# パーソナライゼーションにおけるトピックを意識しない類似度測定

## Unconscious Similarity of Topics in Personalization

長谷川新[*1]
Shin Hasegawa

相澤彰子[*2]
Akiko Aizawa

浜本隆之[*1]
Takayuki Hamamoto

[*1] 東京理科大学
Tokyo University of Science

[*2] 国立情報学研究所
National Institute of Informatics

In personalization, it's necessary to calculate similarity between user's profile and the presented information. But user's profile contains various format and many topics, so we propose unconscious similarity of topics using compression for text.

## 1. はじめに

パーソナライゼーション[1]において、ユーザが求めている情報を判断するには，趣味嗜好を表した情報(プロファイル)との類似度を計算する必要がある．しかし，ユーザの情報は様々な形式やトピックを含むため，文書構造の解析やトピックの分析が必要となり，類似度の計算は容易ではない．そこで，テキスト情報を対象に，近年提案されている NCD[2]のような圧縮に基づく情報間類似度を応用することで，トピックを意識せずに類似度を測る手法を提案し、評価する。

## 2. 圧縮に基づく類似度

本稿では、ユーザの趣味嗜好を表すプロファイルをユーザが好む既知の適合文書の集合 D として表し，この適合文書集合と新たな文書 x の類似度の計算に圧縮を応用する．

A と B の類似性が高いということは，互いに共通の情報を多く含むということである．そのため圧縮の際に，A の情報を用いて B を圧縮した際の圧縮率は高くなり，得られた圧縮率を互いの類似度として表すことが可能である．これを利用すれば，適合文書集合 D の情報を用いて文書 x を圧縮した際の圧縮率を，D と x の類似度として用いることができる．

以下に，本稿で用いる PRDC と呼ばれる類似度尺度について説明し，さらに改良型 PRDC を提案する．

### 2.1 Pattern Representation Scheme using Data Compression（PRDC）

PRDC[3]は，図 1 に示すように，適合文書集合 D に含まれる適合文書 $y_i$ から LZ78 の圧縮により最長一致の部分文字列を集めた辞書 $d_i$ を生成する．生成された辞書 $d_i$ を，文書 x の圧縮に再び用いることで得られる出力長 PRDC(x, $d_i$)を，原文長 L(x)で正規化した圧縮率を各適合文書との類似度としている．このことから，適合文書集合 D に含まれる適合文書 $y_i$と文書 x との類似度 sim(x, $y_i$)は次式のように与えられる．

$$\mathrm{sim}(x, y_i \epsilon D) = \frac{\mathrm{PRDC}(x, d_i)}{L(x)}$$

上記の PRDC の定義に基づき，適合文書集合 D と文書 x の類似度 sim(x, D)を，D に含まれる最も x と類似性の高い(圧縮率の高い)適合文書 $y_i$との類似度を用い，以下のように与える．

$$\mathrm{sim}(x, D) = \mathrm{Minmum}\{\mathrm{sim}(x, y_i \epsilon D)\}$$

連絡先：長谷川新，東京理科大学大学院　工学研究科　電気工学専攻，東京都千代田区九段北 1-14-6，03-3260-4272(内線 6561)，hasegawa@nii.ac.jp

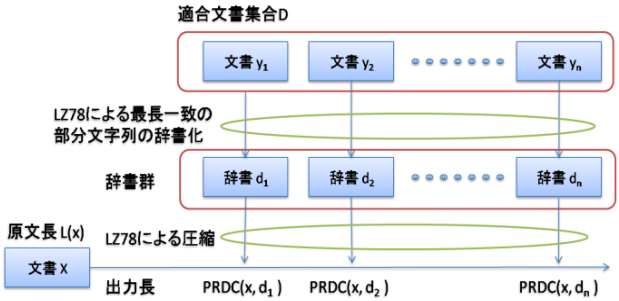


図 1　PRDC の概要

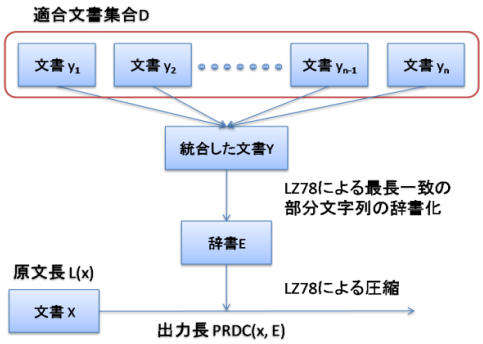


図 2　改良型 PRDC の概要

以上の PRDC は，トピックを意識せずに，適合文書集合に含まれる各適合文書の部分文字列を圧縮により辞書化することで，各文書の特徴を捉えることはできるが，適合文書集合の特徴を捉えることはできない．そのため，適合文書集合の部分文字列を辞書化する改良型 PRDC を提案する．

### 2.2 改良型 PRDC

改良型 PRDC は，図 2 に示すように，適合文書集合 D の文書を全て統合した文書 Y を生成した上で，LZ78 の圧縮により辞書 E を生成する．生成された辞書 E を，文書 x の圧縮に再び用いることで得られる出力長 PRDC(x, E)を，原文長 L(x)で正規化した圧縮率を適合文書集合 D との類似度とする。このことから、適合文書集合 D と文書 x との類似度 sim(x, D)は以下のように与えられる．

$$\mathrm{sim}(x, D) = \frac{\mathrm{PRDC}(x, E)}{L(x)}$$

以上の改良型 PRDC は，適合文書集合に含まれる文書を全て統合しているため，様々なトピックを一つに集約している．これにより，トピックを意識せず一度の圧縮により，類似度を測ることが可能である．さらに，LZ78 により適合文書集合に含まれる

文字列を辞書化しているため，同じ文字列が繰り返し現れるほど，最長一致の部分文字列が辞書に登録される．このことは，任意のトピックに高頻度で出現する文字列を考慮することができることを示しており，同じトピックの文書に対して高い圧縮率を発揮することが期待できる．

また，改良型 PRDC における辞書 E は，適合文書集合に新たな文書が追加されても，再構築する必要がない．既存の辞書 E を用いて新たな文書を圧縮することで，部分文字列を辞書化し，更新することが可能である．

## 3. 類似度の有効性検証

圧縮に基づく類似度の有効性について検証するため，トピックが混ざった適合文書集合を用いて興味のある文書を取得できるかを検証した．さらに，PRDC と改良型 PRDC の比較を行うだけでなく，比較手法に対してどの程度の性能かを検証した．

比較手法として，TF-IDF による重み付けをされた文書ベクトルのコサイン類似度を用いる．具体的には，改良型 PRDC と同様に，適合文書集合 D の全文書を統合した文書 Y を文書ベクトル化し，文書 x の文書ベクトルとのコサイン類似度を計算する．この際，文書ベクトルには，形態素解析器により切り出された名詞のみを用い TF-IDF による重み付けをした．

実験データとして，被験者 10 人の学生が自分の好きなトピックのニュース記事を 20 件収集した．条件としては，記事のサイズは任意で「政治」「経済」「スポーツ」「サイエンス」「芸能」のカテゴリ内で時間的に近い記事とした．

### 3.1 評価方法

10 個のトピックから興味のあるトピックを複数選択し，さらに選択したトピックに含まれる記事を 10 個ランダムに取得した．この時，取得されなかった同じトピックの残りの文書は，未読の適合文書と仮定される．取得した文書集合を適合文書集合 D として，取得した以外の文書 x との類似度を計算し，類似度順にランキングした．このランキングの評価として以下の F 値を用いた．

$$\text{F 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

$$\text{適合率} = \frac{\text{順位 r までに表れた適合文書の数}}{\text{順位 r}}$$

$$\text{再現率} = \frac{\text{順位 r までに表れた適合文書の数}}{\text{全適合文書の数}}$$

順位 r を変化させた際の F 値の最大値を類似度の性能とし，選択したトピックについて 10 回実行し平均化した．この値が高い程，適合文書集合 D に合った適合文書が取得できたことになる．

### 3.2 実験結果

図 3 は，選択したトピック数を 1 から 4 とした場合の PRDC と改良型 PRDC の F 値をプロットしたものである．この図において，中央の赤線よりも改良型 PRDC 側に多くのプロットがあることから，改良型 PRDC のほうが優れていることがわかる。このことから，適合文書集合の部分文字列を辞書化したことで，性能が向上したことが分かる．

図 4 は，選択したトピック数を 1 から 4 とした場合の改良型 PRDC と比較手法の F 値をプロットしたものである．比較手法にプロットの数が多く，選択した全トピックを通した平均の F 値についても，比較手法が 0.918，改良型 PRDC が 0.913 と僅かではあるが劣っている．これは，改良型 PRDC における部分文字列がトピックの特徴を TF-IDF ほど捉えられていないことが考えられる．

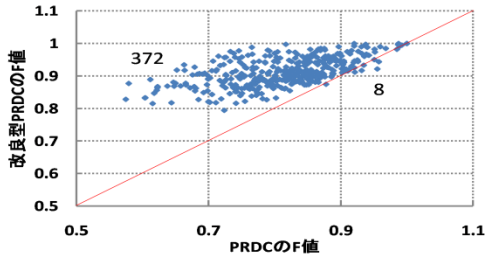


図 3 トピックの数を 1 から 4 個選んだ際の PRDC と改良型 PRDC の性能

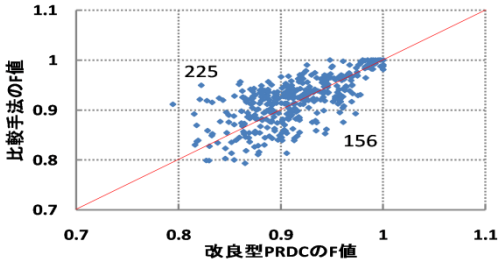


図 4 トピックの数を 1 から 4 個選んだ際の改良型 PRDC と比較手法の性能

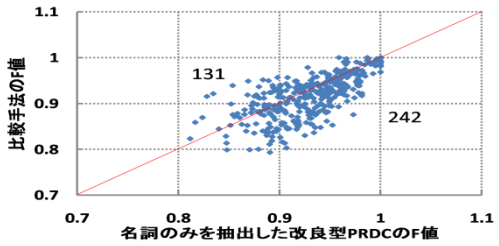


図 5 トピックの数を 1 から 4 個選んだ際の名詞のみを抽出した改良型 PRDC と比較手法の性能

そのため，文献[4]でも行われている前処理による精度向上を検討した．この手法として，改良型 PRDC において統合文書を作成する前に，名詞のみを抽出した場合の結果を図 5 に示した．比較手法よりも多くのプロットがあり，選択した全トピックを通した平均の F 値も 0.931 と性能が向上した．

## 4. むすび

本稿では，改良型 PRDC という圧縮に基づく類似度として，トピックを意識せずに一度の類似度計算を可能にする手法を提案した．また，実験では比較手法に及ばなかったが，平均して 0.913 という高い F 値を示した．また，名詞を切り出す前処理を行ったところ比較手法を上回る性能を示した．今後は，この前処理による効果について詳細に検討する．

## 参考文献

[1] Mobasher B. : Data Mining for Web Personalization, Lecture Notes in Computer Science, Vol.4321, pages 90-136 (2007).

[2] Cilibrasi R., Vitanyi P. M. B. : Clustering by compression, IEEE Transactions on Information Theory, Vol.51, No.4, pages 1523-1545 (2005).

[3] 木村洋章, 渡辺俊典, 古賀久志, 張諾: LZ78 の圧縮性を利用した文書検索手法の提案，情報処理学会研究報告，Vol.2006, No.94, pages 65-70 (2006).

[4] Helmer S. : Measuring the structural similarity of semistructured documents using entropy, In Proceedings of the 33rd international Conference on Very Large Data Bases, pages 1022-1032 (2007).