HistoryPaper:ユーザー個人のブラウザ履歴を用いた毎日の可視化

松枝 知香[a]　　伊藤 貴之[b]
お茶の水女子大学大学院
a) *coco@itolab.is.ocha.ac.jp* b) *itot@itolab.is.ocha.ac.jp*

概要
インターネットを毎日利用する人の閲覧履歴を要約することは，その人の行動や知識の要約につながると考えられる．本報告では，1 日の閲覧履歴の中で特に重要であると判断したいくつかの Web ページの集合を抽出し，それらを新聞のようにレイアウトすることで，ユーザーの毎日の行動や獲得知識を要約表示するシステムを提案する．

キーワード　ブラウザ履歴，可視化

1　はじめに

インターネットの普及に伴い Web ブラウザは技術的に進化したが，その閲覧履歴の表示方法はあまり変化していない．ブラウザの閲覧履歴からは，ユーザー自身の行動や，ユーザーが獲得した知識を知ることができる．しかし，従来のリスト表示による閲覧履歴からそれらを知ることは難しい．本報告では，1 日の履歴を要約する Web ページを一覧表示するシステムを提案する．本手法では，1 日の履歴の中から重要と考えられる Web ページ群を抽出し，我々が毎日読んでいる新聞のようにそれらをレイアウトすることで，ユーザの 1 日を新聞のように表現する．「新聞のようなレイアウト」を実現するために本手法では，Web ページの内容の一部を長方形領域に描画し，その長方形領域の集合を Web ブラウザのウィンドウに充填配置する．

2　提案手法

2.1　代表 Web ページの選出

本手法ではまず，1 日の履歴を構成する各 Web ページの重要度を計算し，その結果から代表 Web ページ群を選出する．ここでは「特定の内容に偏ることなく，多様な内容の Web ページを少しずつ選出する」のが理想的な代表 Web ページ抽出であると定義する．そのために本手法では，まず Web ページ群を内容でクラスタリングし，続いて各クラスタから代表ページを選出する．

2.1.1　本文による Web ページのクラスタリング

Web ページを各々の内容に基づいて以下の手順でクラスタリングする．

1. Web ページのコンテンツ内容を Bag-of-Words 表現に変換する．
2. 潜在的意味解析 (LSA) を用いて Bag-of-Words を次元削減する．
3. 潜在的意味空間に位置する各 Web ページを K-means 法でクラスタリングする．

2.1.2　Web ページの重要度計算

続いて 2.1.1 節の方法で生成した各クラスタについて，以下の変数を用いて Web ページの重要度を計算する．

- t: Web ページの滞在時間
- m: Web ページが実際に検索したキーワードを含んでいる数
- p: アクセスの貴重度 (1 日のアクセス数を最近 1ヶ月のページへのアクセス数 で割った商と定義する)

現時点では記事の重要度を以下の式で定義している．ここで k は経験的に決定される定数である．

$$priority = p(t + km) \tag{1}$$

この値が最大である Web ページを，そのクラスタの代表として選出する．ただし，既に同一ドメインの別の Web ページが別のクラスタから選出されている場合には，次に値が大きい Web ページをクラスタ代表とする．

2.1.3　クラスタの重要度計算

クラスタが含む Web ページ数と，クラスタ代表 Web ページの重要度から，各クラスタの重要を計算する．

2.2　配置決定アルゴリズム

本手法では 2.1 節の手法で選出した 6〜20 程度の Web ページ群を，Web ブラウザの一画面に Web 新聞[1]のように配置する．我々は Web 新聞を観察した結果として，良好な記事配置を実現する基準を以下のように定義する．

- 文字を書くスペースが横長である
- 同じサイズのスペースは隣接している

以下，これらの条件を出来るだけ満たすようなアルゴリズムを提案する．

2.2.1　Web ページの配置領域の理想的なアスペクト比の定義

本手法では，Web ページ配置の理想的なアスペクト比を図 1 の通り定義する．選出された Web ページ群を

[1]例: The New York Times http://www.nytimes.com/

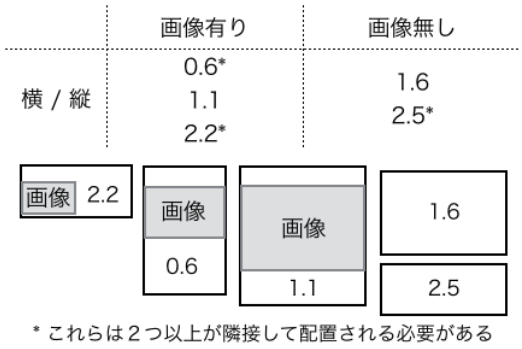

図 1 Web ページの配置領域の理想的なアスペクト比

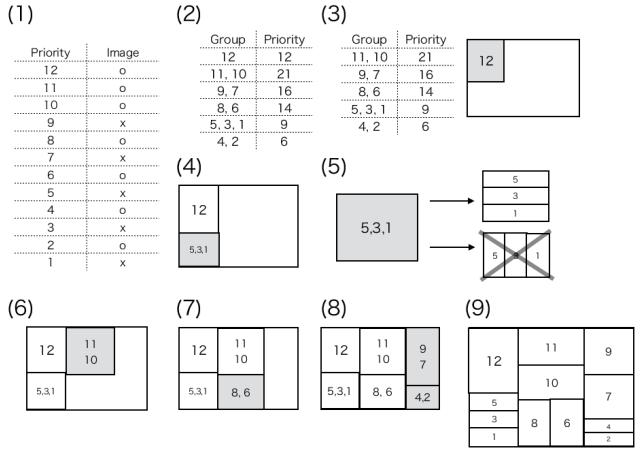

図 2 配置アルゴリズムを適用した例

参考画像の有無でグループ分けするのは，画像を配置する場合としない場合にテキスト表示部分のアスペクト比が変わるため[2]である．

2.2.2 Web ページデータのデータ構造

2.1 節の手法で選ばれた Web ページから，以下で構成されるデータ構造を形成する．

- p: Web ページの重要度 (その Web ページが所属するクラスタの重要度)
- i_s: 理想的な面積
- i_a: 理想的なアスペクト比の集合

2.2.3 Web ページの配置

2.1 節の手法で選ばれた Web ページ群を，Web ブラウザのウィンドウに配置するアルゴリズムを説明する．この配置アルゴリズムによる配置の例を図 2 に示す．

1. トップ Web ページの選出．重要度が最大である Web ページを選出し，これをウィンドウの左上端に配置する．
2. Web ページグループ (図 2(2)) の作成．トップ以外の Web ページ群を画像の有無で 2 分し，その各々について重要度順に Web ページ 2 個または 3 個ずつのグループを作る．以下，グループの集合を G と表す．
3. Web ページグループまたは Web ページグループ群の配置．渡された Web ページグループ，もしくは Web ページグループ群 G の数を n_G，これらを配置する領域を S として，以下の処理を適用する．

- $n_G \geq 4$ の場合 (図 2(3)(6))
  - (i) トップ Web ページが未配置である場合にはトップ Web ページを選び，さもなければ G の中で重要度が最大であるグループを選び，S の左上部に配置する．
  - (ii) 配置した長方形の下にできたスペースを新たに S とし，1 個以上の G を組み合わせた Web ページグループ群を構成する Web ページの i_s の合計が S の面積に 1 番近いグループを選び，それを新たに G として 3. に戻る．
- $n_G = 1$ の場合 (図 2(4))
  - (i) 既に配置した Web ページグループの長方形領域を伸縮してアスペクト比を調整することで，配置領域 S の面積を G 理想面積 i_a の和に近づける．
  - (ii) S に G を配置する．
  - (iii) G を構成する Web ページ群の各々を G として，3. を反復する．
- $n_G = 2$ または 3 の場合 (図 2(5)(8))
  - (i) 式 (2) の値が最も小さくなる方法で，残りのスペースを分割する．ここで分割方法は「縦方向」「横方向」「G が Web ページグループ群かつグループ数が 3 の場合は T 字分割」のいずれかとする．

$$\sum_{i=0}^{L} |\frac{Ss}{n_G} - i_{si}| + min(|\frac{Sw}{Sh} - i_{ai}|) \qquad (2)$$

ただし Ss は S の面積，Sw は S の横幅，Sh は S の縦幅とする．

  - (ii) G を構成する Web ページグループの各々を G として，3. を反復する．

3 今後の課題

今後の課題として，配置アルゴリズムの改良と，記事選出アルゴリズムの精度の検証と改良に取り組みたい．

[2]図 1 の値は，画像のアスペクト比によって変化する．今回は画像を黄金比 (1:1.6) でトリミングした場合の値を表記している．