1-509-35
2017/3/2

データビジュアライゼーションを用いた歴史資料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課題がある．本稿では，階層的クラスタリングにより歴史資料を概念関係を持つ木構造となるように分類し，視覚化することにより歴史資料の発見を支援するシステムを提案する．提案システムではユーザが操作可能な形式である Sunburst と呼ばれる視覚化手法，年表を取り入れることで試行錯誤しながら歴史資料を探すことを支援する．

Building a Historical Records Search Support System by Using Data Visualization

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 exploratory search while grasping the whole image. Therefore, users can't find historical records that they are interested in. In this paper, we propose an exploratory search support system that classifies historical records to be a tree structure with conceptual relation and visualizes classification result. In the proposed system, we support a user to find historical records by using visualization called sunburst and chronology.

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は，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課題がある．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．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，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課題がある．

そこで本研究では，歴史資料の内容を大まかに表す分類から，詳細に表す分類へ，全体像を掴みながら段階的に辿ることを可能にするために，歴史資料を木構造となるように分類する．そして分類の結果をユーザが操作可能な形式で視覚化することで，試行錯誤しながら歴史資料探すことを支援する．これによって，興味のある歴史資料の発見を促すシステムの構築を目指す．

2. 情報探索の支援

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ニーズを明確化する．また探索型検索では，ユー

[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

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．

情報探索を支援する研究として，大河原らは EC サイト上の店舗を特徴別にカテゴリ化し，円形ツリーマップを用いて視覚化することで，EC サイト上の店舗の回遊ができるシステムの構築を行っている [2]．この研究では，EC サイト上の各店舗から商品価格情報，商品カテゴリ情報，レビュア年齢情報，商品テイスト情報を店舗の特徴として抽出している．そして，円形ツリーマップを 3 階層に分けて店舗情報の視覚化を行っている (図 1)．1 階層目には，「レディスファッション」，「メンズファッション」などの取扱商品カテゴリを表示している．2 階層目には，「ナチュラル系」，「エレガント系」などの商品テイストを表示している．また，各店舗のレビュア年齢情報をドーナッツチャートを利用して視覚化している．3 階層目には，「店舗名」，「商品数」などの店舗情報を表示している．この階層では，取扱商品の数をバブルチャートの大きさに対応させている．また，最小商品価格と最大商品価格の平均値をバブルチャートの色に対応させている．実験から，円形ツリーマップを用いて店舗情報を視覚的に表現することは，店舗情報の把握に役に立っているという結果が得られている．また，情報探索が容易だという結果が得られている．そこで，本研究においても円形ツリーマップのようなデータビジュアライゼーションを用いることで，ユーザの探索型検索を支援するシステムを構築する．

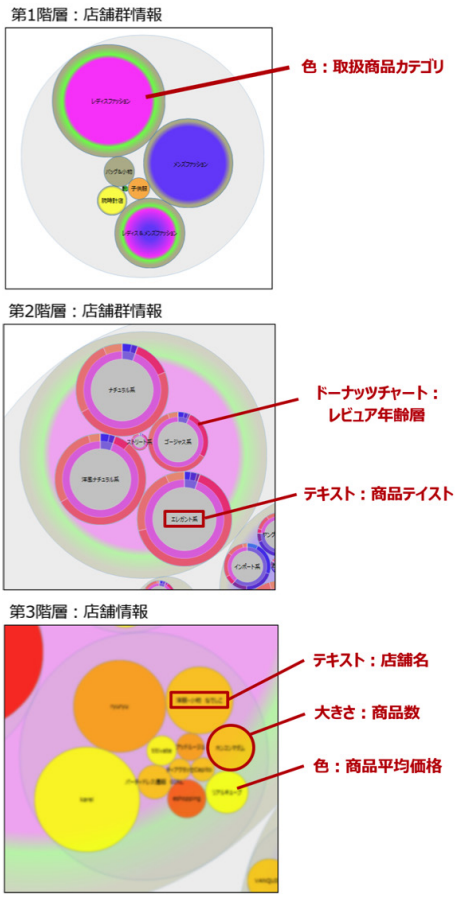

図 1　店舗情報の視覚化 (参考文献 [2] より転載)

Fig. 1　Visualization of store information (Reprinted from the document [2]).

3. 歴史資料探索支援システム

本稿では，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提案する．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課題がある．全体像を容易に把握することができるような仕組みがあれば，探索型検索が容易になると考えられる．また，探索型検索は，繰り返し検索を行うことでニーズを明確にする．そのため，歴史資料の内容を大まかに表す分類から，詳細に表す分類へ，段階的な探索を可能にすることが有効であると考えられる．そこで，歴史資料を概念関係を持つ木構造となるように分類する．そして，ユーザが試行錯誤しながら歴史資料を探すことを可能にするために，分類の結果を対話的に操作可能な形式で視覚化する．

3.1 Sunburst による視覚化

分類の結果の視覚化手法は，木構造の視覚化に適した手法が望ましい．木構造の視覚化に適した手法には，ツリーマップや階層型円グラフがある．ツリーマップは歴史資料の数が多い場合，一度に多くの分類を表示することになり視認性が低下すると考えられる．そのため本研究では，各分類の割合の把握が容易であり，2 階層ずつ表示するため視認性が高い Sunburst を利用する．

Sunburst は，親ノードを中央の領域に表示し，子ノードを親ノードの外側の領域に表示し，孫ノードを子ノードの外側の領域に表示する視覚化手法である (図 2)．Sunburst の子ノードを選択すると，中央の領域に子ノード，内側の領域に孫ノード，外側の領域に孫ノードの子ノードが再描画される (図 3)．

Sunburst を利用して各分類の割合を視覚化することにより，「多くの割合を占めている分類から見る」というような探し方が可能になる．これにより，公開されている歴史資料の全体像を直感的に把握することが容易になると考えられる．

3.2 年表による視覚化

年代が近い歴史資料同士は，表現技法や色使いなどの特徴が類似していると考えられる．そのためユーザは，ある歴史資料に興味を持った場合，同じ年代の歴史資料に関しても興味を持つ可能性が高いと考えられる．また，年代の視覚化に適した手法で視覚化することにより，探索型検索が容易になると考えられる．そこで本研究では，年表形式

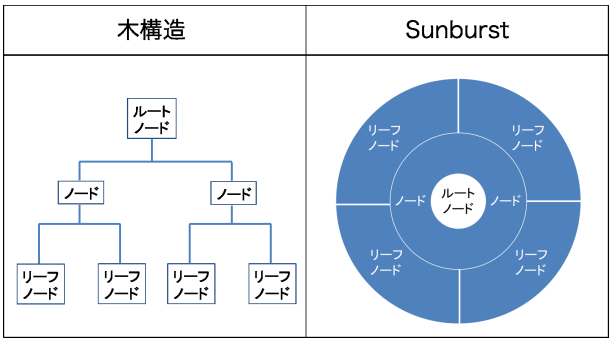

図 2　Sunburst の構造の例
Fig. 2　An example of sunburst structure.

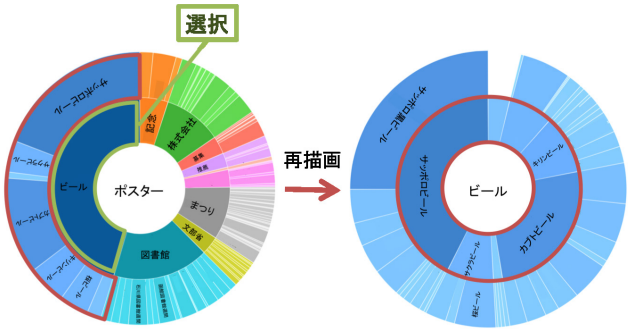

図 3　Sunburst の動作の例
Fig. 3　An example of sunburst.

で視覚化を行う．

システムでは，Sunburst で選択された分類に属する歴史資料を年表形式で年代順に視覚化する (図 4)．年表は，横スクロールすることで操作することができ，年表上の歴史資料を選択することで選択した歴史資料の高精細画像とメタデータを閲覧することができる．ある分類に属する歴史資料が多数存在する場合，多くの縦スクロール操作が必要になる．そのため，ユーザビリティが低下すると考えられる．そこでシステムでは，年表の表示範囲はユーザが操作可能な形式とし，年表に表示されている年代に属する歴史資料のサムネイル画像を表示する．例えば，年表に表示されている年代が 1940 年から 2000 年の場合，表示するサムネイル画像は，1940 年から 2000 年の間に出版された歴史資料とする．

3.3　システム動作

図 5 にシステムの動作を示す．本システムは初期画面に，木構造となるように歴史資料を分類した結果を視覚化した階層型円グラフを表示する．階層型円グラフの各領域には，歴史資料の内容を大まかに表す分類を表示している．初期画面で「札幌」などの内側の分類を選択することで，選択した分類に属する歴史資料のサムネイル画像が表示される．この際に階層型円グラフには，選択されたノードの子ノードである一つ下の階層が再描画される．また，選択された分類に属する歴史資料を年表形式で年代順に視覚化する．

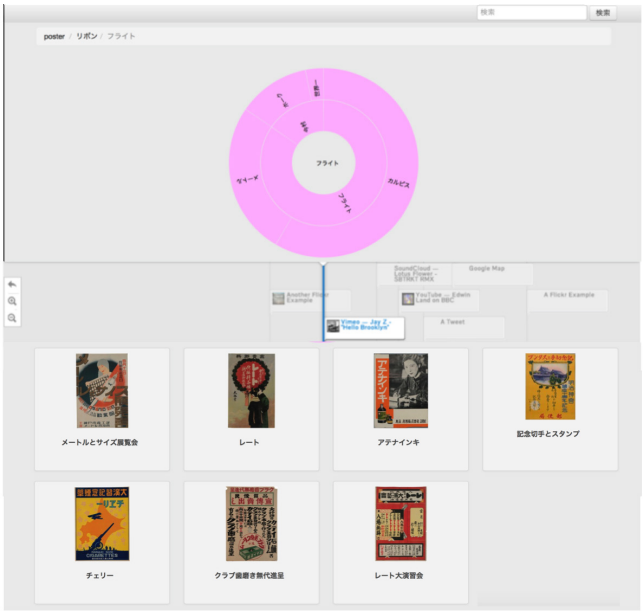

図 4　年表による視覚化のイメージ
Fig. 4　An image of visualization by chronology.

階層型円グラフの中央の領域には，親ノードである一つ上の階層の分類を表示する．階層型円グラフの中央の領域を選択すると，親ノードである一つ上の階層が再描画される．歴史資料のサムネイル画像を選択した際に，高解像度画像，メタデータである目録を表示する．

4.　歴史資料の分類

4.1　木構造の作成

本研究では函館中央図書館デジタル資料館 [3](以下，デジタル資料館) の歴史資料を対象とする．デジタル資料館は歴史資料を死蔵させないことを念頭に置き，目録の完成度が低い資料であっても公開している [4]．ポスターカテゴリの歴史資料の内容説明は，資料に記載されている文字情報のものが多い．しかし，内容説明の項目が空欄の資料が多く存在する．また，デジタル資料館のポスターカテゴリの歴史資料には，ビールやタバコなど特定のテーマに関する資料が多く公開されており，偏りがあるという特徴がある．本研究では歴史資料を以下のように構造化する．資料タイトルが「サッポロビール」である歴史資料の親ノードを，「ビール」などの上位語のキーワードとする．そして，そのキーワードの親ノードを「飲料」などの上位語のキーワードとする．「飲料」などのキーワードの親ノードを上位語のキーワードとすることをルートノードまで繰り返し，多階層の木構造とする．この際にルートノードは「ポスター」などの資料カテゴリとする．

木構造の作成には階層的クラスタリングを利用する．デジタル資料館のビールなどの特定のテーマに関する資料が多いという特徴は，資料タイトルに表れている．そのため，

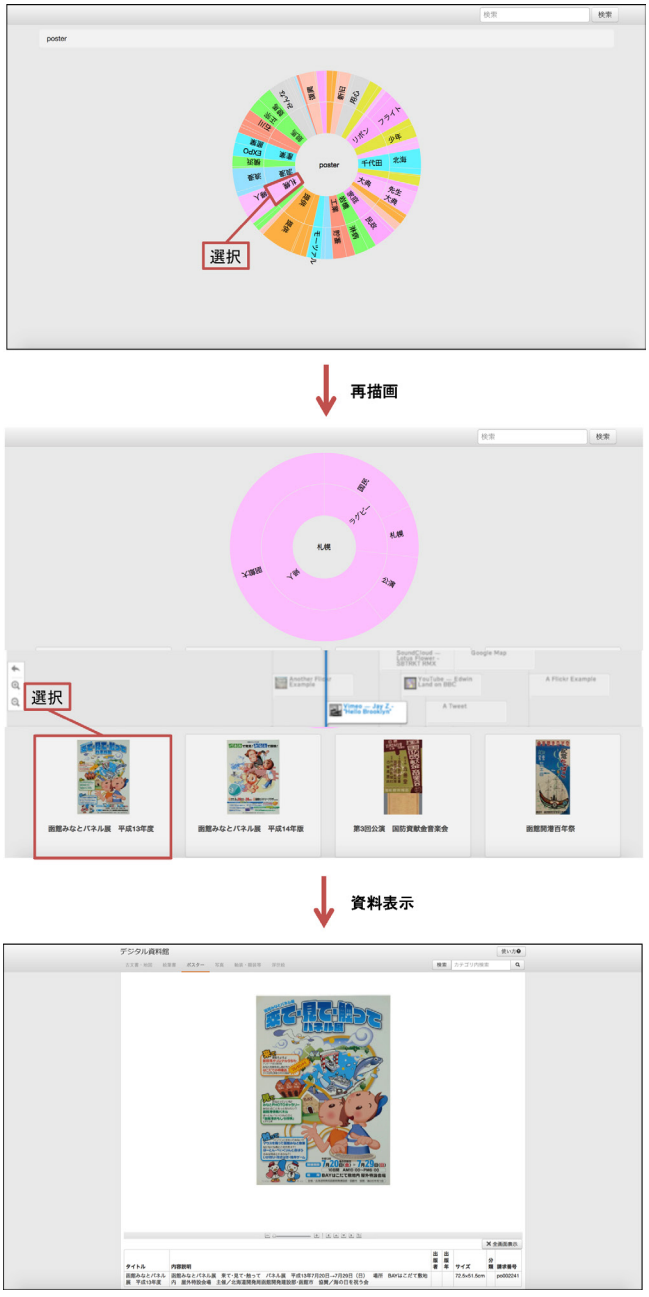

図 5　システム動作
Fig. 5　The system operation.

資料タイトルを利用することで網羅的に歴史資料を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，次の手順で階層的クラスタリングを行う．

(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) 一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 [5] 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．例えば，デジタル資料館には，古文書，絵葉書，ポスター，写真，軸装・額装等，浮世絵の資料カテゴリがある。ポスターカテゴリに含まれる資料は，ビールに関する資料やタバコに関する資料，祭りに関する資料など，分野は多岐にわたる．そのため word2vec の学習データは，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では，分類精度が高い Ward 法を利用する．

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内容を表すラベルを付与する．例えば，「キリンビール」，「カブトビール」，「サッポロビール」から成るクラスタがあった場合，「ビール」をラベルとして付与することを想定している．

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 [6]．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．本研究では，デジタル資料館の歴史資料を対象としている．資料タイトルや内容説明を文書とし，TF-IDF 値を計算するという方法が考えられる．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内容を表すラベルを付与する．

(1) クラスタに属する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単語をラベルとする．

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. まとめ

本稿ではデータビジュアライゼーションを用いて歴史資料の探索型検索を支援するシステムを提案した．提案システムでは，木構造となるように分類した結果を Sunburst によって視覚化する．また，年表形式で歴史資料を年代順で視覚化する．今後は提案システムを用いて評価実験を行い，有効性の評価を行う．

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] 函館市中央図書館デジタル資料館，http://archives.c.fun.ac.jp.
[4] 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)．
[5] Mikolov, T., Sutskever, I., Chen, K., et al. : Distributed representations of words and phrases and their compositionality. In Proc. NIPS, pp.3111-3119 (2013).
[6] 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).