HistoryPaper: ユーザー個人のブラウザ履歴からの代表ページ選択とマガジンスタイルレイアウト

理学専攻　情報科学コース 松枝知香 (指導教員：伊藤貴之)

1　概要

ブラウザの閲覧履歴は，ユーザ自身の行動や，獲得した知識についての情報を含んでいる．そのため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．しかし，現在のブラウザに実装されている閲覧履歴の表示方法だけでは，そのような情報を有用活用することは簡単ではない．本研究では，1 日の閲覧履歴の中で特に重要であると判断した Web ページ群を抽出し，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．提案手法ではまず，閲覧履歴を構成する Web ウェブ群を文書内容でクラスタリングし，続いて検索キーワード，アクセス貴重度から定義される重要度を算出して各クラスタから代表 Web ページを選出する．続いて，最近の Web サイトに多く採用されているマガジンスタイルを模倣するレイアウトアルゴリズムによって，代表 Web ページ群を一画面に配置して一覧表示する．

2　マガジンスタイル

マガジンスタイルは，その名の通り雑誌のようなレイアウトのことである．様々な大きさの四角形を多数画面上に敷き詰め，その中に文字や画像を当てはめてコンテンツを表示させるこのスタイルは，普段から雑誌を読むユーザにとっても見慣れたものであり，多くのコンテンツが詰まっていてもコンテンツの内容を理解しやすい．

マガジンスタイルは俗称であり，それがどのようなレイアウトであるかの明確な定義は我々の知る限り見当たらない．そのため自身の観察結果にもとづいて，本研究におけるマガジンスタイルの生成方針を以下のように定義した．

定義 (1) カラムレイアウト

定義 (2) 重要な記事ほどスペースが大きい

定義 (3) 各長方形領域のアスペクト比は表 1 に準ずる

定義 (4) 同じ大きさの長方形領域が複数隣り合わさる

3　提案手法

3.1　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，その結果から履歴を要約するための代表 Web ページ群を選出する．ここで文書群の要約の目的で代表文書を選出する手法として，tf-idf 法によって測った重要度の高い文書を重複を避けて抽出することで，満足度の高い結果がでる傾向があることが知られている [1]．本手法ではこの知見を利用して，閲覧履歴を構成する Web ページ群に対して文書内容の類似度に基づくクラスタリングを適用し，各クラスタの中から最も重要度の高いものを選ぶことで，履歴の要約となる Web ページ群を構成する．以下に代表 Web ページ群の選出手法を示す．

表 1: 各長方形領域の理想的なアスペクト比

表示タイプ	横 / 縦 (画像位置)
画像有	(上)0.9，1.6 (左)3.0，3.8
画像無	1.0，3.8，5.0

まずはじめに 1 日の履歴にある全ての Web ページの内容を対象として，以下の文書クラスタリング手法を用いて n 個のクラスタに分類する．

1. Web ページの内容を Bag-of-Words 表現に変換
2. 潜在的意味解析で Bag-of-Words を次元削減
3. 潜在的意味空間に位置する各 Web ページを最短距離法を用いてクラスタリング

続いて各クラスタについて，検索キーワードのうち当該 Web ページに載っているキーワードの延べ数とアクセスの貴重度を考慮し各クラスタの代表ページを決定する．また，クラスタに含まれる Web ページの延べ数をクラスタの重要度とし，代表ページを配置する際に専有する面積を決定するときに利用する．

3.2　レイアウトアルゴリズム

3.2.1　予備実験

マガジンスタイルのレイアウトアルゴリズムの開発に先立ち，我々は Squalified Treemap[2] による配置を試みた．その結果，2 章のマガジンスタイルの定義のうち，Treemap は定義 (1)(2) を満たすが，定義 (3)(4) を満たすことが難しいことがわかった．

3.2.2　Web ページ群のデータ構造

本節でのアルゴリズムの説明に先立ち，以下のようにクラスタの集合を定義する．

- R: クラスタの集合 (3.1 節に示した処理で生成する)
- G: 1 個以上のクラスタで構成されるクラスタグループの集合

また，以下の変数を定義する．

- $|S|$: 集合 S に含まれる集合の数
- $area_S$: 集合 S 全体の面積占有度
- $priority_S$: 集合 S 全体の合計クラスタ重要度
- $ratio_S$: 集合 S 全体のアスペクト比群 (表 1 参照)
- $type_R$: クラスタ画像の有無
- W, H: 配置に用いる画面の横幅，縦幅 (定数)
- W_{min}， H_{min}: 各クラスタが配置される長方形領域の最小横幅，縦幅 (定数)

3.2.3 クラスタのグループ化

本手法では以下の式を満たす場合に R_i と R_j の 2 クラスタを同一グループに所属させる (図 1(1))．ただし，R_i が既に他のグループに所属している場合，$type_{R_i}$ と $type_{R_j}$ が違う場合には，この処理を適用しない．

$$priority_{Rj} \leq \lceil (priority_{Ri} \times 1.2) \rceil \qquad (1)$$

(1)

集合	クラスタ重要度	画像
R_1	45	o
R_2	32	o
R_3	30	o
R_4	20	
R_5	11	
R_6	10	
R_7	5	
R_8	2	
R_9	1	

(2)

集合	クラスタ重要度		面積占有度		画像
G_1	45		27		o
G_2	32	30	20	18	o
G_3	20		12		
G_4	11	10	7	6	
G_5	5		3		
G_6	2	1	2	1	

図 1: 面積占有度の算出

3.2.4 各クラスタの面積占有度

各クラスタの重要度をデフォルメした値を，配置する長方形領域の面積占有度として算出する．各クラスタの重要度をそのまま面積占有度としない理由は，3.2.1 節であげた問題点を解決するためである．

$$area_{R_i} = \frac{priority_{R_i}}{priority_R} \frac{WH}{W_{min}H_{min}} \qquad (2)$$

3.2.5 配置アルゴリズム

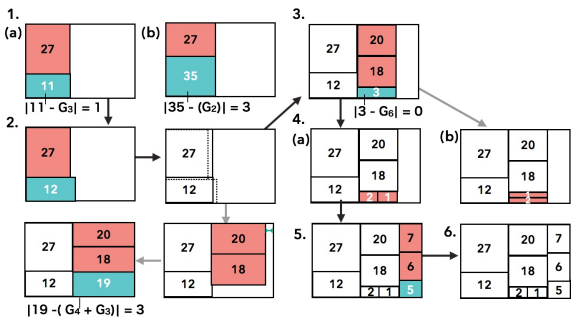

図 2: レイアウトアルゴリズム

1. 面積占有度の平均が最大であるクラスタグループを抜き出し，表 1 で取りうるすべての比率でスペースの左上に仮配置を行う (図 2 1.(a)(b))．左下にできたスペースにできるだけ適応する $|G|$ の要素の組み合わせを選び，左下の面積占有度と $|G|$ の要素の組み合わせの差が 1 番小さい組み合わせを選ぶ (図 2 1.(a)(b) 3.)．

2. $|G|$ が 3 以上の場合 1. を繰り返す．$|G|$ が 2 以下の場合，同じ G の要素は縦もしくは横に等分割する (図 2 4.)．G の要素の組み合わせであれば面積占有度の比率に縦もしくは横分割する (図 2 5. 6.)．

4 実行結果

4.1 アルゴリズムの適用

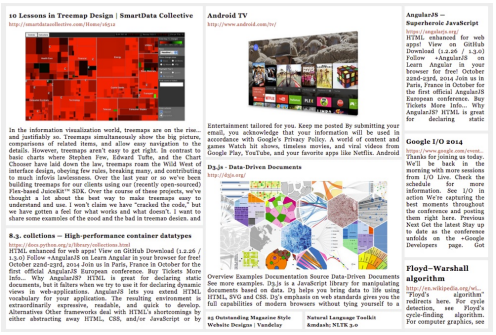

図 3: 配置結果

図 1 のデータを用いて，提案手法を実行した例を図 3 に示す．画面全体を大きく縦に分割し，それをさらに分割して，重要度に応じた面積を割り当てて各ページを表示していることがわかる．また，同じ大きさのページを並べたレイアウトが随所に見られることがわかる．以上により，提案手法が 2 章で論じた定義に近いレイアウトを実現できていることがわかる．

4.2 ユーザテスト

20 代女性 10 人に HistoryPaper を実際に利用してもらい，ユーザアンケートを行った．その結果 9 人のユーザが，HistoryPaper を用いることによって数日前に行ったが既に忘れていた行動を思い出すことがあったと答え，HistoryPaper が 1 日を振り返ることの一助になることがわかった．

参考文献

[1] Genest, P. E., et al. A symbolic summarizer for the update task of tac 2008. In Proceedings of the First Text Analysis Conference, Gaithersburg, Maryland, USA. National Institute of Standards and Technology. (2008)

[2] Bruls, M. et al, Squarified treemaps, Data Visualization 2000, 33-42 (2000).