Vol.2009-NL-194 No.8
2009/11/17

視線情報を用いたユーザプロファイル獲得と文書推薦

長谷川新[†]　相澤彰子[††]　浜本隆之[†]

我々は，ユーザが興味のある文書の集合をプロファイルとして獲得し，このプロファイルとの類似度計算に基づく新たな文書を推薦するシステムを検討している．ここで，一つの文書が複数のトピックを含む場合には，ユーザは文書全体に興味があるとは限らない．ユーザの興味をより細かく反映するために，本稿では，(1) 視線情報を用いることで，文書から興味のある部分のみを抽出し、さらに (2) プロファイルとの類似度計算に，圧縮アルゴリズムに基づく手法を導入することで，対象文書の一部だけに興味がある場合でも有効な文書推薦を提案する．また，新聞記事集合を用いた実験により，圧縮アルゴリズムによる類似度の有効性を示すとともに，予備実験を通して文書推薦において視線情報を用いる効果について考察する．

User Profile by Gaze Information and Document Recommendation

Shin Hasegawa[†]　and Akiko Aizawa[††]
and Takayuki Hamamoto[†]

In this paper, we examine a document recommendation system which is based on a ‘profile’ of a user defined as a collection of documents of the user’s interest. The user profile is used for similarity calculation between incoming documents and the user’s interests. Since documents often have multiple topics, users may not necessarily like the entire document. In this paper, we use gaze information to identify segments of user’s interests. We also propose a similarity measure for document recommendation using a compression-based algorithm. In our experiments, we show the effectiveness of our similarity measure and also, investigate the advantage and issues of gaze information-based profiling for document recommendation.

1. はじめに

我々は，ユーザが興味のある文書の集合をプロファイルとして獲得し，このプロファイルと新たな文書との類似度計算により内容の近い文書を推薦するシステムを検討している．

適合文書を獲得する手法として，メールやウェブの閲覧・作成履歴[1]，ユーザのデスクトップの文書群[2]を利用する手法などがある．文書が複数のトピックを含む場合には，これらの手法を用いて文書全体をユーザの適合文書として扱うことは難しい．例えば，ウェブニュースの記事のページのように記事以外の複数のトピックが含まれており，ユーザはページ全体に興味があるとは限らない．そのため，文書においてユーザの興味がある部分を捉え，適合文書を抽出する必要がある．

本稿では，視線情報を用いて文書から興味のある部分を抽出した文書の集合をプロファイルとして獲得する．さらに，プロファイルとの類似度計算に，近年提案されている NCD[3]のような圧縮アルゴリズムに基づく手法を導入することで，対象文書の一部だけに興味がある場合でも有効な文書推薦を提案する．また，新聞記事集合を用いた実験により，圧縮アルゴリズムによる類似度の有効性を示すとともに，予備実験を通して文書推薦において視線情報を用いる効果について考察する．

2. 視線情報を用いたユーザプロファイル獲得

視線情報により文書中の注視領域からユーザのプロファイルを獲得する手法を提案する．視線情報を文書の注釈に応用した研究[4]では，ユーザが読んだ領域を特定している．本稿では，読んだ領域を特定するだけでなく，読んだ領域をプロファイルとして獲得する. これにより, 文書全体をユーザのプロファイルとして扱うのではなく，興味のある領域を限定した文書の一部をユーザのプロファイルとして扱う．

実際には，ディスプレイにおけるユーザの注視点の座標を取得し，この座標の集合を注視領域として，あるタイミングによりディスプレイの画像を切り抜く．さらに，画像から文字情報を抽出する OCR により切り抜いた画像から，文書情報を取得しプロファイルとして扱う．以下に，その実装と一連の流れを説明する．

[†] 東京理科大学
Tokyo University of Science
[††] 国立情報学研究所
National Institute of Informatics

Vol.2009-NL-194 No.8
2009/11/17

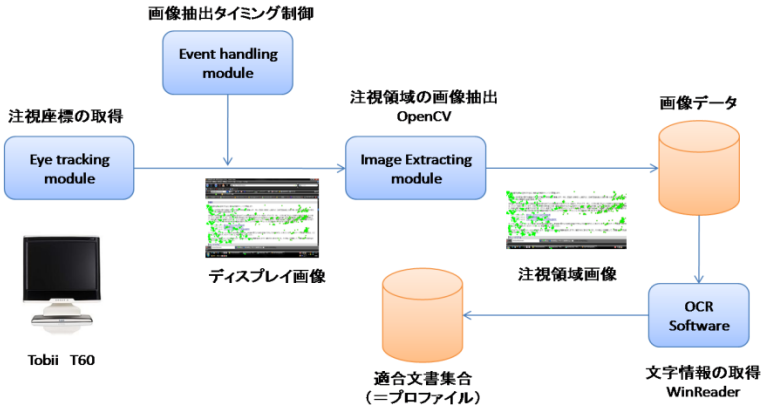


図 1　視線情報を用いたユーザプロファイル獲得までの流れ

2.1 システム構成

本稿における視線情報を用いたユーザプロファイル獲得のためのシステムは，図 1 に示すような構成となっている．それぞれのモジュールの役割と機能について説明する．

(1) Eye Tracking module

このモジュールでは視線検出装置を用いて，ディスプレイにおけるユーザの注視点の座標を取得し，蓄積する．実装には，Tobii 社の T60[a]という視線検出装置及び Tobii 社の SDK を利用している．

(2) Event Handling module

このモジュールでは，マウスやキーボードなどのイベントを把握し，ユーザの操作を監視する．これにより，Eye Tracking module で取得した注視点の座標集合を用いてディスプレイ画像から注視領域を切り出すタイミングを制御する．

ユーザの画面遷移や，スクロールといったイベントにより画面が切り替わってしまうため，注視領域の画像を切り出す際には，画面が切り替わる前に切り出す必要がある．そのため，本稿では，マウスのクリックやスクロールといったイベントを監視することで対応している．また，流し読みを判断するために，連続的なスクロールの際には蓄積した注視点の座標を破棄させる機能も実装している．

(3) Image Extracting module

このモジュールでは，蓄積された注視点の座標集合から，座標集合全体を囲む矩形を求め，ディスプレイ画像から注視領域の画像を切り抜く． 実装には，Intel 社の OpenCV ライブラリ[b]を利用し実装している．

Eye Tracking module で蓄積される視線情報は 60Hz の間隔で蓄積されるため，視線を逸らした際には，注視していた領域とは関係のない外れた点を含む可能性がある．このため，注視点の座標集合から，連続する 5 点毎の注視点について座標の平均を取り，座標の移動量を抑えることでよそ見を抑制している．

(4) OCR Software

このモジュールでは，OCR ソフトを用いて注視領域の画像から文字情報を抽出する．実装には，メディアドライブ社の WinReader Pro[c]というソフトを用いている．このソフトには，フォルダ監視機能が付いているため，Image Extracting module より切り出された注視領域の画像を指定のフォルダに置くことで，文字情報を自動的に抽出することが可能である．

以上の一連のモジュールの処理を経て，文書中のユーザが注目した文字情報を抽出しプロファイルを獲得することができる，また，注視領域画像の切り出しから OCR ソフトを用いた文字認識まで自動的に処理が行われるため，開いたファイルの形式によらずプロファイルを取得することが可能である．

3. 文書推薦のための圧縮アルゴリズムによる類似度

文書推薦では，獲得したプロファイルに基づき新たな文書に対するユーザの興味の度合いを計算するが，ここで，文書全体ではなく一部だけがユーザの興味の対象となる場合も考えられる．文書の構造解析やトピックの分析などの手法によらず，近似的に部分的な一致を計算する方法として，本稿では圧縮アルゴリズムによる類似度に注目し，実験によりその適用可能性を調べる．

圧縮アルゴリズムによる類似度は，遺伝子情報や音楽，文学の解析に利用されている[3]．この類似度において，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る．このことから，得られた圧縮率を互いの類似度として用いることが可能である．そこで本稿では，PRDC[5]と呼ばれる LZ78 の圧縮アルゴリズムを利用した類似度計算の手法に着目し，この手法を元にした改良型 PRDC を提案する．

以降では，ユーザのプロファイルを適合文書集合 G，新たな文書 x との類似度を sim（x, G）として表す．

a) http://www.tobii.co.jp

b) http://opencv.jp
c) http://mediadrive.jp

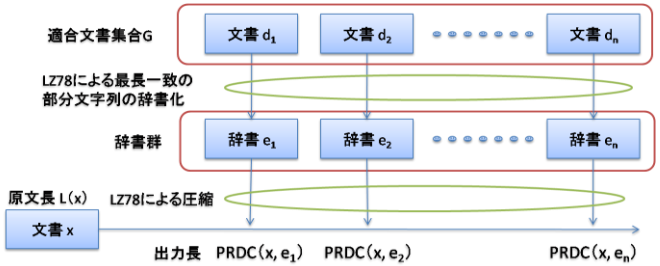


図 2 PRDC の概要

3.1 Pattern Representation Scheme using Data Compression

PRDC は，図 2 に示すように，適合文書集合 G に含まれる文書 d_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 e_i を生成する．さらに，辞書 e_i を利用した圧縮により得られる文書 x の出力長 PRDC（x, e_i）を原文長 L（x）で正規化した圧縮率を各文書との類似度としている．各文書との類似度は以下のように与えられる．

$$\mathrm{sim}(x, d_i) = \frac{\mathrm{PRDC}(x, e_i)}{L(x)}$$

この手法を文書推薦に利用する際, 文書 x と適合文書集合との類似度 sim(x, G)は，適合文書集合 G から生成された辞書毎に圧縮率を求めた上で与える必要がある．しかし，適合文書が増加するほど，辞書も増加し、一つ一つ辞書を読み込んで圧縮することは非効率である．また，LZ78 により最長一致の部分文字列を集めることから適合文書毎の特徴を捉えることは可能だが，適合文書集合の特徴を捉えることができない。よって，適合文書集合全体の部分文字列を辞書化する改良型 PRDC を提案する．

3.2 改良型 PRDC

改良型 PRDC は，図 3 に示すように，適合文書集合 G に含まれる文書 d_i を全て統合した文書 D を生成する．この統合文書 D に対して LZ78 の圧縮により最長一致の部分文字列を集めた辞書 E を生成する．さらに，辞書 E を利用した圧縮により得られる文書 x の出力長 PRDC（x, E）を原文長 L（x）で正規化した圧縮率を統合文書 D との類似度として与えられる．統合文書 D は，そのまま適合文書集合 G を表すため文書 x と適合文書集合との類似度 sim（x, G）は，以下のように与える．

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

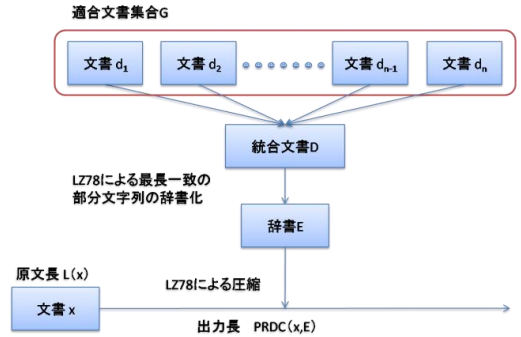


図 3 改良型 PRDC の概要

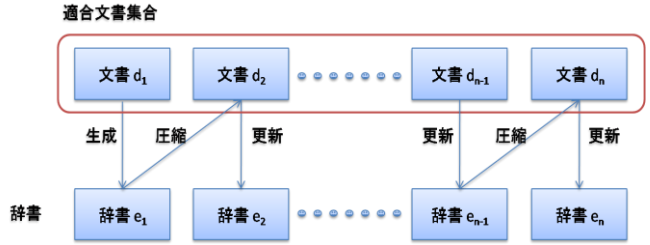


図 4 改良型 PRDC における逐次圧縮による辞書の更新

以上の改良型 PRDC は，適合文書集合を統合することから，様々なトピックを集約している．複数のトピックを持つ適合文書集合を用いて，一度の圧縮により類似度を計算す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するため，トピックごとで頻出するような文字列ほど，最長一致の文字列として辞書に登録される．このことは，同じトピックの文書に対して高い圧縮率を示すことが期待できる．また，改良型 PRDC は図 4 に示すように，文書を逐次圧縮することで辞書を更新しながら生成することが可能である．このことから，新たな文書が適合文書集合に加えられたとしても，辞書を再構築することなく，更新することが可能である．

また，文献[6]でも行われているような，前処理による類似度の精度向上が可能であり，本稿では，適合文書毎に形態素解析器の茶筌[d]により名詞のみを切り出し，統合文書 D を生成する．

d) http://chasen-legacy.sourceforge.jp/

Vol.2009-NL-194 No.8
2009/11/17

4. 改良型 PRDC による類似度の有効性検証

複数のトピックを含む適合文書集合を用いた際の改良型 PRDC による類似度の有効性について検証した．具体的には，毎日新聞 94 年の記事において 20 個のトピックを選択し，それぞれのトピックに関連した記事から適合文書集合を生成した．この適合文書集合 G を用いて，同じトピックの文書を取得できる性能について検証した．この実験では，同じトピックの文書をユーザが興味のある適合文書として仮定するため，性能が高ければユーザにあった文書を推薦する性能が高いということである．以下に，実験の詳細を示す．

4.1 比較手法

比較手法として，図 5 に示すように適合文書集合 G を統合した文書 D の文書ベクトル V と文書 x の文書ベクトル X のコサイン類似度を用いた．統合文書 D は，そのまま適合文書集合 G を表すため文書 x と適合文書集合との類似度 sim（G，x）は，以下のように与えた．

$$\mathrm{sim}(x,G) = \mathrm{sim}(x,D) = \frac{V \cdot X}{|V||X|}$$

文書ベクトルには茶筅によって切り出された名詞のみを使用し，語の重みには，TF-IDF を採用した．

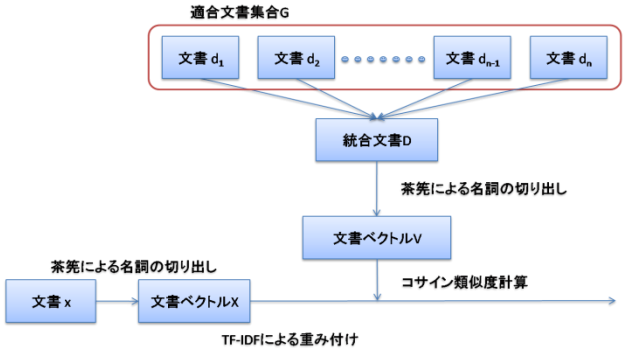


図 5　比較手法の概要

4.2 実験データ

毎日新聞 94 年の記事において 20 個のトピックを選択し，それに関連した記事を実験データとして用いるが，対象データには 102545 個の記事が含まれているため，トピックを人手で選択することは非常に困難である．そこで LDA[7]を用いて，機械的にトピックを抽出し，この結果の一部を人手により検査して評価データとした　　．

表 1　得られた 20 トピックの内容と記事数

トピックの内容	記事数
スポーツ個人	84
首相の行動記録	74
政治改革	71
遺跡・古器	70
インサイダー取引	64
宇宙	64
税制改革	62
景気概況	61
殺人事件	61
汚職事件	58
健康	57
ゴルフ	55
歌舞伎	55
個人献金問題	54
エアバス事故	52
暴力団による事件	51
空港運営	47
オーケストラ	46

LDA のツールとして GibbsLDA++[e]を利用した．このツールにおいて，潜在トピック数 T とトピック数に伴うパラメータ α 及びトピックの更新回数を指定する必要がある．本稿では，毎日新聞 94 年の記事がどれほどのトピック数を有しているかは未知であったため，潜在トピック数 T を 100～300 まで 50 ステップ毎に変化させ LDA を適用した．α は，ツールの例に従い，50000/T とした．また，トピックの更新回数は各実行において共通の 2000 回とした．

次に，以上の予備実験結果から，潜在トピック数を 250 として，記事のトピックらしさを表す確率から単一のトピックのみに属す記事を抽出し，なおかつ，クラスタリング結果を検査する人手の負担も考慮し，40 件から 100 件までの記事を含む 20 個のトピックを選択した．

e) http:// gibbslda.sourceforge.net

表 2　各手法におけるトピック数毎の F 値の平均と性能差

トピックの数	比較手法の F 値の平均	改良型 PRDC の F 値の平均	性能差
1	0.990	0.965	-0.035
2	0.926	0.906	-0.020
3	0.870	0.874	+0.004
4	0.816	0.855	+0.039

4.3 評価方法

20 個のトピックから複数のトピックを選択し，選択したトピック毎に，10 件の適合文書をランダムに取得した．適合文書集合 G を生成する．評価に用いる記事の数を 300 とし，10%に対して，選択したトピックの文書を興味のある適合文書として混ぜた．この際，選択したトピック数で均等に混ぜた．また，残りの 90%に対して，選択しなかったトピックの文書を興味のない不適合文書として混ぜた．この際，選択されなったトピック数で均等に混ぜた．例えば，トピックを 2 つ選んだとして，300 件のうち 30 件は適合文書となるため，15 件ずつ選択したトピックから混ぜる．また，300 件のうち 270 件は不適合文書となるため，15 件ずつ選択されなった 18 トピックから混ぜる．以上の適合文書集合 G と 300 件の各記事 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．

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

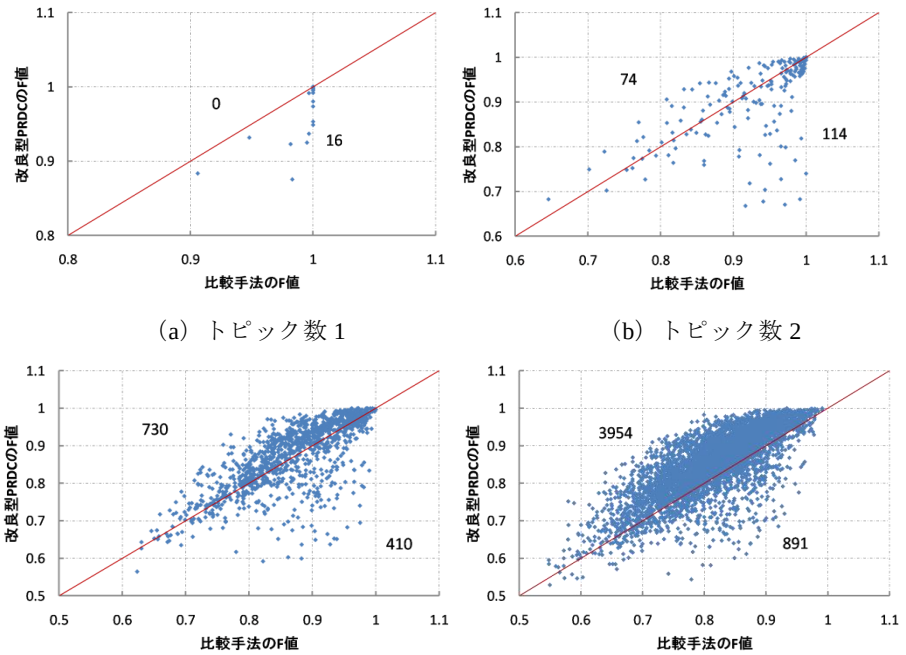
$$\text{適合率} = \frac{\text{順位 r までに現れた適合文書の数}}{\text{順位 r}}$$

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

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，選択したトピックの選び方について 10 回実行し平均化した．この値が高いほど適合文書集合 G にあった適合文書を取得できたことになる．

4.4 実験結果

表 2 は, 選択したトピック数毎での各手法の F 値の平均を示している. この表から, トピック数が少ない状態では，比較手法が改良型 PRDC にわずかに勝っているが，トピックが増えるにつれ，改良型 PRDC が逆転する．また，トピックの増加に対して，比較手法の F 値の下げ幅は，改良型 PRDC に比べ大きい．トピック数 4 においては，改良型 PRDC は高い性能を維持している．　図 6 の（a）から（d）はそれぞれ，選択



（a）トピック数 1　　（b）トピック数 2

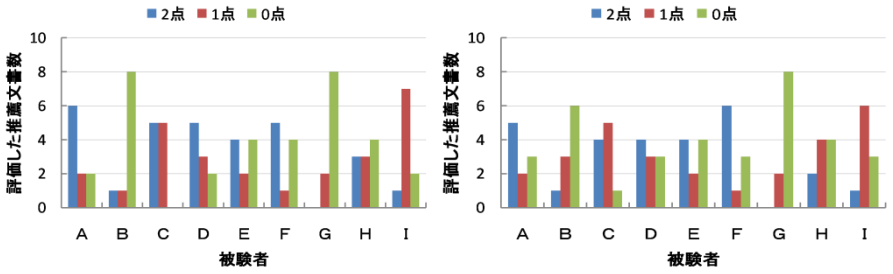
（c）トピック数 3　　（d）トピック数 4

図 6　各トピック数における改良型 PRDC と比較手法の性能

したトピックを 1 から 4 とした場合の比較手法と改良型 PRDC の F 値をプロットしたものである．個々の図において，中央の線よりも軸側に多くのプロットがあれば，性能が高いと解釈できる．表 2 の結果同様，　トピックが少ない状態では，改良型 PRDC は劣るものの，トピックが増えるにつれ比較手法に対して勝っていることが分かる．トピック数 4 に至っては, 全体の約 80%が改良型 PRDC の軸側にプロットされている．

以上の結果から，改良型 PRDC はトピックの増加に対して，頑強性が高いことが分かる．これは，各トピックにおける部分文字列が TF-IDF 以上にトピックの特徴を捉えていることを示す．

この実験から，複数のトピックを含む適合文書集合を用いた改良型 PRDC の類似度は有効に働くと考え，文書推薦のための類似度として用いる．



（a）視線情報による適合文書集合　（b）閲覧ページ全体による適合文書集合

図 7　各適合文書集合を利用した場合の推薦文書に対する評価

5. 予備実験：視線情報を用いた文書推薦の効果検証

5.1 実験内容

9 人の被験者に，Yahoo ニュースのトピックス[f]に含まれる 8 つのトピックのうち 3 つを選んでもらい，そのトピック内の記事を記事数の制限は設けずに 15 分間読んでもらった．この間に，我々が構築したシステムにより閲覧したページと視線情報から閲覧ページ中の注視領域を抽出した文書を保存した．ちなみに，トピックは「国内」「経済」「海外」「地域」「エンターテインメント」「スポーツ」「サイエンス」「コンピュータ」である．

視線情報を用いた文書推薦の効果を検証するため，視線情報から得た閲覧ページの一部を適合文書とした場合と閲覧ページ全体を適合文書とした場合について，2 通りの適合文書集合を用い改良型 PRDC の類似度から文書推薦を行った．なお，実験日の翌日の記事を 10 件ずつ推薦した．

推薦された文書に対する評価は，実験中に読んだ記事に関連しているが 2 点，おおよそ関連しているが 1 点，まったく関連していないが 0 点とした．この実験において，被験者が選んだトピックで読んだ記事は興味のある文書として仮定するため，関連する文書もまた，興味のある文書と仮定する．

5.2 実験結果

図 7 は視線情報を用いて閲覧ページから抽出した適合文書集合と閲覧ページ全体を用いた適合文書集合のそれぞれについて，改良型 PRDC の類似度から推薦された文書に対する評価を点数毎に集計した図である．この予備実験の数値結果では，視線情報を用いたことでの優位性は明らかではないが，閲覧ページ全体を用いた場合について高い評価値が得られた被験者について分析を行い，解決するべき課題として以下の 3 点が見出された．まず一つ目は，抽出する画像の画質改善によって OCR 精度を向上させることである．被験者の閲覧ページから抽出した文書には解析不能になっているものが多々あり，そのほとんどは抽出した画像がぼやけていたためだった．二つ目は，よそ見の抑制方式の工夫である．よそ見の防止対策が不十分であったことから，注視領域が実際より広く取られてしまっていたケースが見受けられた．最後は，十分な実験期間を設定することがある．被験者の興味によっては，多くの記事を参照したり，一つの記事をじっくり参照したりと，被験者によって異なるため，ユーザ毎で興味のある記事の量が異なる．このことから，視線情報を用いた文書推薦の効果を測るには十分に実験時間を設定する必要がある．

f) http://dailynews.yahoo.co.jp/fc/

6. おわりに

本稿では，文書を推薦するためのシステムとして，視線情報を用いたユーザの興味箇所の抽出と圧縮アルゴリズムを利用した文書類似度の計算手法を提案した．まず，複数のトピックを含む適合文書集合に対し，改良型 PRDC を用いることでトピックを解析することなく推薦文書との類似度を計算する手法を提案した．実験では，改良型 PRDC の類似度が，トピックの増加に対して頑強性があることを確認し，その有効性を示した．また，視線情報を用いた文書推薦の効果を検証する予備実験では，システムの改善点や，より長い期間で実験を行い，視線情報を用いた文書推薦の効果を測る必要があると分かった．今後は，システムを改善し，視線情報を用いた文書推薦のユーザ実験を行いたい．

参考文献

1) Jaime T., Susan T. D., Eric H.: Personalizing search via automated analysis of interests and activities, SIGIR, pages 449-456(2005).
2) P. A. Chirita, C. S. Firan, and W. Nejdl: Personalized query expansion for the web, SIGIR, pages 7-14 (2007).
3) Cilibrasi R., Vitanyi P. M.B. : Clustering by compression, IEEE Transactions on Information Theory, Vol.51, No.4, pages 1523-1545 (2005).
4) G. Buscher, A. Dengel, L. van Elst, F. Mittag: Generating and using gaze-based document annotations, Conference on Human Factors in Computing Systems, pages 3045-3050 (2008).
5) 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，情報処理研究報告, Vol.2006, No.94, pages 65-70 (2006).
6) Helmer S.:Measuring the structural similarity of semistructured documents using entropy, In Proceedings of 33rd international Conference on VeryLarge Data Bases, pases 1022-1032 (2007).
7) David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent dirichlet allocation, The Journal of Machine Learning Resarch, Vol.3, March 2003, pages 993-1022 (2003).