社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－AL－100 (3)
2005／3／17

Web検索結果におけるクラスタリングアルゴリズムの研究

丸山 謙志[1], 王 冠超[2], 徳山 豪[1]
[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
(maru,tokuyama)@dais.is.tohoku.ac.jp
[2] 日立製作所
gc-wang@itg.hitachi.co.jp

概要 本論文では，ラベル付きクラスタリングによるWWW検索エンジンの閲覧効率の向上手法を提案する．現存のWWW検索結果クラスタリングアルゴリズムでは，クラスタラベルの重複及びラベルに無関係な文書の混在という欠点があり，ユーザにとって文書発見が困難になることがある．提案するアルゴリズムでは，従来法で得られた初期クラスタリングに対し，ネットワークフローを用いて最適なラベル配置を行う事によりラベル集合のクラスタリングを求める．上記の問題点は，ラベルクラスタリングを利用した文書の再クラスタリングにより解消される．また，アルゴリズムの実装を行い，本手法によるWWW検索エンジンのユーザビリティーの向上を示す実験結果を与える．

Research of Clustering for Web Search Results

Kenji Maruyama[1], Guanchao Wang[2], Takeshi Tokuyama[1]
[1] GSIS, Tohoku University, Sendai, Japan.
(maru,tokuyama)@dais.is.tohoku.ac.jp
[2] Hitachi,Ltd.
gc-wang@itg.hitachi.co.jp

Abstract. We propose a method for improving usability of WWW search engines via labeled clustering. Existing algorithms for clustering WWW search results have a couple of defects to prevent users from finding documents they want: Duplication of cluster-labels and existence of irrelevant documents to the label in each cluster. In our algorithm, we generate a network from an initial clustering and apply a network flow algorithm to give an optimal label assignment to have a clustering of the set of labels. By using the label clusters, we modify the document clustering to resolve the above defects. We implemented the algorithm and give experimental results to show that our algorithm improves usability of WWW search engines.

1 はじめに

WWWが登場して以来，Web上に混在する様々なリソースが増大してくるにつれ，これらのリソースの中から有用な情報を効率的に探索する優れたインタフェースの需要性がますます高まってきた．

Web上の情報を探索するのに主としてWebサーチエンジンとWebディレクトリサービスという二つのアプローチがある．Webディレクトリサービスとして代表的なものにYahoo[2]，Open Directory Project[3]があり，これらは人手によってウェブページに情報を添加し，それを基に階層的なカテゴリーに分類する．ユーザはカテゴリー情報を参照に意図するウェブページを探すことができる．一方，Webサーチエンジンはユーザからのキーワード入力を基に関連するウェブページに重み付けし，これらの重みよって自動的にランキングされたウェブページのリストを返す．このような検索エンジンの代表的なものにGoogle[1]がある．現在のWeb検索のデファクトスタンダードはWebサーチエンジンであり，以下本研究はサーチエンジンのみについて考える．

現在主流のサーチエンジンからユーザに出力される検索結果は一般的に一次元リストになっている．しかしながら，[6]によるとユーザの入力が統計的に2，3語と短く，このような短いキーワードに対してサーチエンジンは膨大な検索結果を返す．検索結果における

ランクが近いページ間は必ずしも関連性があるわけではなく，このような膨大な検索結果をなんらかの手法で処理し，ユーザが目的のページを効率的に探索できるようにする必要がある．

膨大な検索結果処理に対する一つのアプローチにクラスタリングがある．似ている Web ページ同士を同じグループに，そして，異なるグループ間の類似度をできるだけ小さくするのがクラスタリングのねらいである．Web 検索結果のクラスタリングは検索結果をあるトピックに基づく自動グループ化と位置づけできる．古典的な文書クラスタリングとは異なり，検索結果クラスタリングは，ユーザのキーワード入力後に前処理に時間をかけることなくリアルタイム処理を行う．また，クラスタリングの対象も全検索結果ではなく一部の検索結果のみである．

検索結果クラスタリングを用いることにより，グループ（クラスタ）に付加されたラベル情報を頼りにユーザはより効率的に検索結果から目的のページまで辿り着くことができるようになる．検索結果クラスタリングの主な既存研究として，Pedersen，Hearst らの Scather/Gather System[4]，STC(Suffix Tree Clustering) による Zamir らの Grouper System[7][8]，潜在的意味インデキシング (LSI) や多言語対応にした Zhang や Dong らの SHOC(Semantic Hierarchical Online Clustering)[5] などがある．

本研究は SHOC から発想を得て，ベクトル空間モデルを基に LSI を実現するための一つの手法である特異値分解 (SVD) を用いて Web 検索結果のクラスタリングを行う．しかしながら，SHOC アルゴリズムでは生成されたクラスタにラベル付けをするときに，複数のクラスタに同一のラベルが与えられるというラベル重複問題が生じることがしばしばあり，ユーザにとって見やすい出力が得られない欠点があった．そこで本研究ではラベルに基づく階層的クラスタリング手法 LBHC を提案する．LBHC では，ラベルに関する排他的クラスタリングを用いて文書の非排他的クラスタリングの性能向上を行う．まずラベル重複問題を最小コストフロー問題に帰着して解くことにより，異なるクラスタ同士は互いに異なるラベル集合付けと対応するようにした．クラスタにユニークなラベル付けをすることでユーザにとってクラスタの判別が明確になることが実験によって示されている．さらにこのようなラベル集合とマッチングする文書集合を同じクラスタの要素とするように再クラスタリングすることでラベルの特徴を強調することができた．

2 LBHC アルゴリズム

LBHC アルゴリズムは 4 つのフェーズから成る．まず Web 検索結果に対して前処理を行い，特徴語を抽出する．次に特徴語-文書行列を作成し，SVD により非排他的ラベルクラスタを生成する．ここまでは SHOC アルゴリズムとほぼ一緒である．3 番目の処理として，クラスタ間においてクラスタに属すラベル集合の要素が重複しないように再配置を行う．最後に，クラスタ内のラベルに基づいて文書集合とのマッチングにより文書の再クラスタリングを行う．またこのとき，生成したクラスタ内のラベル間においてマッチングする文書集合同士の要素を比較する．要素間に包含関係がみられた場合，クラスタを階層構造になるよう更新する．

2.1 フェーズ 1: 特徴語抽出

Web 検索結果クラスタリングにおける入力は Web ページ全体の内容ではなく Web ページのタイトルおよび要約文である．これらの入力から無意味な特徴語が抽出されないためには，適切な前処理が重要である．また，ここで抽出される特徴語はクラスタのラベル候補となるため，後の処理に大きな影響を及ぼす．そのため前処理として，HTML 文から,HTML タグや非文字記号を除去し, 句読点によってフレーズに分割する．分割された文書集合を $D = \{d_1, d_2, ..., d_m\}$(ただし,$d_i$ はそれぞれ分割されたフレーズの集合) とする．そして，文書集合 D から形態素解析により名詞を抽出して TF-IDF 法で重み付けし, 閾値 k 以上の重みをもつ特徴語集合 $T = \{t_1, t_2, ..., t_n\}$ を抽出する．

2.2 フェーズ 2: ラベルクラスタの生成

文書集合と特徴語集合との関係をベクトル空間モデルを用いて表現する. 具体的には, 特徴語 t_i・文書 d_j を用いて, 以下のように特徴語・文書行列を構築する．文書 d_j は, 次のようなベクトルで表される．

$$\text{文書ベクトル}：\boldsymbol{d_j} = \begin{pmatrix} a_{1,j} \\ a_{2,j} \\ \vdots \\ a_{m,j} \end{pmatrix}$$

ここで，$a_{i,j}$ は tf-idf 法による重みである。また，文書集合全体は，次のような $m \times n$ の特徴語・文書行列

Aによって表現することができる.

$$A = [\boldsymbol{d_1}, \boldsymbol{d_2}, \cdots, \boldsymbol{d_n}] = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

特徴語・文書行列の第j列はj番目の文書に関する情報を表し,同様に,特徴語・文書行列の第i行はi番目の特徴語に関する情報を表しているベクトルである.

ベクトル空間における特徴語ベクトルの次元は,文書から抽出される特徴語の総数と等しいので,文書数が増えると,特徴語の数も膨大となり,計算機のメモリに入り切れなくなるだけでなく,文書に含まれる不必要な特徴語がノイズにもなる.そこで,本研究では,特徴語・文書行列Aを特異値分解することで,特徴語ベクトルの次元を削減し,文書と特徴語を表現するための最適な非排他的クラスタを求める.

特徴語・文書行列Aはフェーズ1より,入力の文書集合とそれから抽出された特徴語の集合とを用いて生成される.この行列Aを用いて非排他的クラスタを求める.その前にまず非排他的クラスタの定義を行う.

Definition 1 m個の要素$t_1, t_2, ..., t_m$を持つ集合の非排他的クラスタC_kはm次元の単位ベクトルx_k, $|x_k| = 1$を用いて表すことができる.ベクトル要素$x_k(i)$はクラスタC_kにおける要素t_iの適合度を指す.x_kをクラスタC_kのクラスタベクトルと呼び,クラスタC_kと同義に扱う.

本アルゴリズムでは,適合度によるクラスタリングを行列Aの特異値分解によって求める.

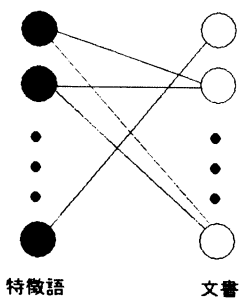

図 1: 二部グラフ

Aは特徴語と文書との関係を表す隣接行列であり,図1のような二部グラフで表現できる.二部グラフの節点はそれぞれ特徴語と文書を表し,枝には特徴語が文書に属す適合度を付加する.適合度はTF-IDF法による重みである.また,二部グラフからは特徴語と文書との二重強化関係がわかる.例えば,同じ文書と辺を結ぶ特徴語同士は意味的に近く,同じ特徴語と辺を結ぶ文書同士は意味的に近いと考えられる.これから,二部グラフのリンク関係から特徴語と文書とが最も意味的に近い部分構造を取り出し,一つのクラスタとして形成すべきだと考えるのが自然である.しかし,グラフの最適クラスタ分解はNP完全であり,また,非排他性にも対応できない.そこで,部分構造を部分グラフでなく,正規化された重みベクトルで表現する.クラスタにおいて特徴語と文書との間の関係の度合いは次のように定義できる.

Definition 2 $\boldsymbol{x_g}$をAの列ベクトルで表現されるクラスタ(特徴語クラスタ)とする.クラスタ$\boldsymbol{x_g}$の密度を$|\boldsymbol{x_g^T}A|$で定義する.同様に行ベクトル$\boldsymbol{y_g}$で表現されるクラスタ(文書クラスタ)の密度は$|A\boldsymbol{y_g}|$である.

求めたいのは高い密度をもつクラスタである.なぜなら,これらは意味が類似する文書集合をよく表しているからである.$\boldsymbol{x_1}$を最も密度の高いクラスタ,$\boldsymbol{x_2}$をその他のクラスタとすると,
$\boldsymbol{x_2}$は$\boldsymbol{x_2} = \eta \boldsymbol{x_1} + \left(\sqrt{1-\eta^2}\right)\boldsymbol{z}$と線形和で表せる.ただし,$\eta(0 \leq \eta \leq 1)$はスカラー,$\boldsymbol{z} \perp \boldsymbol{x_1}$, $|\boldsymbol{z}| = 1$とおく.
よって,$\boldsymbol{x_2}$のクラスタの密度は

$$|\boldsymbol{x_2^T A}| = \sqrt{\eta^2 |\boldsymbol{x_1^T A}|^2 + (1-\eta^2)|\boldsymbol{z^T A}|^2}$$

となることから,ηの値が大きければ,$\boldsymbol{x_2}$のクラスタ密度も増すことになる.もし,$\boldsymbol{x_2}$に対して制約がなければ,$\boldsymbol{x_1}$にいくらでも近づけることが可能となる.これでは$\boldsymbol{x_2}$と$\boldsymbol{x_1}$は同じクラスタを表すことになってしまうため,ここで新しく意味のあるクラスタ$\boldsymbol{x_g}$を得るには$\boldsymbol{x_g}$を制限する.つまり,すでに見つかったクラスタベクトルに対して直交するよう次のように定義する.

Definition 3 Aの列(行)における直交クラスタリングとは$\{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_k}\}$であるクラスタベクトルの集合を見つけることである.ただし,求める$\boldsymbol{x_g}(1 \leq g \leq k)$は,最大密度のクラスタであり,$\{\boldsymbol{x_1}, ..., \boldsymbol{x_{g-1}}\}$と直交している.

直交クラスタリング問題を解くために次の定義と定理を紹介する.

Definition 4 任意の $m \times n$ 行列 A, $rank(A) = r$ とし，$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$ は $AA^T(A^TA)$ の r 個の非負固有値,$\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_m}(\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_n})$ は正規直交の固有ベクトルとすると，
A の特異値分解 (*singular value decomposition;SVD*) は

$$A = U \begin{pmatrix} \Sigma & O \\ O & O \end{pmatrix} V^T$$

ただし，Σ=$diag(\sigma_1, ..., \sigma_r)$, $\sigma_g = \sqrt{\lambda_g}(g = 1, 2, ..., k)$ を A の特異値と呼び，$\boldsymbol{U} = [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_m}]$, $\boldsymbol{V} = [\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_n}]$, $\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_m}$ $(\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_n})$ を A の左 (右) 特異ベクトルと呼ぶ.

Definition 5 M を $m \times m$ の実対称行列として，M の $\boldsymbol{x}(\boldsymbol{x} \in R^m \backslash 0)$ に対する Rayleigh 商 (Rayleigh quotient) とは，

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}^T M \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$$

のことをいう.

Theorem 2.1 M が $m \times m$ の実対称行列で，その固有値を $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ とし，それに対応する固有ベクトルを $\boldsymbol{p_1}, \boldsymbol{p_2}, \cdots, \boldsymbol{p_m}$ とする.
もし $\forall \boldsymbol{x} \in R^m$ ならば,

$$\max_{\boldsymbol{x} \neq 0} R(\boldsymbol{x}) = R(\boldsymbol{p_1}) = \lambda_1, \min_{\boldsymbol{x} \neq 0} R(\boldsymbol{x}) = R(\boldsymbol{p_m}) = \lambda_m$$

もし $\forall \boldsymbol{x} \in L(\boldsymbol{p_g}, \boldsymbol{p_{g+1}}, ..., \boldsymbol{p_h}), 1 \leq g \leq h \leq m$ ならば,

$$\max_{\boldsymbol{x} \neq 0} R(\boldsymbol{x}) = R(\boldsymbol{p_g}) = \lambda_g, \min_{\boldsymbol{x} \neq 0} R(\boldsymbol{x}) = R(\boldsymbol{p_h}) = \lambda_h$$

SVD の定義により，行列 A の特異値分解 (SVD) を行うことで A の列ベクトル（または，行ベクトル）の直交クラスタリング問題を解決できる．つまり，以下の定理が成り立つ.

Theorem 2.2 特徴語・文書行列 A の左 (右) 特異ベクトルは，A の行 (列) ベクトルの直交クラスタリングによるクラスタベクトルである.

Proof: AA^T は $m \times m$ の実対称行列であるから，$\boldsymbol{x} \in R^m$ に関して AA^T の Rayleigh 商は

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}^T \left(AA^T\right) \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \frac{|\ \boldsymbol{x}^T A\ |^2}{|\ \boldsymbol{x}\ |^2}$$

となる．$\boldsymbol{c_g}$ が A の行ベクトルのクラスタを表すとき，

$$|\boldsymbol{c_g}| = 1, R(\boldsymbol{c_g}) = \frac{|\ \boldsymbol{c_g}^T A\ |^2}{|\ \boldsymbol{c_g}\ |^2} = |\ \boldsymbol{c_g^T} A\ |^2$$

とかけるので，$|\ \boldsymbol{c_g^T} A\ | = \sqrt{R(\boldsymbol{c_g})}$ となり，すなわち $\boldsymbol{c_g}$ のクラスタ密度は $\boldsymbol{c_g}$ に関して AA^T の Rayleigh 商の平方根に等しい．直交クラスタリングの定義によれば，$\boldsymbol{c_g}$ は最大密度のクラスタであり，$\{\boldsymbol{c_1}, ..., \boldsymbol{c_{g-1}}\}$ と直交していなければならない．Rayleigh の定理 (定理 2.1) より，$\boldsymbol{c_g}$ は AA^T の g 番目の固有ベクトル $\boldsymbol{p_g}$ であり，また,

$$AA^T = \underbrace{\left(U\Sigma V^T\right)}_{\text{特異値分解}} \underbrace{\left(U\Sigma V^T\right)^T}_{\text{特異値分解}} = \overbrace{U\Sigma^2 U^T}^{\text{固有値分解}}$$

となるので，$\boldsymbol{c_g}$ は行列 U の g 番目のベクトル，即ち行列 A の g 番目の左特異ベクトルである．A の列ベクトルついても同様に示すことができる. □

Definition 6 特異値分解により求められたクラスタベクトルに負の要素が現れる場合もあるが，それぞれのクラスタ $\boldsymbol{x_g}$ に対して，$\sum_{i=1}^{m} \boldsymbol{x_g}(i) \geq 0$ という制約をつける.

次に，適切なクラスタメンバの決定問題について扱う．最初に，クラスタ行列を定義する.

Definition 7 $\boldsymbol{x_g}$ のクラスタ行列を $X_g = \boldsymbol{x_g}(\boldsymbol{x_g^T} A)$, 同様に，$\boldsymbol{y_g}$ のクラスタ行列を $Y_g = (A\boldsymbol{y_g})\boldsymbol{y_g^T}$ とする．クラスタ行列は，特徴語・文書行列 A の中でその対応する部分を表わしている.

Theorem 2.3 $A = \sum_{k=1}^{r} (\sigma_k \boldsymbol{x_k} \boldsymbol{y_k^T})$

Theorem 2.4 $X_g = Y_g = C_g = \sigma_g \boldsymbol{x_g} \boldsymbol{y_g^T}$

Proof: $X_g = \boldsymbol{x_g}(\boldsymbol{x_g^T} A) = \boldsymbol{x_g}\left(\sum_{k=1}^{r} (\sigma_k \boldsymbol{x_g^T} \boldsymbol{x_k} \boldsymbol{y_k^T})\right)$
特異ベクトル $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_m}$ は直交ベクトルであるので,

$$\boldsymbol{x_g}^T \boldsymbol{x_k} = \begin{cases} 0 & \text{if } g \neq k \\ 1 & \text{if } g = k \end{cases}$$

$$\sum_{k=1}^{r} (\boldsymbol{x_g}^T \boldsymbol{x_k} \sigma_k \boldsymbol{y_k}^T) = \sigma_g \boldsymbol{y_g^T}$$

よって，$\boldsymbol{X_g} = \boldsymbol{x_g}\sigma_g\boldsymbol{y_g}^T = \sigma_g\boldsymbol{x_g}\boldsymbol{y_g}^T$
同様に，$\boldsymbol{Y_g} = (\boldsymbol{Ay_g})\boldsymbol{y_g}^T = \sigma_g\boldsymbol{x_g}\boldsymbol{y_g}^T$

□

この定理は，$\boldsymbol{x_g}$ と $\boldsymbol{y_g}$ によって表される二つのクラスタが実は同じトピックに関するものであることを意味する．即ち，$\boldsymbol{C_g}$ に対応するクラスタを特徴語ベクトルで表現したものが $\boldsymbol{x_g}$，文書ベクトルで表現したものが $\boldsymbol{y_g}$ である．また，入力 A に対して k 次元まで削減された行列 A は，

$$\boldsymbol{A_k} = \sum_{g=1}^{k} \boldsymbol{C_g} = \sum_{g=1}^{k} (\sigma_g \boldsymbol{x_g} \boldsymbol{y_g^T})$$

とかけ，直交クラスタリング $\{\boldsymbol{x_1}, ..., \boldsymbol{x_k}\}$ の精度は A_k/A の比率に反映される．

Definition 8 k 次元までの直交クラスタリングにおける A の精度は

$$q(\boldsymbol{A}, k) = \frac{\| \boldsymbol{A_k} \|_F}{\| \boldsymbol{A} \|_F} = \frac{\sqrt{\sum_{g=1}^{k} (\sigma_g^2)}}{\sqrt{\sum_{g=1}^{r} (\sigma_g^2)}}$$

と表される．ただし，k は $1 \leq k \leq r$ を満たす．クラスタの精度の閾値 q^*(例:90 %) を与えると，求めるクラスタ数 k^* は q(A,k)$\geq q^*$ を満たす最小の k である．

SHOC アルゴリズムと同じように LBHC アルゴリズムも特徴語・文書行列における直交クラスタリングを行う．その際クラスタの適合度における閾値 t を与え，g 番目の文書を成分とするラベルクラスタを U_g とする．なお，U_g のクラスタベクトル $\boldsymbol{x_g}$ は閾値 t 以上のもののみ成分にもつ．ここで選ばれるラベルの特徴として最も値が大きいものを選ぶというよりも，閾値以上のものを複数選ぶこととする．また，クラスタ生成の特徴より同じラベルが複数のクラスタに存在することからクラスタ同士は直交するが，同じラベルを複数のクラスタが含むためラベル配置を改めて最適化する問題が生じる．

2.3 フェーズ 3: ラベルクラスタの最適化

ラベル配置問題の定義を行い，最小コストフロー問題に帰着する．SVD を用いた直交クラスタリングでは文書の多義性を有効に利用できる非排他的クラスタを生成する反面，クラスタのラベルづけ時において不都合が生じる．すなわち，SHOC ではもっとも適合度が高いものをラベルづけするが，ラベルクラスタも非排他的であるため異なるクラスタ間において同一のラベルになることがしばしばおきる．一つクラスタに対しての最適なラベルを一つとるという手法では，異なるクラスタにおいて同一ラベル付けされた場合，クラスタの選定ができないばかりでなく，最適ではないが意味あるラベルを無視してしまう恐れがある．そこで本研究では，ひとつのクラスタにおいて複数のラベルをもつようにし，また全てのラベルに対して排他的にクラスタに属すようにラベルを配置する．ラベルは次に定義するラベル配置問題を解くことで最適なクラスタに配置できる．

Definition 9 ラベルが属する非排他的クラスタの集合を E=$\{\boldsymbol{e_1}, \boldsymbol{e_2}, ..., \boldsymbol{e_m}\}$，クラスタベクトル $\boldsymbol{e_i}$=$\{x_1, x_2, ..., x_n\}$ 内におけるラベルの適合度の和が重み $w(\boldsymbol{e_i}) = \sum_{k=1}^{n} x_k$ として与えられたとき，互いに素であるクラスタからなる集合 $S = \{\boldsymbol{s_1}, \boldsymbol{s_2}, ..., \boldsymbol{s_h}\}$ ($\boldsymbol{S} \subseteq \boldsymbol{E}$) の中で重み $w(\boldsymbol{S}) = \sum_{s_i \in S} w(\boldsymbol{s_i})$ が最大となるクラスタの組み合わせ (ただしラベルの数は最大で p とする)

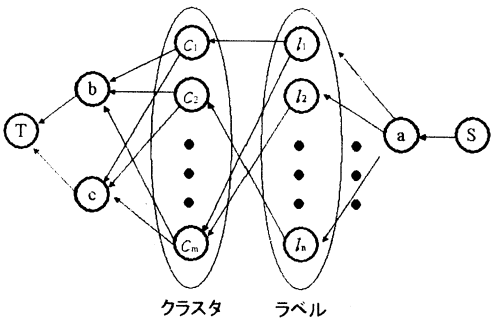

図 2: ネットワークフロー

与えられたラベルとクラスタとの 2 部グラフを次の手順に従ってネットワークフローを構成する．図 2 にあるように，ラベルとラベルとが属するクラスタの関係はラベルから属するクラスタに有向辺で接続することで表すことができる．また，次のように点や辺，辺の容量制約，単位流量あたりのコストを定義することで，ネットワークフローを作ることができる．

Definition 10 ラベル集合を $L = \{l_1, l_2, .., l_n\}$，クラスタ集合を $C = \{c_1, c_2, ..., c_m\}$ とすると，図 2 に

あるように，ネットワークフローの点集合は $V = L \cup C \cup \{a, b, c, S, T\}$ とかける．なお，$|L| = n$，$|C| = m$ とする．(ただし，$m \leq n$)．また，ラベル l_i からクラスタ c_j への辺を e_{ij} とし，辺集合を E とする．

供給点 S からラベル集合へのフローの中継点として頂点 a は S および L の各頂点と接続する．さらに，ラベル $l_i(1 \leq i \leq n)$ がクラスタ $c_j(1 \leq j \leq m)$ に属すなら，ラベル l_i とクラスタ c_j とを辺で接続する．さらに，C の各頂点は b，c と接続し，b と c は需要点 T と接続する．

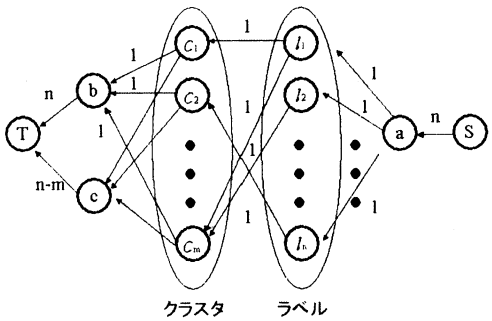

図 3: ネットワークフロー:上限容量

次に，辺の容量制約を定義する．図 3にあるように，S から a のフローの上限容量を $f_{max}(S \to a) = n$，a から L の各頂点へのフローの上限容量を

$f_{max}(a \to l_i) = n, (1 \leq i \leq n)$,

l_i と c_j との間で存在するフローの上限容量を 1，C から b の各頂点へのフロー上限容量を

$f_{max}(c_j \to b) = 1, (1 \leq j \leq m)$,

C から c の各頂点へのフロー上限容量の和を

$\sum_{C \to c} f_{max}(C \to c) = n - m$,

最後に，b から T，c から T へのフロー上限容量を

$f_{max}(b \to T) = m$,

$f_{max}(c \to T) = n - m$ とする．

単位流量あたりのコストは次のように定義できる．図 4にあるように，l_i と c_j とが接続していれば，定義 1より単位流量あたりのコストをラベル i のクラスタ j における適合度 $x_j(i)$ とする．またクラスタのすべての要素から c へはコスト 0 より大きくする．なぜなら b へ優先的にフローを流し，残余流を c へ流すためである．その他の辺における単位流量あたりのコストを 0 とする．

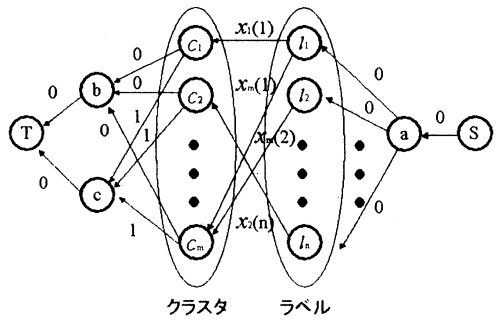

図 4: ネットワークフロー:単位流量あたりのコスト

Theorem 2.5 ラベル配置問題を最小コストフロー問題に帰着することができる．

Proof: ラベル配置問題の入力として与えられる適合度はネットワークフローでは，定義 10 よりクラスタと接続するラベルの辺の単位あたりのコストとして与えられる．ラベル配置問題が求める解はクラスタ集合に対して互いに素になるようなラベル集合であり，かつ，条件として個々のラベルがクラスタにおける適合度の和が最大となるようにすることである．これに対して定義したネットワークフローにおいて，流量保存則よりラベル l_i から c_j へはたかだか 1 しかフローが流れないので，ラベル l_i はたかだか一つのクラスタのみに属すといえる．さらに，ラベルとクラスタとの接続する辺以外のコストは 0 であり，かつ，定義 6 と定義 1 より，ラベル i のクラスタ j におけるコスト $x_j(i)$ は $0 \leq x_j(i) \leq 1$ を満たすことから，元問題の適合度の和が最大になるときのラベル集合を求めるにはコストを負の値に反転して，最小コストフロー問題の目的関数である

$$\sum_{e_{ij} \in E} cost(e_{ij}) \lambda(e_{ij})$$

が最小となる流量 λ を求めればよいことがわかる．よって，ラベル配置問題を最小コストフロー問題に帰着することができる． □

2.4 フェーズ 4: ラベルによる階層的再クラスタリング

前のフェーズより，求められたラベル集合に対して特徴語・文書行列 A を再度用いて，ラベル集合と文書集合とのマッチングを行い，ラベルに基づいて文書の再クラスタリングを行う．

それぞれのクラスタに属すラベル集合に対して，一つのラベルに対して一つの列をなすクラスタ行列 $q_i(1 \leq i \leq n, n$ はクラスタの数) を生成する．ただし，q_i の成分はフェーズ 3 より求められたラベルの適合度である．q_i の行の長さは特徴語ベクトルの長さ，列の長さはクラスタのラベル集合の要素数になる．クラスタ集合に対してクラスタ行列集合 $Q = \{q_1, q_2, ..., q_n\}$ を生成する．

$L_i = q_i^T A(1 \leq i \leq n)$ とすると，行列 L_i の要素 l_{jk} は j 番目のラベルにおける k 番目の文書の従属度を示す．ここで，閾値 w を定めて文書の従属度が w 以上のものを要素とする L_i' とし，これらの要素をラベルとマッチングした文書集合の要素とする．このような $\{L_1', L_2', ..., L_n'\}$ が求まったとして，さらに同じラベルクラスタ内において，ラベルとマッチングする文書集合同士が包含関係にある場合，ラベル間の関係を親と子の関係にした木構造を生成して，クラスタ内においてラベルの階層化を行う．

3 実行時間の評価

取得ずみの 200 件の Web 検索結果を用いて，10 回ずつクラスタリングを行い，各実行フェーズにおける平均実行時間を測定した．結果は以下の表 1 になっている．200 件の Web 検索結果に対して，システム全体の平均実行時間は 1〜2 秒程度である．そのうち，全体の処理時間の 7 割を特異値分解が占めている．特異値分解の計算時間は $m \times n$ の行列に対して，$O(mn^2)$ になる．

表 1: LBHC 実行時間 (検索結果 200 件)

フェーズ:	平均時間 [s]	比率
特徴語抽出:	0.316	26%
クラスタの生成:	0.852	73%
クラスタの最適化:	0.017	1%

4 クラスタリング結果の有効性検証

いくつかキーワードを検索エンジン Google に入力し，出力されたデータのクラスタリング処理を行った．ここではラベルとマッチングする文書番号に着目してクラスタリングの有効性を検証する．なお，文書番号とは Google における文書の出現順番をさす．以下一例として，キーワード {**検索エンジン**} に対する 200 件のクラスタリング結果を示す．

表 2: 生成されたクラスタ

検索キー	クラスタ数	ラベルの例
検索エンジン	26	ディレクトリ型 カテゴリー サイト登録

表 3: クラスタのラベルとメンバー

クラスタのラベル (文書数)	文書番号
ディレクトリ型 (15)	4,14,23,25,41,60, 62,77,92,100,157, 161,165,177,187
カテゴリー (7)	6,29,34,39,43, 124,132
サイト登録 (5)	40,71,155,161,169

統計的な調査報告より，検索エンジン利用者の 7 割はトップ 20 位までのページのみ閲覧することがわかっている．しかしながら，表 2 と表 3 とからわかるように，本システムで生成したクラスタラベルを用いることでユーザの欲する情報が下位ランクのページであってもラベルからすばやく辿ることが確認でき，提案するアルゴリズムは有効であると考えられる．

5 まとめ

本研究で提案する LBHC アルゴリズムの特徴として以下三つの点が挙げられる．一つ目は意味の類似するラベル集合を得るために SVD を利用したことである．二つ目の特徴としては生成した非排他的クラスタにおけるラベル重複問題に対して最小コストフロー問題に帰着したことである．これによりクラスタを明確に区別できるようにした．三つ目は最適化されたラベル集合に基づいて文書とのマッチングを行ったことである．ラベルと完全に対応する文書がクラスタの要素となることでラベルから Web ページへ確実に辿ることができるようにした．

今後の課題として 2 つ挙げる．一つはクラスタの階層化では単純に完全な包含関係のみに適用したが，他にも様々な手法が考えられ，比較改善の余地が残されている．二つ目に，システム全体の実行時間の 7 割以上を占める SVD に変わる手法の検討が挙げられる．サイズが大きな Web 検索結果の入力に対しても数秒以内にクラスタリングできるような手法が望ましいと考えられる．

参考文献

[1] Google Search Engine <http://google.com>

[2] Yahoo Search Engine <http://yahoo.com>

[3] Open Directory Project <http://dmoz.org>

[4] D.Cutting,D.karger,J,Pedersen,J.W.Tukey. Scatter/Gather:A Cluster-based Approach to Browsing Large Document Collections. *Proceedings of the 15th Annual International ACM/SIGIR Conference*, 1992.

[5] Dell Zhang and Yisheng Dong. Semantic, Hierarchical, Online Clustering of Web Search Results. Accepted by 3rd International Workshop on Web information and data management, Atlanta, Georgia.2001

[6] Krishna Bharat and Monika R. Henzinger. Inproved Algorithms for Topic Distillation in a Hyperlinked Environment. *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[7] Oren E. Zamir. Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results. Doctoral Dissertation, University of Washington, 1999.

[8] Oren Zamir and Oren Etzioni. Grouper: A Dynamic Clustering Interface to Web Search Results. WWW8/Computer Networks, Amsterdam, Netherlands, 1999.