CD-004

新聞アーカイブシステム KENBUN での 類似ランドマーク表示によるブラウジング支援
Browsing Support by Displaying Similar Landmarks for the Newspaper Archive System KENBUN

中尾 優太† 中島 誠‡
Yuta Nakao Makoto Nakashima

1. はじめに

図書館が所蔵する古い時代の新聞は，様々な分野の活動や社会情勢を伝える重要な情報源であり，新聞記事への容易なアクセスを可能にするアーカイブシステムへの要求が高まっている[1]．現在開発中の KENBUN[2]では，記事の内容を簡潔に表現しており，利用者が注視しやすい見出し，写真，図，広告をランドマークとして抽出して表示し，利用者が選択したランドマークと類似したランドマークを表示することで新しい気づきを与えるブラウジングを提供する．本稿では，ランドマークの抽出方法と類似ランドマークの選定方法，表示方法について述べ，評価実験から，ブラウジングの支援に対する有効性を示す．

2. 新聞アーカイブシステム KENBUN

本章では新聞アーカイブシステム KENBUN[2]の概要について説明し，ブラウジングの問題点について述べる．

2.1 KENBUN の概要

KENBUN は，大分県立図書館から提供された明治 9 年（1876 年）から昭和 42 年（1967 年）までに大分県内および周辺地域で発刊された計 25 社分の地方新聞のマイクロフィルムをスキャニングして得た 177,576 頁の紙面画像を保持している．ブラウジングを中心とした記事探索によって利用者が望んだ新聞紙面を探すことができる，ブラウジング指向の新聞アーカイブシステムであり，プロトタイプの KENBUN は，現在，大分県立図書館，中津市立小幡記念図書館，大分大学図書館，別府市立図書館にて稼働中である．

KENBUN のブラウジング画面では，図 1 に示すように，新聞紙面画像を発刊年月順に画面左上から右下にかけて一か月分を折り返しながら配置している．

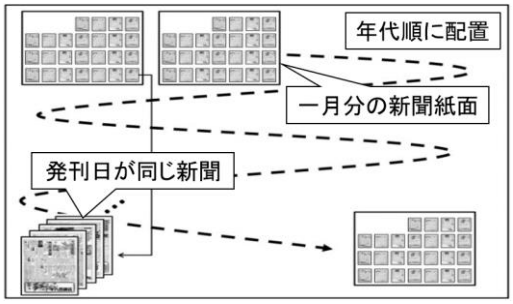

図 1　KENBUN の新聞紙面の配置方法

2.2 ランドマーク表示

前述したとおり，KENBUN の初期画面では限られた表示領域に新聞紙面を表示するために，月ごとに新聞紙面画像をまとめて表示している．図 2 に示すように，利用者は目的の新聞紙面がある月または日を選択することで，新聞紙面を拡大して閲覧することができるが，拡大していない状態や，一か月分を拡大している状態だと，新聞紙面の内容を把握することが難しい．大量の情報を含む新聞紙面画像が配置されている中から，記事の内容を容易に把握できなければ，円滑なブラウジングに支障が出てしまう．

新聞紙面内でも，見出しや写真，図，広告など小さくとも内容が視認しやすく，記事内容を示唆するものがある．これらをランドマークとして利用して，記事の内容を容易に把握することができるようにして，興味を引く記事をなるべく多く目に触れやすくする方法を提案する．

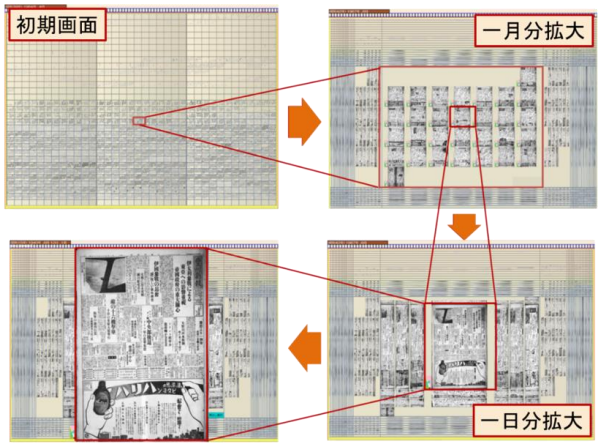

図 2　KENBUN の拡大の流れ

図 3 に，提案する類似ランドマーク表示の例を示す．ブラウジング中の利用者は好きなタイミングで，新聞紙面の中からランダムに選ばれたランドマークを表示することができる（図 3 右）．利用者が表示されたランドマークの中から気になったものを選べば，そのランドマークを含む紙面を拡大表示する．さらに，次のブラウジングのきっかけとするために，利用者が選択したランドマークと見た目あるいは内容が類似したランドマークを表示する（図 3 左）．

† 大分大学大学院工学研究科　Oita University Graduate School of Engineering
‡ 大分大学理工学部　Oita University Faculty of Science and Technology

図 3　ランドマークのランダム表示例（右）と
類似ランドマーク表示例（左）

3. 実装方法

類似ランドマーク表示機能の実現には，ランドマークの抽出，ランドマークの類似度判定，表示機能の実装が必要となる．

3.1 ランドマークの抽出

新聞紙面画像からのランドマークの抽出にはニューラルネットワークである YOLOv3[3]を用いた．教師データは，アノテーションツールである LabelImg[4]を使用して，見出し，写真，図，広告の 4 種のランドマークを矩形で選択することで作成した．作成した教師データ 769（見出し 6,863，広告 3,127，写真 1,656，図 460）個に，画像の色味を変化させたものとノイズを加えたものを加えて 10 倍に水増ししたものを用意して学習させた．表 1 に主要な新聞社の新聞紙面画像 90 枚を対象にテストを行った時の結果（再現率と適合率と F 値）を示す．図以外では F 値が 0.5 を超えた．教師データが少なかった図についても，適合率では高い値を示せた．この学習したネットワークを使って，約 17 万枚の新聞紙面から 4 種のランドマークを約 210 万個抽出した．ランドマークごとの内訳を表 2 に示す．

表 1　テスト結果の再現率と適合率と F 値

	再現率	適合率	F値
見出し	0.51	0.89	0.65
広告	0.37	0.78	0.50
写真	0.66	0.79	0.72
図	0.20	1	0.33

表 2　抽出したランドマークの内訳（個）

見出し	1,301,594
広告	511,571
写真	311,778
図	49,629
合計	2,174,572

3.2 ランドマークの類似度判定

多様な記事に気づくブラウジングのきっかけとするために，利用者が選択したランドマークと見た目が類似したランドマークを表示する．そのために，抽出したランドマーク間で見た目の類似度判定を行った．また，見出しについては，そこの書かれたテキスト（見出し文）の意味に関する類似度判定も行った．

3.2.1 ランドマークの見た目の類似度判定

見た目の類似度判定には Average Hash[5]を用いた．Average Hash では画像をリサイズ，グレースケール変換し，各ピクセルを RGB 値の全体の平均値をもとに二値化する．二値化したデータからハッシュ値を求め，比較したい画像同士のハッシュ値を比較することで，類似度（[0,1]）の判定を行う．値が 1 に近いほど類似している画像となる．今回は閾値を 0.5（見出しは 0.6）とし，閾値以上の類似度が判定されたものを類似画像とした．図 4 に見た目の類似画像の例と基準画像に対するそれぞれの類似度を示す．

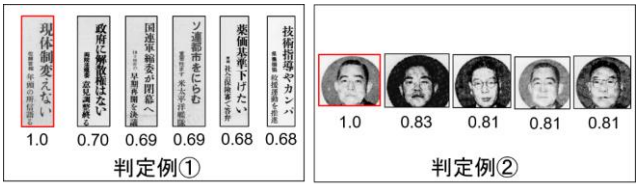

図 4　見た目の類似度判定例（赤枠は基準画像）

3.2.2 ランドマークの意味の類似度判定

ランドマークのうち，見出しについては，見出し文に含まれるテキストを OCR で抽出すれば，意味的な類似度をもとにした．見出しの内容の類似度判定もできる．OCR は Google ドライブの OCR 機能[6]を使用した．内容の類似度判定は word2vec で学習済みの日本語 Wikipedia エンティティベクトル[7]を使用して，見出し文のベクトル同士の類似度を，見た目の類似度同様にコサイン類似度で求めた．図 5 に意味的な類似度の判定結果例を示す．今回は見た目の類似度が高い見出し同士のみで，意味的な類似度の判定を行ったため，それほど意味的な類似度の高いものはない．

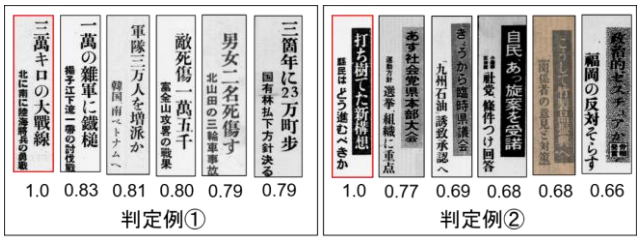

図 5　意味的な類似度判定例（赤枠は基準画像）

3.3 ランドマーク表示機能の実装

抽出したランドマーク画像と類似度の判定結果を用いて KENBUN での類似ランドマーク表示機能の実装を行った．

3.3.1 ランドマークの表示

利用者はランドマーク表示ボタンを押すことで，任意のタイミングでランドマークを表示することができる．始めは，見出し，写真，図，広告からランダムに表示を行う．利用者がその中から，ランドマークを選択することで，選

択したランドマークを含む新聞紙面を拡大してブラウジング画面に表示する．さらに，見た目が類似しているランドマークがある場合は，それらを類似度が高い順に表示する．また，意味的な類似度が高いランドマークがある場合は，図 6 に示すチェックボックスを使用できるようになり，見た目の類似表示から，意味的な類似表示に切り替えられる．類似したランドマークがある紙面は，図 7 で示すように，ブラウジング画面上でハイライト表示して，見つけやすくする．見た目が類似したランドマークがない場合は，利用者が選択したランドマークと同じ種類で類似度が高いランドマークをランダムで表示する．

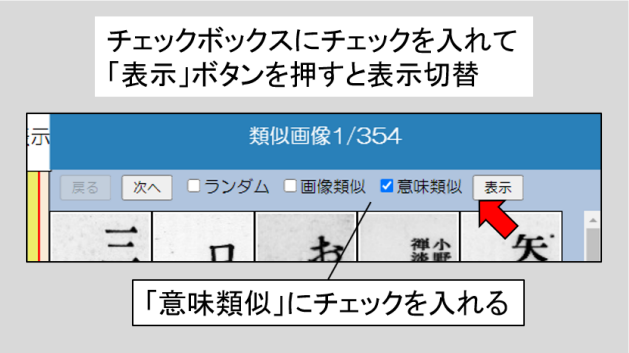

図 6　意味的な類似表示への切り替え

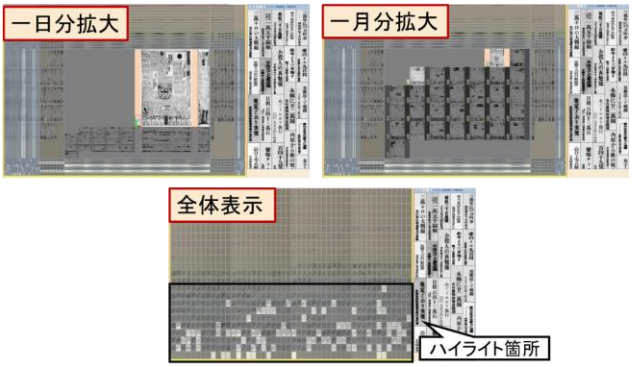

図 7　ブラウジング画面上でのハイライト表示

4. 評価実験

類似ランドマーク表示によるブラウジング支援の有効性を検証するために，理工学部の学部生および大学院生，計 12 名を被験者として評価実験を行った．

4.1 実験概要

被験者には KENBUN と類似ランドマーク表示機能の操作方法を説明し，類似ランドマーク表示を使用しない場合（タスク 1）と使用した場合（タスク 2）で 10 分間ずつ興味深いと思う記事の探索を行ってもらった．興味深い記事があった場合は，その新聞紙面の新聞社，発行年月日，刊区分，頁，どの記事かを用紙に記入してもらった．タスク終了後，アンケートに回答してもらった．回答には 7 段階のリッカート尺度を用いた．

4.2 実験結果

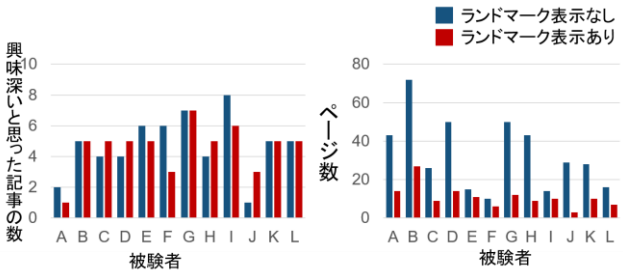

図 8　興味深いと思った記事の数（左）と
閲覧したページ数（右）

図 8 に被験者が時間内に閲覧したページ数と興味深いと思った記事の数を示す．タスク 1 とタスク 2 で興味深いと思う記事の数に差はなかったが，タスク 2 では，全被験者がタスク 1 に比べて閲覧したページ数が少なかった．ランドマーク表示を使用した方がより少ない労力で興味深い記事を発見できたことが分かる．

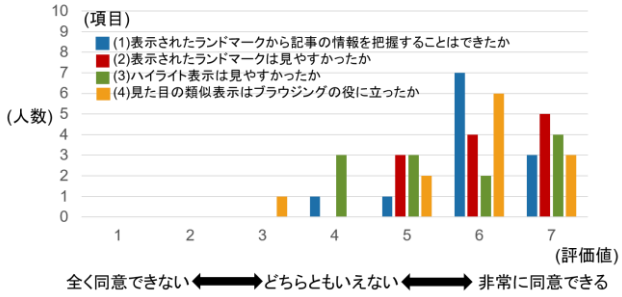

図 9　アンケート結果

図 9 にアンケートの結果を示す．（1）から（4）までの 4 つのアンケート項目において，評価値 7，6，5 を肯定意見，3，2，1 を否定意見と捉えると，いずれも優位に肯定意見が多いことが確認できた（二項検定，有意水準 0.05）．また，タスク 1 とタスク 2 ではどちらの方がブラウジングしやすかったかの質問には 12 名中，10 名がタスク 2 と答えた．なお，今回，意味的な類似度利用した被験者が 1 名であったため，意味的な類似表示についての評価は得られなかった．

4.3 考察

アンケート項目の（1）で有意に肯定意見が多いことおから，今回，自動的に抽出した 4 種のランドマークは正しい領域のものであり，それらを表示することで，紙面内の情報を把握することに役立っていることが分かった．また，ランドマークは表示領域内に隙間なく敷き詰めるように配置しており，アンケート項目の（2）で肯定意見が多いことからランドマークの配置方法もランドマークの閲覧に有効であることが分かった．さらに，アンケート項目の（4）で有意に肯定意見が多いことから，ブラウジングのきっかけとしては，内容に注目していない，見た目の類似ランドマーク表示でもブラウジングの支援に有効であることが分かった．

アンケート項目の（3）ではどちらともいえないと答えた被験者が数名いた．これはハイライト表示によって，対象の紙面以外が暗く表示されることにより，類似ランドマークを含まない他の紙面のブラウジングが制限されてしまったためと考えられる．今後，ハイライト表示以外の対象紙面の提示方法を検討する必要がある．

図 10 と図 11 に被験者がタスク 1 とタスク 2 で興味深いと思った記事のランドマークの典型的な例を示す．被験者が選んだ記事のランドマークを赤枠で示している．タスク 1 でブラウジングを行う場合，図 10 で示しているように，目につきやすい，大きな見出しや広告を興味深い記事として選ぶ被験者が多くいた．それに対して，タスク 2 では，図 11 のようにランドマークの大きさに関係なく，興味深い記事を選択していた．また，使用後の感想（自由記述）でも「ランドマーク表示を使用せずにブラウジングを行う場合は，広告や図などに目が行く」という意見が得られた．これらのことから，ランドマーク表示を使用することで，通常のブラウジングでは見落としてしまう記事を気づき易くでき，ブラウジングの支援に繋がっていると考えられる．

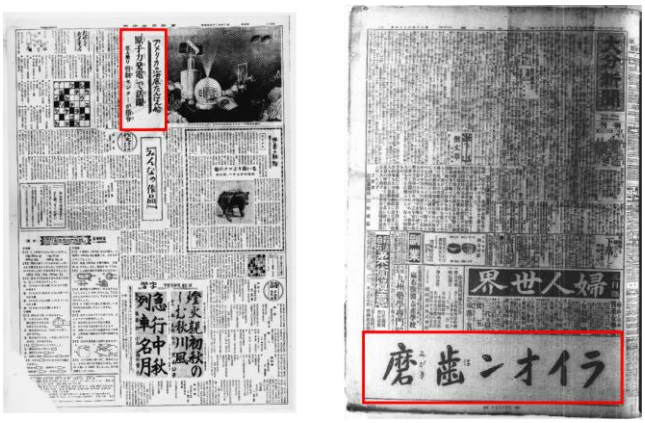

図 10　タスク 1 で興味深いと思った記事の例

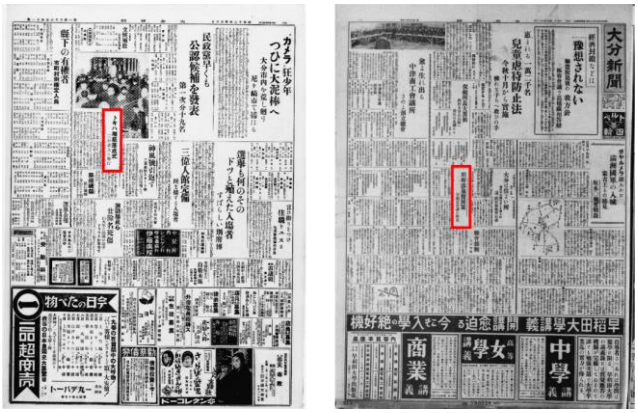

図 11　タスク 2 で興味深いと思った記事の例

その他に「タスク 1 では，選択した月の周辺の時期の紙面を見てしまうため，偏った月日の新聞しか見ることができなかった」という意見や，「ランドマーク表示を使用しないで記事のブラウジングを行う場合，紙面の 1，2 頁目に気になる記事がないとその次の頁を見ようと思わなくなる」といった意見が得られた．図 12 に被験者が興味深いと思った記事を含む紙面のページを示す．タスク 1 では，興味深いと思った記事を 1 頁目から選択している被験者が多く，それに対して，タスク 2 では興味深い記事を見つけた紙面の頁にばらつきがみられる．そのため，ランドマーク表示を利用することで，様々なページの記事を閲覧できていることが分かった．

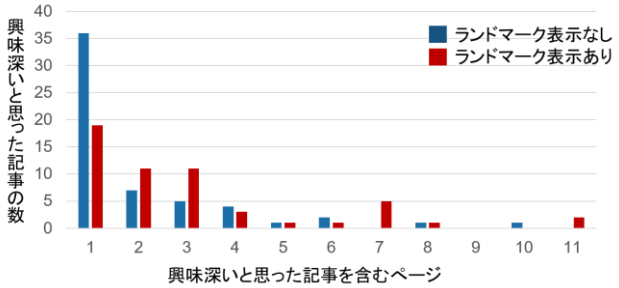

図 12　被験者が選んだ記事を含む紙面のページ

今回の実験で意味的な類似度の高いランドマークの表示を使用する被験者が少なかったのは，意味的に類似したランドマークがある場合とない場合でのチェックボックスのレイアウトの違いが少なく，意味的に類似したランドマークがあるかどうかが分かりづらかったためだと考えられる．そのため，類似ランドマークの表示の切り替え方法は今後検討していく必要がある．また，今回の実験では，意味的な類似度が高いランドマークのブラウジングの支援への有効性を検証できなかったため，今後検証を行っていく必要がある．

5. おわりに

本稿では，新聞アーカイブシステムでのブラウジングを提供するにあたって，記事内容を把握しやすい 4 つのランドマークを設定し，これらを表示することで，利用者により広い範囲でのブラウジングを促す方法について提案した．被験者実験から，見た目が類似したランドマークを表示することで，目的なく記事のブラウジングを行いながら，興味深い記事を探すきっかけとして有効であることが確認できた．今後の課題としては，実験で評価できなかった，意味的に類似したランドマークを表示することの有効性と，ランドマークの表示切替方法，ハイライト以外のランドマークを含む紙面の提示方法の検討などが挙げられる．

参考文献

[1] 佐々木美穂，“英国とオランダの国立図書館にみる新聞資料デジタル化プロジェクト”，カレントウェアネス，No.309，CA1750（2011）．
[2] 中尾優太，中島誠，“新聞アーカイブシステム KENBUN でのブラウジングのためのランドマーク表示”，2020 年度電気・情報関係学会九州支部連合大会（第 73 回連合大会）講演論文集，06-1A-10（2020）．
[3] J. Redmon，A. Farhadi，“Yolov3: An incremental improvement”，*arXiv*，*1804.02767*（2018）．
[4] tzutalin，”LabelImg”，https://github.com/tzutalin/labelImg..
[5] siny，“機械学習で画像分類をしてみよう【python】”，https://sinyblog.com/%E6%A9%9F%E6%A2%B0%E5%AD%A6%E7%BF%92/average_hash/.
[6] Google ドライブ，https://www.google.com/intl/ja_jp/drive/.
[7] Masatoshi Suzuki，“WikiEntVec”，https://github.com/singletongue/WikiEntVec/releases.