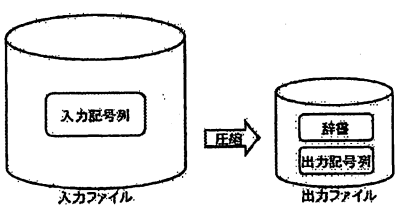

図 1: 辞書に基づく符号化

表 1: 入力記号列 A の辞書

入力記号	出力記号
aa	01
ab	001
ba	0001
bb	00001

表 2: 入力記号列 B の辞書

入力記号	出力記号
bb	01
ba	001
ab	0001
aa	00001

ここで，入力ファイル A に対して B で生成された辞書を用いて圧縮を行うと，出力記号列は「000010000100001001010000100010000 1」となり，圧縮率は 33bit/112bit=0.295 となる．同様に入力ファイル B に対して A で生成された辞書を用いて圧縮を行うと，「001000010001000010100001001000001」となり，圧縮率は 32bit/112bit=0.286 となる．つまり，圧縮に用いる辞書を変更することで圧縮率は変化する．さらに，別のファイル C,「aaabaaabbaabaaab」を辞書 A，辞書 B を用いて圧縮すると出力記号列はそれぞれ「010010100100010010100 1」，「000010000100001000100100010000100001」となり，圧縮率は 22bit/112bit=0.196，34bit/112bit=0.304 となる．以上で求めた圧縮率（表 3）を A，B の辞書を軸とする二次元座標上にプロットすると図 2 のようなグラフになる．座標上での距離を調べると，ファイル C はファイル B よりもファイル A の方が近い位置にあるため，C は A に対して B よりも類似性があるということがわかる．これがデータ圧縮を用いた文書の関係分析方法の原理となる．

表 3: 使用辞書による圧縮率の変化 [%]

	ファイル A	ファイル B	ファイル C
辞書 A	20.5	21.4	19.6
辞書 B	29.5	28.6	30.4

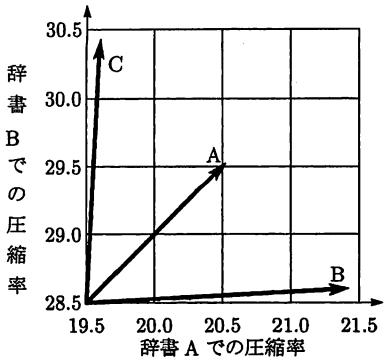

図 2: 圧縮率ベクトル

3 文書の関係分析

複数の文書を想定したとき，各文書が持ち得る共通のトピックの重なりは，以下のように表すことができると考えられる．例えば図 3 のように，文書 A が文書 B，C と斜線部のような共通のトピックを持っていたとき，文書 B，C は文書 A と部分的包含関係にある．また，文書 D が他の A，B，C と共通するトピックを持っていないとき，図 3 のように孤立し，いずれの文書とも関係がない．

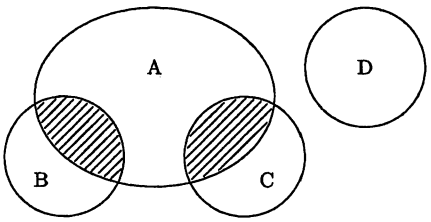

図 3: トピックの重なり

このような文書のトピックの重なりを圧縮率で分析する方法を考える．まず，対象となる複数の文書を入力とし，符号化を行いそれぞれについて辞書を得る．(今回は圧縮符号法として LZW 符号法を用いる.) 次に得られた辞書を利用して，入力文書すべてに対し再度符号化を行う．ここで，入力文書を N とし，その文書を符号化して得られる辞書を D_N とする．また，テキスト N を辞書 D_N で圧縮したときの圧縮率を C_{ND_N} とし，次式で定義する．L_N は N の文字数，K_N は辞書 D_N での符号化回数である．

$$C_{ND_N} = \frac{K_N}{L_N}$$

ここで圧縮率は，通常のデータ圧縮では考慮される辞書部のデータサイズは含めない．また，一般に LZW では，辞書に登録する単語数に上限を設け，上限を越えたら新たな単語登録は行わないが，ここでは文書全体を必ず圧縮させたいため，この制限は設けない．

以上の結果次の圧縮率表を得る（表 4）．この圧縮率表から文書の関係を分析する．例えば，図 3 において提案手法で分析を行うと，文書 A で生成された辞書 D_A は，共通のトピックを持つ文書 B，C をその分だけよく圧縮することができる．逆に，文書 B の辞書 D_B についても，包含する分だけ文書 A を圧縮することができる．しかし文書 C，D に対しては，共通トピックを持たないため，文書 A ほどには圧縮できない．文書 C についても同様なことがいえる．一方，文書 D においては，他と共通するトピックを持たないため，辞書 D_D での他の全ての文書の圧縮度は悪くなる．以上のようにして，文書のトピックの重なりの関係を分析することができる．

表 4: 圧縮率表

	N_1	N_2	N_3	$\cdots$	N_n
D_{N_1}	$C_{N_1D_{N_1}}$	$C_{N_2D_{N_1}}$	$C_{N_3D_{N_1}}$		$C_{N_nD_{N_1}}$
D_{N_2}	$C_{N_1D_{N_2}}$	$C_{N_2D_{N_2}}$	$C_{N_3D_{N_2}}$	$\cdots$	$C_{N_nD_{N_2}}$
D_{N_3}	$C_{N_1D_{N_3}}$	$C_{N_2D_{N_3}}$	$C_{N_3D_{N_3}}$		$C_{N_nD_{N_3}}$
$\vdots$		$\vdots$			$\vdots$
D_{N_n}	$C_{N_1D_{N_n}}$	$C_{N_2D_{N_n}}$	$C_{N_3D_{N_n}}$	$\cdots$	$C_{N_nD_{N_n}}$

4 仮想文書

実際の文書中のトピックは，読み手の解釈やその目的によって多種多様であり，トピックの完全に客観的な抽出は難しい．このような実際の文書をそのまま用いた，提案手法の関係分析能力を検証するのは容易でない．そこで，提案手法の検証を行いやすいように，人工的に文書を以下の方法で作成することにした．

ある文書中における主要な話題（トピック）は，特定のフレーズや単語（以下，基本フレーズ）が複数回繰り返され構成されていると考えられる．そこで，意図的にこの基本フレーズを作成し，またその出現回数を任意に変化させ，それらの基本フレーズの間を、基本フレーズに含まれることのない文字（ノイズ文字）で埋めることとした．このノイズ文字は，文書中に繰り返しが表れないように，充分多くの文字をあらかじめ割り当てることで実現する（図 4）．今回は，このノイズ文字を 7000 種集め，この文字集合からランダムに一文字ずつ抽出し，基本フレーズの間を埋めていった．

このようにして，基本フレーズの種類や出現頻度を変化させることで，自由にトピックを形成させた仮想文書を作成することができる．基本フレーズの種類と出現頻度を調整することで，意図的に文書間でトピックが重なり合うような関係を作り出すことができる．これらの仮想文書に対し提案手法を適用することで，その特性を把握することができる．また，実際の文書間の関係を分析する足がかりとなる．

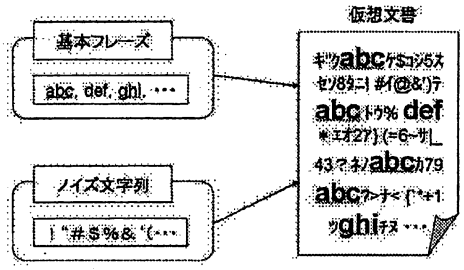

図 4: 仮想文書

5 実験

文書中のトピックは偏在的に集中している場合や，離散的に存在している場合が考えられる．そこで本節では，文書中にトピックが順序だてて語られている場合と，同時に複数のトピックが語られている場合を想定した．実験 1 では仮想文書について，実験 2 では実際の文書について実験を行った．また，実験 2-2 においては，文書間の類似性について，形態素解析を用いた方法との比較を行った．

5.1 実験 1

ここでは仮想文書を用いて，文書中にトピックが順序だてて語られている場合（実験 1-1），また，同時に複数のトピックが語られている場合（実験 1-2）を想定し実験を行った．

5.1.1 実験 1-1

この実験では，基本フレーズを「abc」，「def」，「ghi」とし，ノイズ文字数を 1000 文字に固定し，表 5 のように「abc」，「def」，「ghi」に対応したトピックを 1 つだけ持つ文書 T1，T2，T3 を作成した．次に，得られた文書 T1 と T2 を結合させ，1 つの文書 A を作成する．つまり，1 つの文書が順序だてて 2 つのトピックを持っている場合を人工的に再現したものである．また文書 A の他に B，C，D を用意した．その構成は以下に示した通りである．文書 B，C は A に包含されており，D は他とは異なったトピックを持っている．これらの文書 A，B，C，D についての実験結果を表 6 に示す．

- 仮想文書 A：T1・T2
- 仮想文書 B：T1
- 仮想文書 C：T2
- 仮想文書 D：T3

表 5: 仮想文書

Text ID	構成
T1	「abc」：50 回出現
T2	「def」：50 回出現
T3	「ghi」：50 回出現

表 6: 実験 1-1 での圧縮率　[%]

	A	B	C	D
D_A	48.35	48.43	48.26	100.00
D_B	74.22	48.43	100.00	100.00
D_C	74.13	100.00	48.26	100.00
D_D	100.00	100.00	100.00	47.83

結果の表 6 について考察する．第 1 行は，文書データ，第 1 列がその文書から生成された辞書であり，表の要素は辞書で文書を圧縮した時の圧縮率 [%] を表す．まず表中の文書 A の列に注目する．C_{AD_B} (74.22 [%]) と C_{AD_D} (100.00 [%]) の圧縮率を比較すると，D_B の方が文書 A をよく圧縮している．C_{AD_C} (74.13 [%]) と C_{AD_D} (100.00 [%]) に関しても D_C の方が文書 A をよく圧縮している．つまり，文書 A は文書 B，C と深い関係を持つことが明らかにされている．また C_{AD_B} (74.22 [%]) と C_{AD_C} (74.13 [%]) の圧縮率がほぼ等しいことから，文書 A は文書 B，C と同程度の関係を持つとみれる．一方，圧縮率が 100 [%] と高くなっている C_{AD_D} については，文書 A は文書 D と無関係であるとみれる．他の列 B，C，D でも同様に検証していくことで，文書 A は文書 B,C と強い関係にあり，文書 D は他と関係がないことがわかる．以上のように，意図的に生成した仮想文書については，圧縮率表からその関係を読み取ることが示された．

5.1.2 実験 1-2

同じく仮想文書を使って，基本フレーズを実験 1-1 と同様に「abc」,「def」,「ghi」とし，ノイズ文字数を 1000 文字に固定した．但し前実験と異なり，文書を結合させることなく，基本フレーズを混在させて仮想文書を生成した．これは，ある 1 つの文書が同時に 2 つのトピックについて語られている場合を再現したモデルである．実験に用いた文書は以下の通りである．実験 1-1 と同様に文書 B，C は A に部分的に包含されており，D は他とは異なったトピックを持っている．結果を表 7 に示す．

- 仮想文書 A：「abc」,「def」それぞれ 50 回出現
- 仮想文書 B：「abc」50 回出現
- 仮想文書 C：「def」50 回出現
- 仮想文書 D：「ghi」50 回出現

表 7: 実験 1-2 での圧縮率　[%]

	A	B	C	D
D_A	45.92	91.22	91.30	100.00
D_B	92.23	47.48	100.00	100.00
D_C	92.23	100.00	47.65	100.00
D_D	100.00	100.00	100.00	48.26

実験 1-1 の結果に比べ表 7 は，圧縮率の値は変化したものの，同じ傾向を示している．つまり，文書 A は文書 B,C と関係があり，文書 D は他と全く関係ないことがわかる．よって，1 つの文書で同時に 2 つのトピックが並行的に語られている場合においても，圧縮率表からその関係を読み取ることが示された．

5.2 実験 2

ここでは，実際の文書を用いて実験を行った．データは Web ページ上から取得した．実験 2-1 は実験 1-1 に対応した，トピックが順序だてて語られている場合について，実験 2-2 では実験 1-2 に対応する，トピックが混在して同時並行に語られている場合についてである．

5.2.1 実験 2-1

表 8 のように 3 つのトピックを用意し，次のように構成した．文書 T1 と T2 を結合させ，文書 A を作成する．また文書 B，C，D はそれぞれ，T1，T2，T3 をそのまま用いた．結果を表 9 に示す．

- 実文書 A：T1・T2
- 実文書 B：T1
- 実文書 C：T2
- 実文書 D：T3

表 8: トピック

Topic ID	概要
T1	平成 18 年 7 月豪雨
T2	レバノン空爆
T3	2006W 杯イタリア優勝

表 9: 実験 2-1 での圧縮率　[%]

	A	B	C	D
D_A	44.95	44.17	45.80	88.95
D_B	67.73	44.61	92.87	94.17
D_C	68.56	88.48	46.91	91.66
D_D	88.84	88.48	89.06	42.99

表 9 より，それぞれの圧縮率の値は，仮想的に作成した実験 1-1 と多少差はあるものの，同様の傾向が確認できる．つまり，文書 A は文書 B,C と関係があり，文書 D は他の文書との関係が弱いと判断される．以上より，実際の文書においても，その関係が圧縮率に表れることが確認できた．

5.2.2 実験 2-2

ここでは実験 1-2 に対応する，トピックが混在して同時に語られている場合について，実際の文書を用いて実験を行った．文書の概要を以下に示す．A はスマトラ沖地震についての報告，B は津波についての論述，C はインドネシアの観光案内，D は 2006 年サッカー W 杯でイタリアが優勝した時についての記事である．文書 A は，「インドネシアで」,「津波が起きた」という少なくとも 2 つのトピックが存在している．一方，文書 B は「津波について」，文書 C は「インドネシア」について語られており，それぞれ文書 A と共通のトピックを持つ，部分的包含関係にある．また文書 D は A，B，C とは異なった，「サッカー」に関するトピックを持っている．結果を表 10 に示す．

- 実文書 A：スマトラ沖地震
- 実文書 B：津波について
- 実文書 C：インドネシア観光情報
- 実文書 D：サッカー W 杯イタリア優勝

表 10: 実験 2-2 での圧縮率　[%]

	A	B	C	D
D_A	41.34	75.23	76.38	79.95
D_B	81.44	43.69	79.42	83.04
D_C	75.15	75.11	39.56	72.16
D_D	96.10	92.19	89.59	43.14

表 10 について考察する．まず表 10 中の文書 A の列だが，C_{AD_D} (96.10 [%]) の圧縮率が突出している．つまり，文書 A は文書 D とは異なるトピックで構成されていると考えられる．文書 B 列に関しても，C_{BD_D} (92.19 [%]) での圧縮率が高い，また，興味深いことに C_{BD_A} (75.23 [%])，C_{BD_C} (75.11 [%]) での圧縮率が同程度となった．これは，文書 B の津波についての論述を読んでみると，事例としてスマトラ沖地震についての言及があった．このため，文書 A についてはもちろんのこと，それら地域の名称が，文書 C のインドネシア観光情報の内容と重なったためだと推測できる．また文書 C 列についても同様なことがいえる．最後に文書 D 列だが，C_{DD_A} (79.95 [%]) は C_{BD_A} (75.23 [%])，C_{CD_A} (76.38 [%]) より少し大きく，C_{DD_B} (83.04 [%]) は，C_{AD_B} (81.44 [%])，C_{CD_B} (79.42 [%]) より少し大きい，よって，文書 A，B とは関係が弱い．但し，C_{DD_C} (72.16 [%]) は逆転している．これは文書間に共通した助詞や，助動詞などの影響だと予想しているが，定かではなく，今後検討する必要がある．しかしの A，B，C 列の最終行，C_{AD_D} (96.10 [%])，C_{BD_D} (92.19 [%])，C_{CD_D} (89.59 [%]) が大きいため，文書 D は他と異なる内容であることが推測できる．

5.2.3 形態素解析法との比較

提案方式との比較のため，実験 2-2 の文書を用いて，文書間の類似性を形態素解析によって求めた．提案手法では，表 10 の列をベクトルとみなし，それぞれのなす角 θ を用いて $(1-cos\theta)$ を算出した (表 11)．形態素解析を用いた方は，文書中の自立語を抽出し，その出現回数を頻度ベクトルとしたコサイン類似度を用いて，(1 − コサイン類似度) と算出した (表 12)．双方とも文書間の非類似度を表し，値が小さいほど類似している．表 11 と表 12 の値の差が大きいのは，算出時のベクトルの要素値の値域の違い (圧縮率 (0〜1) と自立語の出現回数 (0〜100)) によるものである．表 11，12 より，文書間の類似性は同じ傾向を示しており，提案手法によって，伝統的な形態素解析を用いる手法と同様の結果を導ける可能性がうかがえる．

表 11: 提案手法 文書間非類似度

A				
B	0.056			
C	0.055	0.057		
D	0.098	0.094	0.077	
	A	B	C	D

表 12: 形態素解析 文書間非類似度

A				
B	0.62			
C	0.53	0.57		
D	0.82	0.81	0.75	
	A	B	C	D

6 まとめ

本論文では，データの圧縮率に着目して，文書間の関係を分析する方法を提案した．トピックの重なりの関係が圧縮率表に表れることを示した．提案手法の検証のために，意図的に文書間の関係を構成できる仮想文書を導入した．提案手法の特徴は，形態素解析などの言語処理を全く行わないため，大規模な辞書を作成することなく，あらゆる言語に適用できる点である．

謝辞　本研究を進めるにあたり，有益なご指導，ご助言を頂いた，電気通信大学大学院の北村浩客員助教授に深く感謝致します．

参考文献

[1] 斉藤和己，“パラメトリック混合モデル (PMM) による多重トピック抽出，” NTT 技術ジャーナル 2004.6, pp.10-13, 2004.

[2] 馬場肇，“Namazu システムの構築と活用，” ソフトバンクパブリッシング株式会社，2003.

[3] Toshinori Watanabe, Ken Sugawara, and Hiroshi Sugihara, “A New Pattern Representation Scheme Using Data Compression,” IEEE TransPAMI, Vol. 24, No. 5, May. 2002.

[4] 植松友彦，“文書データ圧縮アルゴリズム入門，” CQ 出版社，1994.