RA-006

リストページ自動分割問題の最適グラフ分割を用いた解法の提案と評価

Design and Evaluation of a Solution for an Automatic List-Page Decomposition Problem by Using an Optimal Graph Decomposition

志賀 琴枝* 徳山 豪*
Shiga Kotoe Takeshi Tokuyama

1 はじめに

Amazon に代表される商品カタログサイトのように、動的に利用者の要求の結果を表示する HTML ページが作成される Web サイトが広く利用されている。多くの場合、サイトは次の２つに分類できる WEB ページの集合で構成されている。１つは多くの商品の情報をリストやテーブル形式で持っており、この構造を含むページを”リストページ”と呼ぶことにする。もうひとつは、ひとつの商品の付加情報を持っており、この構造を含むページを”ディテールページ”と呼ぶことにする。リストページから抽出した商品情報リストに対し、これを各商品に対応した部分リストに分割する問題は、Web のリストページの自動分割問題と呼ばれており、膨大な商品情報をデータベース化する目的で研究されている。
既存研究 [1] ではこの問題を制約充足問題として定式化し、数理計画法のパッケージ等を用いて解を求めているが、これは NP 完全問題への定式化であり、理論的には多項式時間解法ではない。一方、本論文で示すように、グラフの最適連続分割問題に定式化すると、多項式時間で自動分割を行うことができるが、良い分割を得るためにはグラフの重み付けを適切に行う必要がある。　本論文では様々な重み付け手法について考察を行う。

2 問題の定義

Web のリストページの自動分割問題のモデルを以下のように定義する。$L = l_1, ..., l_n$ を、ひとつのリストページから抽出された特徴語列 とする。リストページに対応する m 個の商品各々のディテールページの集合を $D = \{D_1, ..., D_m\}$ として、r_j をディテールページ D_j から抽出された特徴語列であるとする。これをディテールページに対応したレコードと呼ぶ。r_j は重複を許し、特徴語の出現順は D_j と同一とする。集合 $\mathbf{r} = \{r_1, ..., r_m\}$ は各ディテールページのレコードからなる集合であり、レファレンス集合と呼ぶ。
L の部分リストが与えられたとき、その語列とディテールページのレコードには関連度が与えられる。リスト分割問題では、L の部分列 $\mathbf{I} = I_1, ..., I_k$ への分割を考え、$\mathbf{I}$ と $\mathbf{r}$ の間の関連度と、付随したいくつかの制約条件に依存したマッチングを行い、最適なマッチングを与える分割を求める問題である。ここで $I_j = [a_j, b_j), a_1 = l_1, b_k = l_{n+1}, a_{j+1} = b_j$ とし、すなわち I_j はリストの部分区間であり、左から順にインデックスされているとする。

$$l_1, l_2, l_3, l_4, l_5, l_6, l_7, l_8, l_9, l_{10}, l_{11}, ...$$
$$\downarrow$$
$$\underbrace{\{l_1, l_2, l_3\}}_{I_1}, \underbrace{\{l_4, l_5, l_6\}}_{I_2}, \underbrace{\{l_7, l_8, l_9, l_{10}\}}_{I_3}, \underbrace{\{l_{11}, ...\}}_{I_4}, ...$$

図 1: Web のリストページの自動分割問題のイメージ

3 既存研究 [1] の説明

語句と特徴語集合の間の関係を表現する二値行列 $X = (x_{ij})$ を以下のように定義する。

$$x_{ij} = \begin{cases} 1 & (l_i \in r_j) \\ 0 & (else) \end{cases} \quad (1)$$

*東北大学大学院情報科学研究科システム情報科学専攻

既存研究 [1] では、x_{ij} を用いて制約充足問題（ＣＳＰ）に変換して解く。レコードと特徴語の対応において、l_i が r_j に複数個存在するときは、その存在位置を表すパラメータを $pos_j(l_i)$ と置き、語句 l_j を r_j の t 番目の語に対応させるとき、$pos_j(l_i) = t$ と定義する。存在しない場合には値を返さないものとする。制約条件は $pos_j(l_i)$ を用いて、以下の３つのように表現する。

連続条件：　一つのレコードに割り当てられるのは、リストページの連続した区間である

$x_{nj} = 0$ となるような n が $k < n < i$ で存在するとき、$x_{ij} + x_{kj} \leq 1$

単独条件：　各特徴語 l_i はただ１つのレコード r_j に割り当てられる

$\sum_j x_{ij} \leq 1$

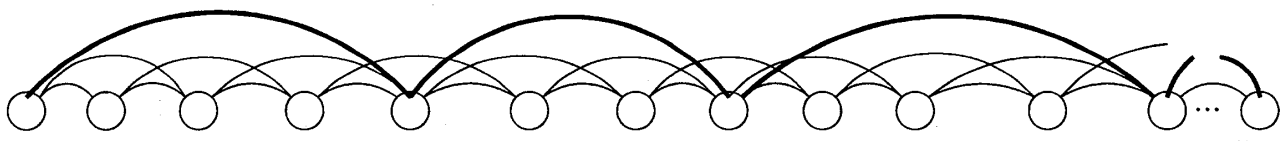
位置条件：　リストで異なる位置にある同一語句 l_i,l_k を含む区間がレコード r_j に存在するのなら、それらはレコードの異なるポジションに対応する

$pos_j(l_i) = pos_j(l_k)$ のとき、$x_{ij} + x_{kj} = 1$

残念なことに、上記の制約充足問題を多項式時間で解くアルゴリズムは知られておらず、論文 [1] では数理計画法のパッケージを利用して解を求めている。また [1] では、学習モデルを利用した問題解決のアプローチも紹介されている。

4 提案手法

我々が提案する手法では、この問題をグラフの最適連続分割問題に変換してリスト分割を行い、多項式時間で解を与える。グラフの最適連続分割問題の定義は以下のようなものである。$H = (V, E)$ を $(n+1)$ 個の頂点集合 $V = \{1, 2, \ldots, n+1\}$ を持ち辺集合 $E = \{(i, i+p) : 1 \leq i \leq n, 1 \leq p \leq s\}$ を持つ重み付きグラフとする。ここで、s はリスト分割によって得られるレコードに含まれる最大の特徴語数であり、辺の和は $O(ns)$ である。頂点を数直線上の $n+1$ 個の点として１番目から $n+1$ 番目まで並べると、辺は自然に L の区間に対応する。各々の辺 $e \in E$ に対して、e に対応する L の区間 $I(e)$ ともっともよくマッチするレコード r_j との適合度を利用した重み $w(e)$ を与える。すると、L の区間への分割は、H 上での１から $n+1$ へのパスに対応し、分割の品質は、パス上の辺の重みに対応する。従って、１番目から $n+1$ 番目までの最短路が最適なリスト分割に対応する。辺の重みが非負で、点が直線的に配置されるグラフの最短路は下記の動的計画法を利用して、$O(ns)$ で解ける [2]。
i 番目の点までの最短路を D_i, かつ $D_1 = 0$ とおくと $D_{i+1} = \min_{0 \leq k \leq s} \{D_{i-k} + edge[i-k][k]\}$ である。これを利用して、帰納的に $n+1$ 番目の点までの最短路を求めることができる。　しかしながら、実用的に良い自動分割アルゴリズムを与えるためには、独立条件と位置条件に対応する制約を上記の最短路問題に付加する必要があり、そのために、グラフの重み付けに工夫をする必要がある。



	v_0	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{n+1}
L		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	l_{10}	l_{11}	...
		草枕	夏目漱石	525 円	山椒魚	井伏鱒二	420 円	斜陽	太宰治	420 円	在庫なし	川端康成	...
I		I_1			I_2			I_3				I_4	...
x_{i1}		1	1	1	0	0	0	0	0	0	0	0	...
x_{i2}		0	0	0	1	1	1	0	0	1	0	0	...
x_{i3}		0	0	0	0	0	1	1	1	1	1	0	...
...		...	...	...	...	...	...	...	...	...	...	...	...

表 1: テーブルの自動分割の例

4.1 連続条件

本研究では、二値行列 X の列に対応した列ベクトルの分割エントロピーを利用して辺の重み付けを行うことにより、すこし弱い形の連続条件を与える。直感的には、辺 $e=(i,i+k)$ に対応する辺と r_j の適合度は、X の j 列に対応するベクトル $\mathbf{x_j}$ の i 番目から $i+k$ 番目の部分 $\mathbf{x_j}[i,i+k]$ での 1 の出現頻度が、$\mathbf{x_j}$ 全体での 1 の出現頻度よりも多い場合に、その適合度が高く、辺の重みが小さくなるように設定する。

x 0000111011000010
$n=15$ $n_1=6$
$4-9$: $p_j=6/30$ $p_{1j}=4/6$ $p_{2j}=2/24$

図 2: 01 シークエンスの具体例

ベクトル $\mathbf{x'_j}$ を $\mathbf{x_j}$ の後ろに 0 を n 個付加した数列とする。p_j を $\mathbf{x'_j}$ での 1 の出現確率、p_{1j} を $\mathbf{x_j}[i,i+k]$ での 1 の出現確率、p_{2j} を、$\mathbf{x'_j}$ から $\mathbf{x_j}[i,i+k]$ を除いた部分での 1 の出現確率とする。ここで、ベクトルの後ろに 0 を付加するのは $p_j=1$ になる事を避ける工夫である。また $E(p)=-\{plog_2p+(1-p)log_2(1-p)\}$ とする。このとき、辺 e の重みを

$$w(e)=\min_j\left\{\frac{kE(p_{1j})+\alpha(2n-k)E(p_{2j})}{2nE(p_j)}\right\}$$

とする。α は $0\leq\alpha\leq1$ の定数であり、$E(p_{2j})$ の $w(e)$ への影響を小さくするために導入した。すべてのディテールページについて調べるため、これを算出するのに $O(smn)$ かかる。

4.2 単独条件

単独条件に対応する条件を考察してみる。単独条件が成立するとすると、$\sum_j x_{i_1j}=\sum_j x_{i_2j}=1$ かつ $x_{i_1j}=x_{i_2j}=1$ かつ $i_1=i_2$ ならば l_{i_1} と l_{i_2} は同一クラスタに含まれるべきなので、i_1 から i_2 までの点をひとつのブロックとしてまとめて考え、ブロックの間で分割する点を選んでよい。更に、$\sum_j x_{i_1j}=\sum_j x_{i_2j}=1$ かつ $x_{i_1j_1}=x_{i_2j_2}=1$ かつ $j_1\neq j_2$ ならば l_{i_1} と l_{i_2} は異なるクラスタに含まれるべきなので、i_1 番目の点が始点で、終点が i_2 番目以降の点の辺の重さは非常に大きく実際は最短経路に含まれる可能性は低いと考えてよいため、辺を取り除いた。
また、この条件から s の推定を行う。s は、すべてのクラスタの最大サイズより大きいという条件下で最小化する。$first$ と $last$ をそれぞれすべてのブロックのうち最初の点と最後の点の番号の配列とする。$first[k+1]-last[k-1]+1$ は $k-1$ 番目のブロックの最後から $k+1$ 番目のブロックの最初までに含まれる点の数であり、クラスタにはただ一つのブロックしか含まれないことを考えると、s はこれよりも小さくなければならない。このことから、$s\leq\max_{k\geq1}\{first[k+1]-last[k-1]+1\}$ となる。

4.3 位置条件

この条件が満たされた場合、i_1 と i_2 は異なるクラスタに存在すべきなので i_1 より前の点を始点とし、i_2 番目以降の点を終点に持つ辺の重みを大きくする。しかしながら観察の結果、一つのブロックに同じ商品情報が複数個含まれている場合もあるので、この条件は必ずしも正しいとは言えない。 従って、我々は位置条件をすこし弱めた条件に対応する条件を考え、グラフの重み付けを工夫する。具体的には、連続条件と単独条件で算出した重み $w(e)$ を条件を満たす毎に 2 倍するものとした。結果は次の章で示す。
以上の条件によるグラフの作成も含めて、$O(smn)$ でこの問題を解くことが可能となる。

5 実験結果

実際に WWW 上にある 5 つのサイトで実験して、精度と実行時間について評価を行った。実験に用いたサイトの n,m,s の値については表 2 の通りである。
実際値とは、L を実際に観察して確認したすべてのレコード

site	n	m	s	実際値	s の誤差
Amazon	110	10	17	11	35.29
Q-pot	175	30	8	8	0.00
catchbon	63	10	5	4	20.00
altech	112	20	8	6	25.00
toysrus	102	12	11	8	27.27

表 2: 実験に用いた Web サイトの s の値について

に含まれる最大の特徴語数である。s の誤差は様々だが、実際値よりも必ず大きく取ることができていることが確認できた。

5.1 精度評価

精度の評価は、レコード分割した特徴語を以下の 4 種類に分類して 3 種類の指標 P,R,F を用いて行った。

1. Cor：正しくレコード分割された特徴語の数
2. $InCor$：正しくレコード分割されなかった特徴語の数
3. FN：正しく抽出されなかった特徴語の数

4. FR：特徴語でないのに抽出された文字列の数

$$P = Cor/(Cor + InCor + FP)$$
$$R = Cor/(Cor + FN)$$
$$F = 2PR/(P + R)$$

精度に関しては、α と s の精度による影響を確認するため２種類の実験を行った。
まず、α の値をランダムに変えて実験を行った。$\alpha = 1.0$ と $\alpha = 0.04$ の場合の結果は表 3, 4 のようになった。$\alpha = 0.04$ を選択したのは、これが最も平均的に指標が高かったからである。

site	Cor	InC	FN	FP	P	R	F
Amazon	69	33	8	0	0.68	0.89	0.77
Q-pot	174	1	0	0	0.99	1.00	1.00
catchbon	38	2	23	0	0.95	0.62	0.75
altech	92	19	1	0	0.82	0.99	0.90
toysrus	80	9	18	0	0.95	0.82	0.88
average					0.88	0.87	0.87

表 3: $\alpha = 1.00$ の場合の精度評価

site	Cor	InC	FN	FP	P	R	F
Amazon	91	11	8	0	0.89	0.92	0.91
Q-pot	173	2	0	0	0.99	1.00	0.99
catchbon	38	2	23	0	0.95	0.62	0.75
altech	111	0	1	0	1.00	0.99	1.00
toysrus	79	5	18	0	0.94	0.81	0.87
average					0.95	0.87	0.91

表 4: $\alpha = 0.04$ の場合の精度評価

この実験から、精度は α にかなり依存していることが確認できた。α の値を様々変えて実験してみたところ、すべての入力に対して精度を向上させるような α を見つけることはできなかった。しかし、どの入力に対してもかなり精度の高い結果を返す α が存在することは確認できた。この関係について理論的解析をしていくことが今後の課題である。どちらの場合も

	P	R	F
$\alpha = 1.00$	0.88	0.87	0.87
$\alpha = 0.04$	0.95	0.87	0.91
CSP	0.85	0.84	0.84

表 5: 既存研究との比較

既存の手法よりは精度が向上していることが確認された。このことから、グラフ問題への定式化は既存のＣＳＰによる手法よりも適当であったと考えられる。

次に、s による精度比較をするために、レコードに含まれる特徴語の最大数の実際値を入力してレコード分割をした。この結果は表 6 のようになった。s が実際値に近いとレコード分割もかなり高い精度でできた。この結果から、s の精度を高めることがレコード分割の精度の向上への課題であることが確認された。

ここで、Amazon の実験結果が比較的悪かったことに関して考察する。Amazon のページから抽出した入力データを確認したところ、ディテールページに複数のアイテムの情報が含まれていた。これは多様化する商品カタログサイトには決して珍しくない傾向である。このことにより前提条件であった単独条件による制約が悪影響を受け、精度の低下につながったものと考えられる。ただし、α を調整することによって連続条件による制約からある程度精度の向上を図る可能性も確認できた。この点に関してはこれからの課題である。

site	Cor	InC	FN	FP	P	R	F
Amazon	83	19	8	0	0.81	0.91	0.86
Q-pot	173	2	0	0	0.99	1.00	0.99
catchbon	38	2	23	0	0.95	0.62	0.75
altech	111	0	1	0	1.00	0.99	1.00
toysrus	83	1	18	0	0.99	0.82	0.90
average					0.92	0.87	0.89

表 6: s が実際値の場合の精度評価

5.2 時間評価

5.1 で用いたのと同一のデータを提案手法による手法でレコード分割した場合の実行時間を３０回繰り返して測定し、その平均値を調べた。結果は表 7 のようになった。

site	time [ms]
Amazon	57.53
Q-pot	35.93
catchbon	10.42
altech	82.60
toysrus	36.13

表 7: 実行時間の評価

同じデータを GLPK というソルバー [3] を利用して既存研究である CSP でレコード分割したところ、数分から数十時間かかった。この結果から、計算時間は劇的に短縮できていることが確認された。

6 まとめ

リストページの自動分割問題は、グラフ問題として定式化し直すことにより既存の手法よりも高い精度で解を求められる。このことから、制約充足問題の定式化には問題があったと考えられる。また、グラフ問題に変換することによって多項式時間で解くことができるようになり、実行時間も大幅に短縮できた。この点は提案手法の大きな成果である。ただしこの精度は α に依存し、入力データにより適正値が異なった。α の値によっては更なる精度の向上が期待できるため、この相関関係について理論的に解析していくことが今後の課題である。また、s の精度を高めることも同様の理由から今後の課題として挙げられる。

参考文献

[1] Kristina Lerman, Lise Getoor, Steven Minton, Craig Knoblock Using the Structure of Web Sites for Automatic Segmentation of Tables *Proceedings of the 2004 ACM SIGMOD Conference*119-130

[2] Optimal Sequential partition of graphs, *Journal of ACM* **18** (1971) 34-40.

[3] GLPK **<http://www.gnu.org/software/glpk/glpk.html>**