ツリーマップを用いたデジタルアーカイブ探索支援システムの構築

河辺 雅史[1,a)]　奥野 拓[2]

概要：近年，歴史資料のデジタル化・公開を行うデジタルアーカイブが多数公開されている．しかしデジタルアーカイブには，公開されている歴史資料の全体像を把握しながら情報探索を行う仕組みがない．そのため，興味のある歴史資料の発見は容易ではないという問題がある．本稿では，歴史資料の分類の視覚化により歴史資料の探索をすることができるシステムを提案する．本研究では，歴史資料の資料タイトルのベクトル表現を特徴ベクトルとして階層的クラスタリングを行い，歴史資料を分類する．また，ツリーマップと呼ばれる視覚化手法を取り入れることで歴史資料を探すことを支援する．

Building a Digital Archive Search Support System by Using Treemap

Masashi Kawabe[1,a)]　Taku Okuno[2]

Abstract: In recent years, many historical records are digitalized, and there are many websites of digitalizing and publishing historical records. However, users can't search historical records while grasping the whole image. Therefore, there is a problem that a user can't find historical records easily. In this paper, we propose an exploratory search support system that a user can search historical records by visualizing classification of historical records. In this research, we build a classification method by hierarchical clustering with vector of historical records title as feature vectors. Moreover, we support a user to search historical records by using visualization technique called treemap.

1. はじめに

近年，歴史資料のデジタル化・公開を行うデジタルアーカイブが多数公開されている．これらのデジタルアーカイブには，ポスター，絵葉書，写真，浮世絵など様々なカテゴリのデジタル化された歴史資料の画像がメタデータと共に公開されている．デジタルアーカイブにおいて歴史資料を探す方法は，メタデータに対するキーワード検索が主流となっている．しかし，キーワード検索は予備知識を必要とし，適切なキーワードを入力しなければユーザが求める歴史資料を発見することができないという問題がある．また，必ずしもユーザの検索対象が明確であるとは限らないため，興味のある歴史資料の発見は容易ではない．キーワード検索の他にも，時代や分野などを選択することで，歴史資料を探すことができる．特定の時代や分野の歴史資料が充実しているような場合は，時代や分野を選択する際の判断基準の一つになると考えられる．しかし，デジタルアーカイブには膨大な数の歴史資料が公開されており，資料分類とその割合等の全体像を把握しながら歴史資料を探すことは容易ではないという問題がある．

そこで本稿では，全体像の把握を容易にし，興味のある歴史資料の発見を支援するデジタルアーカイブ探索支援システムを提案する．本手法では，デジタルアーカイブに公開されている歴史資料の全体像の把握を容易にするために，歴史資料を分類する．そして，分類結果を歴史資料の全体像の把握が容易な形式で視覚化することで歴史資料を探すことを支援する．

[1] 公立はこだて未来大学大学院
Graduate School of Future University Hakodate
[2] 公立はこだて未来大学
Future University Hakodate
[a)] g2116012@fun.ac.jp

2. 関連研究

2.1 探索型検索の支援

情報要求が曖昧な状況における検索を支援する仕組みとして，探索型検索 (Exploratory Search) という概念がある [1]．探索型検索は，ユーザの情報要求が曖昧であり適合判断が困難である．そのため，ユーザは繰り返し検索を行うことで情報要求を明確化する．また探索型検索では，ユーザは検索対象に関する知識が乏しいなかで検索を行うため，クエリの修正が容易である必要がある．

2.2 大量画像の探索支援

Bederson らは，ツリーマップを用いて大量画像のブラウジングをサポートするシステム PhotoMesa の構築を行っている [3]．ツリーマップとは，二次元平面上の領域を入れ子状に分割した矩形を生成することで，木構造データを視覚化する手法である．ツリーマップで生成する各矩形の縦横比が 1 に近いほど、一般的に魅力的であると言われている [3]．しかし，各矩形の配置の順序が整列されるにつれて細長い矩形が生成されやすくなり，縦横比は 1 から離れるという特徴がある (図 1 (左))．また，縦横比を 1 に近づけるにつれて、各矩形の配置に突然の角度変化が発生し，順序が予測しにくくなるという特徴がある (図 1 (右))．この研究では，入力として与えられたツリーマップで生成する各矩形における，画像等のオブジェクトを含むための十分な大きさの矩形を生成する Quantum Treemap [4] を利用することで大量画像の視覚化を行っている．

五味らは，大量画像群の一覧表示と詳細度制御のためのインタフェースを持ち合わせた可視化手法 CAT の提案を行っている．CAT では，画面空間充填による階層型データ可視化手法「平安京ビュー」[5] を利用している．この研究では，各画像に 1 個以上のキーワードが付与されていると仮定している．まず，WordNet Similarity によって算出したキーワード間の距離に基づいて画像をクラスタリングする．次に，各クラスタに属する画像の色情報および周波数情報に基づく特徴量から特徴ベクトルを算出し，ベクトル間の余弦に基づいてクラスタリングすることで画像を階層的に分類している．また，各クラスタに対して代表画像を選定している．システムでは，ズームアウト時には代表画像を表示する．そして，ズームイン時には低階層の代表画像を表示する．さらにズームインを進めた際には，最下層クラスタの各々のサムネイル画像を表示する (図 2)．平安京ビューは Quantum Treemap との比較実験において，生成する各矩形の縦横比を 1 に近づけることができるという結果が得られている [5]．しかし，二次元平面上の領域と各矩形の間に空白が形成されやすい傾向にあり，スペース効率に課題があると考えられる．

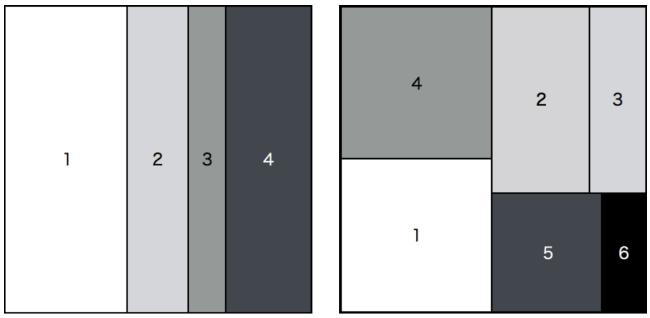

図 1　ツリーマップで生成する矩形の例

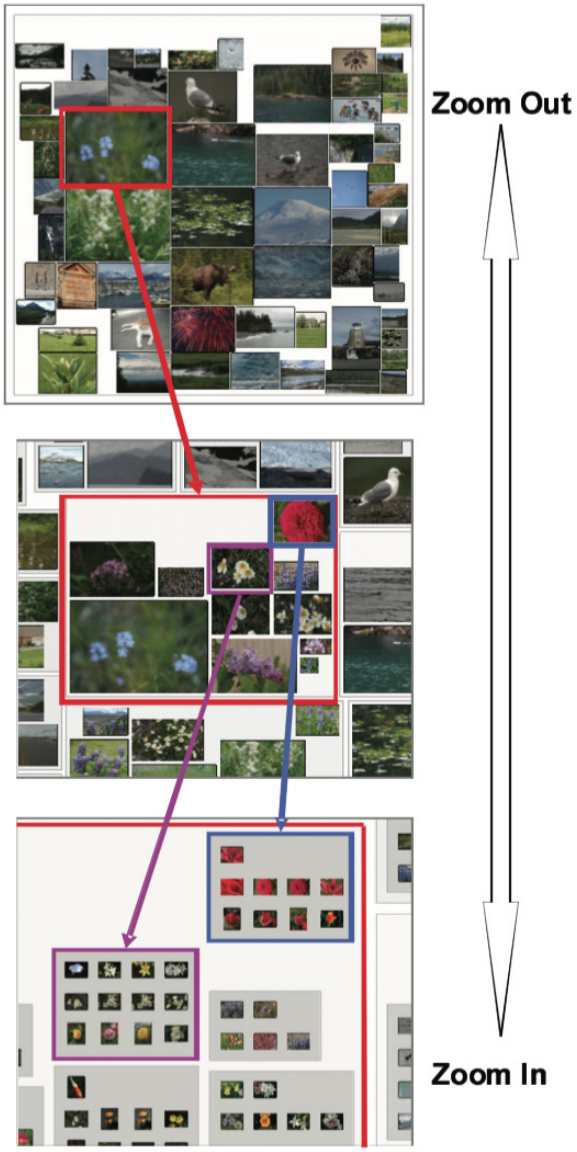

図 2　CAT の概観 (参考文献 [6] より転載)

3. 歴史資料探索支援システム

本研究では，試行錯誤しながら探索型検索を行うことを可能にし，興味のある歴史資料の発見を支援するシステムを構築する．デジタルアーカイブは検索対象が明確なユー

ザには有用である．しかし，歴史資料の数が膨大なため，全体像の把握が困難であるという問題がある．また，歴史資料は分類されているが視覚化されておらず，それぞれの分類の割合等の全体像を直感的に把握することが容易ではないという問題がある．歴史資料の分類が視覚化されている場合は，検索対象が曖昧な場合においても適切な検索クエリを考える必要はなく，歴史資料の探索ができると考えられる．そこで，歴史資料を分類し，分類の結果を全体像の把握が容易な形式で視覚化する．

3.1 ツリーマップによる視覚化

デジタルアーカイブでは膨大な数の歴史資料を取り扱っている．そのため，分類結果の視覚化手法は，膨大な数の画像の視覚化に適した手法が望ましい．そのため本研究ではツリーマップを利用する．

ツリーマップにはいくつかのアルゴリズムがあり，それぞれ特徴がある．ツリーマップを生成する 5 つの主要なアルゴリズムとその特徴を表 1 に示す．

本研究で扱う各分類には順序関係は無いため，ツリーマップで生成される各矩形の順序は重要な要素ではない．そのため，本研究ではツリーマップで生成される各矩形の縦横比を最も 1 に近づけることができる Squarified Treemap[8] を利用する．また，ツリーマップでは各矩形を色分けすることで視認性が向上する [9]．そのため，ツリーマップの各矩形を色分けする．

表 1 ツリーマップを生成する 5 つのアルゴリズムと特徴 (参考文献 [7] を参考に作成)

Algorithm	Order	Aspect ratios	Stability
BinaryTree	partially ordered	not very good	stable
Ordered	partially ordered	medium	medium stability
SliceAndDice	ordered	very bad	stable
Squarified	unordered	best	medium stability
Strip	ordered	medium	medium stability

3.2 システム動作

図 3 に提案システムの動作を示す．本システムは分類一覧画面に，歴史資料を分類した結果を視覚化したツリーマップを表示する．ツリーマップの各矩形には，デジタルアーカイブに公開されている歴史資料の特徴を表す分類，分類名，分類に属する歴史資料のサムネイル画像を表示する．

画像を含む分類をクリックすることで，クリックした分類に属する歴史資料の画像と資料タイトルが表示される．資料一覧画面で並べ替え項目の「資料名」をクリックすることによって，表示されている歴史資料を資料タイトル順に並べ替える．また，並べ替え項目の「新着」をクリックすることによって，歴史資料をデジタルアーカイブに追加された日時の降順に並べ替える．また，並べ替え項目の

図 3 システム動作

「年代」をクリックすることによって，歴史資料を出版年代の降順に並べ替える．また，並べ替え項目の「人気」をクリックすることによって，歴史資料をお気に入り登録数の降順に並べ替える．表示されている歴史資料をクリックすることによって，クリックされた歴史資料の高精細画像，メタデータを表示する．また，資料タイトル横のハートアイコンをクリックすることによってお気に入り登録する．

お気に入り登録した歴史資料は，画面上部のナビゲーションバーの「お気に入り」を選択することで表示される．また，ナビゲーションバーの「みんなのお気に入り」を選択することで，他のユーザがお気に入り登録した歴史資料がお気に入り登録数の降順で表示される．

4. 歴史資料の分類

4.1 階層的クラスタリングによる分類

本研究では函館市中央図書館デジタル資料館 [10](以下，デジタル資料館) の歴史資料を扱う．デジタル資料館は，歴史資料を死蔵させないということを念頭に置き，目録の完

成度が低い歴史資料であっても公開している [11]．そのため，内容説明の項目が空欄の歴史資料が多く存在する．デジタル資料館のポスターカテゴリの歴史資料には，ビールや包装紙などの特定のテーマに関する資料が多く公開されており，テーマに偏りがあるという特徴がある．このようなテーマ毎に歴史資料を分類することで，デジタルアーカイブに公開されている歴史資料のテーマとその割合の全体像の把握が容易になると考えられる．そこで本研究では，テーマ毎に歴史資料を分類する．

デジタル資料館におけるビールなどの特定のテーマは，資料タイトルに含まれている場合が多い．そのため，資料タイトルを利用することで歴史資料をテーマで分類することができると考えられる．そこで本研究では，資料タイトルのベクトル表現を特徴ベクトルとし，階層的クラスタリングを行うことにより歴史資料を分類する．分類は以下の手順で行う．

(1) 資料タイトルを形態素解析し，名詞を抽出する．
(2) 抽出した名詞のベクトル表現を獲得する．
(3) 1 つの資料タイトルから複数の名詞が抽出された場合，名詞のベクトル (語ベクトル) から句ベクトルを生成する．
(4) 獲得した語ベクトル，句ベクトルを正規化する．
(5) 正規化した語ベクトル，句ベクトルを特徴ベクトルとして距離を計算し，階層的クラスタリングを行う．

本研究では，名詞のベクトル表現の獲得には word2vec [12] を利用する．word2vec とは Mikolov らによって提案された，ニューラルネットワークを用いて単語の分散意味表現を獲得する自然言語処理の手法である．デジタルアーカイブには，幅広い分野の歴史資料が公開されている場合がある．例えば，デジタル資料館のポスターカテゴリに含まれる歴史資料は，図書館，会社，祭りなど，分野は多岐にわたる．そのため word2vec の学習データに用いるコーパスは，幅広い分野を網羅したコーパスである必要がある．そこで本研究では，日本語版 Wikipedia のダンプデータから作成したコーパスを利用する．階層的クラスタリングには，分類精度が高い Ward 法を利用する．階層的クラスタリングの実行後，クラスタの切断を行い分類を作成する．

4.2 クラスタに対するラベリング

階層的クラスタリングの実行後，得られたクラスタに歴史資料の特徴を表す分類名を付与する．この際，クラスタの特徴を表す分類名が複数ある場合がある．また，ユーザによって分類の解釈が異なる場合があると考えられる．例えば，「サッポロビール」，「キリンビール」，「カクテル」，「清涼飲料水」から成るクラスタがあった場合，「ビール」の分類と解釈するユーザや，「酒」の分類と解釈するユーザ，「飲料」の分類と解釈するユーザがいると考えられる．この際，複数の分類名が付与されていれば，歴史資料の特徴の把握が容易になると考えられる．付与する分類名の数は予備実験から，3 単語の場合網羅性が高いという結果が得られた．そのため本研究では，「サッポロビール」，「キリンビール」，「カクテル」，「清涼飲料水」から成るクラスタがあった場合，「ビール 酒 飲料」を分類名として付与することを目指す．

Treeratpituk らは，階層的クラスタリングで得られたクラスタに含まれる文書中に出現する単語の TF-IDF 値からスコアを求めてラベリングを行っている [13]．TF-IDF 値からスコアを算出しラベリングする手法は，メタデータである目録の完成度が高い場合は有効であると考えられる．しかし，資料タイトルは TF-IDF 値を計算するには記述量が少ない．また，内容説明の項目が空欄の資料が多く存在するため，TF-IDF 値からスコアを算出し，ラベリングする手法はデジタル資料館においては有効ではないと考えられる．そこで本研究では次の手順で，階層的クラスタリングによって得られたクラスタに分類名を付与する．

(1) クラスタに属する全ての歴史資料の資料タイトルから名詞を抽出する．
(2) 抽出した名詞のベクトル表現を獲得する．
(3) 獲得したベクトルから句ベクトルを生成する．
(4) 句ベクトルを正規化する．
(5) 正規化した句ベクトルに最も近いベクトルを持つ 3 単語を分類名とする．

本研究では，ベクトル表現の獲得，正規化した句ベクトルに最も近いベクトルを持つ単語の獲得には word2vec を利用する．word2vec の学習データには，幅広い分野を網羅している日本語版 Wikipedia のデータから作成したコーパスを利用する．

5. 評価実験

提案システムが興味のある歴史資料の発見の支援に有効か検証するため，デジタル資料館と提案システムの比較実験を行った．実験では，デジタル資料館に公開されているポスターカテゴリの歴史資料を利用した．

5.1 実験方法

28 名の学生に対して，興味のある歴史資料を探すという課題を課し，デジタル資料館と提案システムの 2 つのシステムをそれぞれ 5 分間操作してもらった．興味のある歴史資料を発見した際には，記録をしてもらった．提案システムでは，興味のある歴史資料を発見した際には，お気に入り登録をしてもらった．デジタル資料館には，歴史資料をお気に入り登録する機能がないため，興味のある歴史資料を発見した場合は，歴史資料のスクリーンショットを撮影してもらうこととした．システムの操作順序による影響を防ぐため，デジタル資料館，提案システムの順で操作してもらう被験者 14 名と，提案システム，デジタル資料館の

順で操作してもらう被験者 14 名の 2 グループに分けて実験を行った．実験終了後，以下の質問項目でアンケート調査を行った．各質問に対して，5 が最高評価，1 が最低評価の 5 段階評価で回答を得た．

質問 1: デジタル資料館は興味のある歴史資料を発見しやすかったですか？

質問 2: 提案システムは興味のある歴史資料を発見しやすかったですか？

質問 3: 提案システムは分類の割合を把握しやすかったですか？

質問 4: 提案システムの歴史資料の分類は妥当でしたか？

質問 5: 提案システムの各分類の分類名は妥当でしたか？

さらに，自由記述で意見を得た．また，被験者毎に各システムでの閲覧資料数，お気に入り登録資料数を記録した．

5.2 実験結果

アンケート結果を図 4 に示す．質問 1，質問 2 の回答より，提案システムはデジタル資料館と比較して高い評価が得られた．また，質問 1，質問 2 に対する評価値に有意差があるか分析するために対応のある t 検定を行い，片側検定における P 値を求めた．その結果，$2.08 \times 10^{-7} (p < 0.05)$ であり有意差が示された．また，差の大きさの程度を示すために効果量として Cohen's d を求めた．この値は 0.2 以上が小さい，0.5 以上が中程度，0.8 以上が大きいと判断される [14]．効果量を算出した結果，1.25 であった．また，質問 3 の回答より，分類の割合をとても把握しやすかったと思った人 (39%)，把握しやすかったと思った人 (39%) が多いことがわかった．質問 4 の回答より，歴史資料の分類が妥当であると思った人 (61%) が多いことがわかった．質問 5 の回答より，分類名が妥当であると思った人 (25%) が少ないことがわかった．

被験者毎の両システムでの閲覧資料数を図 5 に示す．提案システムの閲覧資料数が，デジタル資料館の閲覧資料数より多い被験者は 17 名 (61%) であった．また，提案システムの閲覧資料数とデジタル資料館の閲覧資料数が等しい被験者は 1 名 (3%) であった．また，提案システムの閲覧資料数が，デジタル資料館の閲覧資料数より少ない被験者は 10 名 (36%) であった．総閲覧資料数を比較した結果，デジタル資料館と比べて提案システムの方が 33.67%多かった．

被験者毎の両システムでのお気に入り登録資料数を図 6 に示す．提案システムのお気に入り登録資料数が，デジタル資料館のお気に入り登録資料数より多い被験者は 14 名 (50%) であった．また，提案システムのお気に入り登録資料数とデジタル資料館のお気に入り登録資料数が等しい被験者は 3 名 (11%) であった．また，提案システムのお気に入り登録資料数が，デジタル資料館のお気に入り登録資料数より少ない被験者は 11 名 (39%) であった．総お気に入り登録資料数を比較した結果，デジタル資料館と比べて提案システムの方が 16.77%多かった．

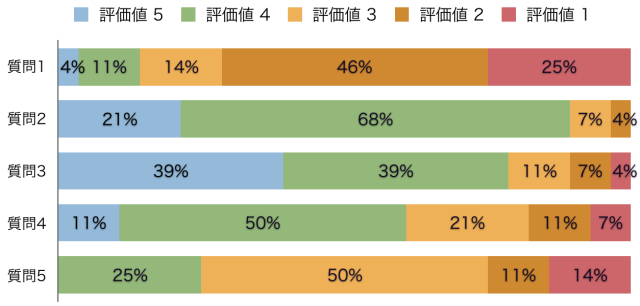

図 4 アンケート結果

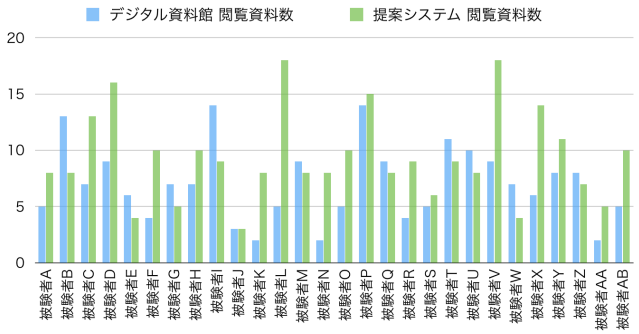

図 5 両システムの閲覧資料数

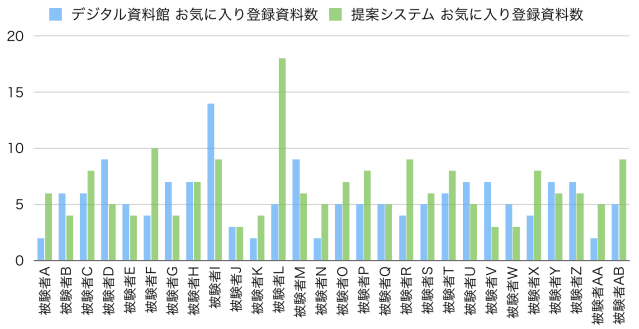

図 6 両システムのお気に入り登録資料数

5.3 考察

質問 1，質問 2 の回答，t 検定による分析の結果から，提案システムは興味のある歴史資料の発見の支援に有効であることがわかった．また，効果量を算出した結果 1.25 であったことから，提案システムはデジタル資料館と比較して，興味のある歴史資料の発見に大きな効果があったと考えられる．また，質問 3 の回答から，提案システムは分類の割合の把握が容易であることがわかった．質問 4 の回答から，歴史資料の分類の結果が妥当であることがわかった．また，自由記述に「資料の分類の割合を視覚的に見られるのは、興味を持つ可能性のある資料に辿り着くヒントになった」という記述や，「予め分類分けされており、興味のある資料の集まりが直感的に見てわかりやすいと思った」という記述があった．そのため，ツリーマップによる

分類の視覚化は有効であると考えられる．しかし，質問 5 の回答から，分類に対して付与した分類名が一部妥当ではないことがわかった．そのため，現状より細かい粒度で分類し，より類似度の高い歴史資料から成る分類を作成することでラベリングの精度を向上させる必要があると考えられる．また，自由記述に「見つけた歴史資料の中で興味深いワードを見つけて以降はデジタル資料館のキーワード検索が便利だと感じた」という記述や「キーワード検索機能が無いため，もう少し分類を絞りたいときに不便だと感じた」という記述があった．そのため，興味のある歴史資料を発見した後，効率よく興味のある歴史資料を発見するためにはキーワード検索機能が必要であると考えられる．

6. まとめ

本稿では，歴史資料の探索型検索を容易にするための歴史資料探索支援システムを提案した．提案システムの有用性を調べるため，提案システムと既存システムであるデジタル資料館における，興味のある歴史資料の発見の容易性の比較を行った．また，システムのユーザビリティの評価を行った．その結果，提案システムは興味のある歴史資料の発見が容易であることがわかった．

今後は，階層的クラスタリングで生成したクラスタを切断する高さを調整し，類似度の高い歴史資料から成る分類を作成することでラベリングの精度向上を図る．また，キーワード検索機能の追加等システムの改善を行う．システムの改善後，デジタルアーカイブの利用者を被験者とした実験を行う予定である．実験では，システム利用ユーザに対する，1 件以上資料をお気に入り登録したユーザの割合を算出することにより評価を行う予定である．

参考文献

[1] White, R. and Roth, R. : Exploratory Search : Beyond the Query-Response Paradigm, Morgan and Claypool Publishers (2009).
[2] 大河原一輝，平野廣美，益子宗，星野准一，ショッピングモール型 EC サイトのための店舗情報視覚化システム，情報処理学会論文誌，Vol.56，No.3，pp.847-855 (2015).
[3] Benjamin B., Bederson. : PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps, User Interface Software and Technology, pp.71-80 (2001).
[4] Bederson B., Schneiderman B. : Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies, ACM Transactions on Graphics, Vol.21, No.4, pp.833-854 (2002).
[5] 伊藤貴之，山口裕美，小山田耕二，長方形の入れ子構造による階層型データ視覚化手法の計算時間および画面占有面積の改善，可視化情報学会論文集，Vol.26，No.6，pp.51-61 (2006).
[6] 五味愛，宮崎麗子，伊藤貴之，Li, J.，CAT: 大量画像の一覧可視化と詳細制御のための GUI，画像電子学会誌，Vol.37，No.4，pp.436443 (2008).
[7] Shneiderman, Ben. and Catherine Plaisant. : Treemaps for space-constrained visualization of hierarchies (1998).
[8] Bruls Mark., Kees Huizing., and Jarke J. Van Wijk. : Squarified treemaps, VisSym (2000).
[9] Jul, S., and Furnas, G.W.: Critical Zones in Desert Fog: Aids to Multiscale Navigation. In Proc. of User Interface and Software Technology (UIST 98) ACM Press, pp.97-106 (1998).
[10] 函館市中央図書館デジタル資料館，http://archives.c.fun.ac.jp.
[11] 出口貴也，中原裕成，高橋正輝，奥野拓，川嶋稔夫，地域の記録と市民の記憶を共有するディジタルアーカイブ CMS，第 84 回デジタルドキュメント研究会 (2011).
[12] Mikolov, T., Sutskever, I., Chen, K., et al. : Distributed representations of words and phrases and their compositionality. In Proc. NIPS, pp.3111-3119 (2013).
[13] Treeratpituk, P. and Callan, J. : Automatically Labeling Hierarchical Clusters. In Proc. of the Sixth National Conference on Digital Govern-ment Research, pp.161-176 (2006).
[14] 水本篤，竹内理，効果量と検定力分析入門―統計検定を正しく使うために―，外国語教育メディア学会関西支部メソドロジー研究部会 2010 年度報告論集，pp.47-73 (2010).