ハイパーリンクの共起性を用いたクラスタリング手法

高橋　　功†　　三浦　孝夫†

† 法政大学 工学部 電気電子工学科　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †{c01d3089,miurat}@k.hosei.ac.jp

あらまし　Web ページのようなハイパーリンク構造を持つ文書を取り扱う場合, そのハイパーリンクが共起しているページは内容が酷似している特性がある. この考えにもとづき, 本論文ではハイパーリンクの共起性とベクトル空間モデルにおける類似度を考慮したクラスタリング手法を提案し, その手法の有効性を評価する.
キーワード　Web クラスタリング, Web マイニング

Clustering Web Contents Based on Correlation of Hyperlinks

Kou TAKAHASHI† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †{c01d3089,miurat}@k.hosei.ac.jp

Abstract　In this investigation, we propose a novel technique of Web clustering based on an idea that similar documents share hyperlinks. We extract characteristic hyperlinks as well as index words to distinguish from each other and show some experimental results.
Key words　Web Clustering, Web Mining

1. 前　書　き

これまで数多くの Web クラスタリング手法が提案されている [1]. その目的は様々であり,Web 上でのクラスタリング, Web ログ・セッション入手,Web セッションクラスタリング, Web コミュニティ検出 (Authority や Hub),Web 文書クラスタリング, 検索エンジン結果の集約など多岐に渡っている.

クラスタリング (Clustering) はオブジェクト集合へのグルーピング手法であり, 同じクラスタ内のオブジェクトは類似し異なるクラスタのオブジェクトは似ていない様に振り分ける [7]. つまり, クラスリング技法は”類似性”の定義とその実行方法に依存して, 隠れたパターンをどれだけ見出せるかを競い合っているといえる. これまで知られたクラスタリング技法は, 大きく分割方式 (オブジェクト集合を分割し, ある基準で評価する), 階層化 (オブジェクト集合をある基準で階層的に分解する), 密度に基づく手法 (結合度・密度関数による評価) などに大別され, 類似性の定義は距離の定義として考察されることが多い.

Web **クラスタリング** は Web 情報の利用度の向上,Web 探索経路の短縮, 利用者要求への対応・応答の向上, 検索性能 (Recall/Precision) の向上, 内容提示の質的向上, 利用者動作の意図の理解, データ表現標準への対応,Web 情報構造の改善などを目的としたものであり, 上述クラスタリングと変わるものではない. Web 文書をクラスタリングすることによって, 相互に関連する Web ページのグループを検出し Web コミュニティを得るものや, 類似した内容をもつページ集合にまとめ, Web 検索の性能向上・検索結果の質的向上を図ることができる.

対象とするデータは,Web 文書 (ページ内容) だけでなくて, 利用者動作を記述するログ情報も含まれる. このとき,Web 文書集合をグループ化するための特徴値として何を用いればよいのであろうか？これまで, 検索エンジンの結果を解析するという立場からの提案は多い. 検索エンジンへの要求に何らかの共通性があり, これを手がかりに”強く関連した文書は同じ質問に反応しがち” という特性から, 利用者意図の表現を探る多くのクラスタリング手法が提案されている. 例えば Scatter/Gather [2], STC [10], Carrot2 [9] などが代表的である.

本稿では, 情報検索手法を用いてカテゴリカルクラスタリングを Web 文書に適用する手法を提案する. 本稿で扱うデータのほとんどはカテゴリカル (”Gold, Silver, Blonde” など) であり, 数値データではない. このため距離や順序概念が考えにくく, 過去に提案された多くのアプローチがなじみにくいの理由のひとつになっている. 本稿ではハイパーリンクおよび Web 文書内に生じる索引語の分布頻度や共起性を分析し, また関連性によるグラフ構造分析を用いたクラスタリング手法を提案する.

第 2 章は Web 文書のクラスタリング手法を, 続く 3 章では提案手法を示す. 実際の Web 文書データを用いた実験結果を第 4 章で述べる.

2. Web 文書クラスタリング

Web **文書クラスタリング**は類似した内容の Web 文書集合を得ることを目的とするクラスタリングである．Web 文書に対して，”文書特性”と”ハイパーリンク”による構造を利用したクラスタリングが適用される．

文書特性を利用したクラスタリングでは, Web 文書は (通常のテキストクラスタリングと同様に) ”単語の多重集合” (Bag of Words) として表現される [7]．各文書はベクトルで表され、全体としてベクトル空間を構成する．ベクトルの各要素は対応する単語の出現頻度に対応し, 文書間の非類似度を対応するベクトル間の余弦 (cosine) 値を用いて記述する．文書に生じる各語については，語幹を抽出し (stemming) 不要語 (stop word) を除去するなどの事前操作により，文書の特定を効率よく行う必要がある．しかし，文書数が増えるにつれ高次元化していくという問題点があるため，特徴的な語 (索引語 index term) を選び出して次元数を限定するなどの工夫が必要である．

一般の文書と比較して，Web 文書に際立つ特徴について配慮せねばならない．例えば，少ない語だけで特徴的な Web 文書(注 1)や，空間配置，CSS/XML, 色彩, フォント, マルチメディアといった Web 文書の特殊性を吸収する必要があろう．これらの特性のうち**ハイパーリンク** (他 Web 文書への参照) は，Web 文書間の意味的な結びつきを明示的な構造で表すと言う点で重要である．ハイパーリンク構造は有向グラフで表現することができる．頂点が Web ページ, 辺がハイパーリンクに相当し, 参照の数はトピックの注目度に対応する．ただ良く知られているように，参照/非参照の頻度 (構造情報) によるクラスタリングを行うと，巨大サイトへの参照のみでクラスタが形成されることが多い．つまり，少数の巨大・準巨大クラスタと多数の泡沫クラスタが生成されやすく，実質的に特徴的なクラスタ集合を得にくいという問題点がある．

3. 提 案 手 法

本稿で提案する Web 文書クラスタリング手法は, Web 文書の文書特性とハイパーリンク構造を反映したものであり, 直感的で単純な方法である．HITS アルゴリズム [8] の解析等でよく知られているように, オーソリティ(authority) とは非参照 Web 文書 (ページ) のうち特定のトピックにおける的確な情報を持つと承認されているものを意味する．このため同じ authority を参照する Web 文書は同一トピックに言及している可能性が高く, 当該トピックにに関して類似していると考えてよい．

この考え方を用いて, 本稿ではハイパーリンクの共起性を利用したクラスタリング (Link クラスタリングと呼ぶ) を行う．同時に，索引語により Web 文書をベクトル化し (当該ベクトル空間上で) ベクトル集合をクラスタリング (VSM クラスタリングと呼ぶ) を生成する．この 2 つの結果を”重ね合わせる”ことにより, 同一のトピックを参照し, かつ文書の酷似しているクラスタへと分割する．

(注 1)：例えば“ 飛べ赤星！”とだけ記述された Web 文書がある．

3.1 Link クラスタリング

はじめに Link クラスタリングを定義する．このため有向グラフ G を用いた形式化を行う．有限集合 $N = \{a_1, .., a_n\}$ および $E \subseteqq N \times N$ が与えられたとき $G = \langle N, E \rangle$ を有向グラフという。ただし，N の要素を頂点，$(a, b) \in E$ を始点を a, 終点を b とする辺という．G では，始点および終点それぞれが同じである辺は唯一つしかなく，またサイクル (a, a) は無いとする．頂点 a から出る辺の集合 $From(a) = \{b \in N \mid (a\, b) \in E\}$ を a からの出辺集合 (要素数を出次数)，逆に頂点 b へ入る辺の集合 $To(b) = \{a \in N \mid (a\, b) \in E\}$ を入辺集合 (要素数を入次数) という．頂点 (node) を Web 文書に, 辺 (arc) をハイパーリンクに対応させれば, Web 文書集合上のハイパーリンク構造は有向グラフで表現することができる．参照数は入次数に対応しており，トピックの注目度に対応する．一般に $|From(a)| \gg 0$ となる a はハブ (hub), $|To(b)| \gg 0$ となる b はオーソリティに対応する．なお，本稿では出次数が 0 の頂点は (トピックが独立しており) 除外する．

Link クラスタリングの手順を示す．実際の手順は完全連結法を用いた，階層型クラスタリングによる．2 つの頂点 a_i, a_j に対して，a_i と a_j の値 d_{ij} を次式で与える．

$$d_{ij} = 1 - \frac{2\,|From(a_i) \cap From(a_j)|}{|From(a_i)| + |From(a_j)|} \tag{1}$$

d_{ij} は a_i, a_j の双方から参照されている頂点数 (共起数) の割合を用いて定義されていることに注意したい．実際，この距離は頂点の辺の数に依存せず同じ終点への辺の割合と対応している．d_{ij} は，共起の割合が大きいほど 0.0 に，少ないほど 1.0 に近づく．このため，(共起性に関する) 非類似度と呼ぶ．$n \times n$ の行列 $D = ((d_{ij}))$ を非類似度行列と呼ぶ．定義から D は対称行列である．

非類似度行列 D を用いてクラスタリングを行う [**?**]．このクラスタリング手法を *Link* **クラスタリング**，その結果の各クラスタを ”Link クラスタ” と呼ぶ. 以下では，クラスタのうち要素数が閾値 θ 以下のものを破棄する．実際，頂点の数少ないクラスタは内容を判断することができず，誤った判断を招く可能性が高い。

前節で指摘したように，Link クラスタリングを実行するとき，(検索エンジンサイトなどの) 巨大なハブ・オーソリティだけからなるクラスタが検出され，効果的で意味のある結果が得られないことが多い．本研究では，Zipf の法則を用いて，効果的なハイパーリンクだけを抽出する(注 2)．

(注 2)：Zipf の法則とは高次元化を抑制するための次元縮小技法の一つで, 高頻度の単語で成り立つ Zipf の第 1 法則と, 低頻度の単語で成り立つ Zipf の第 2 法則がある．低頻度の単語をどの程度削除するかの基準として, まず「中程度の頻度」を決める必要がある．頻度 1 の単語数を F_1 とすると,2 つの法則を同時に満たす中程度の単語頻度 f_k は, 以下の式で求められる．

$$f_k = \frac{\sqrt{8F_1 + 1} - 1}{2} \tag{2}$$

ここで得られた出現頻度 f_k が索引語の頻度順位において中間地点であることを仮定すれば, 以下の手順で索引語数を決定できる．

（1） 出現頻度 f_k を持つすべての語を索引語とする

［例 1］ ここでは Link クラスタリングの例を示す．図 1 のように 6 個の頂点 $a_1 \cdots a_6$ があるとき Zipf の法則によりいくつかの頂点を破棄する．この例では閾値 $\theta = 1$ としている．頂点

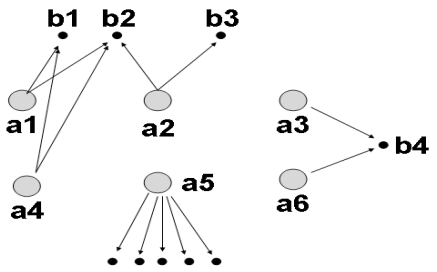

図 1　例 1:Link クラスタ

$a_1, .., a_6$ の出次数はそれぞれ 2,2,1,2,5,1 である．Zipf の法則より，$f_k = \frac{\sqrt{8 \cdot 2+1}-1}{=} 1.56$ であり, a_5 (ハブとみなすことができる) が除去される．これ以外の頂点の非類似度行列を D とする。これは以下のように求められる．Link クラスタリングを行う．

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 2 & 3 & 4 & 6 \end{array} \\ \begin{pmatrix} 0 & 0.5 & 1 & 0 & 1 \\ 0.5 & 0 & 1 & 0.5 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0.5 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

このクラスタリングによりふたつのクラスタ $A_1 = \{a_1, a_2, a_4\}$, $A_2 = \{a_3, a_6\}$ が生成できる．頂点 a_5 は削除されたため、孤立点 (1 点だけからなるクラスタ) とみなして削除する．

□

3.2　VSM クラスタリング

Web 文書クラスタリングでは Web 文書からアンカータグを取り除いたプレーンテキストを対象にする．ここでは大量の文書を扱うために, ページごとの単語の出現頻度を扱うと，Web 文書の文字数による偏りが生じる可能性がある．このため，本稿では各 Web 文書の重みを 0(未出現), 1(出現) の 2 値で表現したベクトル $\vec{p_i}$ を用いる．

m 個の Web 文書集合 $P = \{\vec{p_1}, \cdots, \vec{p_m}\}$ に対し $\vec{p_i}$ と $\vec{p_j}$ の非類似度 d_{ij} を次で定義する．

$$d_{ij} = 1 - \frac{(\vec{p_i} \cdot \vec{p_j})}{|\vec{p_i}\cdot| \, |\vec{p_j}|} \tag{3}$$

この d_{ij} から非類似度行列 $D = ((d_{ij}))$ を定義し，先と同様に完全連結法による階層型クラスタリングを行う．この手法を *VSM* **クラスタリング**，その結果のクラスタを ”VSM クラスタ” と呼ぶ．

（2） 第 1 順位から $f_k - 1$ 個の頻度を持つ語までのすべてを索引語とする．全部で K 個の語があるとする

（3） $f_k + 1$ 以下の出現頻度の語のうち, 上位 K 個を索引語とする

［例 2］ VSM クラスタリングの例を示す．6 個の Web 文書 $a_1, \cdots, a_6$ に対応して文書ベクトルが図 2 で与えられているとする．このとき,VSM クラスタリングを行う、ここで閾値 $\theta = 1$ とする．

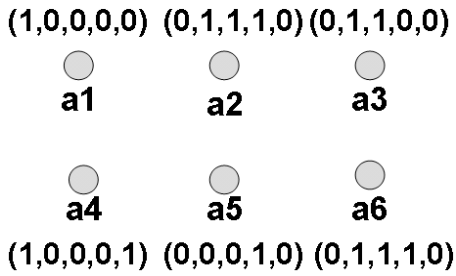

図 2　例 2:VSM クラスタ

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0.5 & 1 & 1 \\ 1 & 0 & 0.67 & 1 & 0.67 & 0.67 \\ 1 & 0.67 & 0 & 1 & 0.75 & 0.75 \\ 0.4 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0.67 & 0.75 & 1 & 0 & 0.75 \\ 1 & 0.67 & 0.751 & 1 & 0.75 & 0 \end{pmatrix} \end{array}$$

各ベクトルの非類似度を上記の行列 D で表し，階層型クラスタリングを行う．この結果，2 つのクラスタ $B_1 = \{a_1, a_4\}$, $B_2 = \{a_2, a_3, a_5, a_6\}$ が生成される．

□

3.3　クラスタの重ね合わせ

文書特性とハイパーリンク構造の特性の双方を備え, さらに Link クラスタ結果を効果的に分割するために，2 つのクラスタリング結果を重ね合わせる方法を考える．

Link クラスタリングによる n 個のクラスタ $\mathcal{A} = \{A_1 \cdots A_p\}$, VSM クラスタリングによる m 個のクラスタ $\mathcal{B} = \{B_1 \cdots B_q\}$ に対して，さらに A_0 と B_0 をそれぞれの手法で破棄された頂点の集まりとする．このとき, **クラスタの重ね合わせ** を次式で定義する．

$$C_{ij} = A_i \cap B_j \tag{4}$$

ただし C_{ij} の要素数が閾値 θ を下回れば破棄する．このとき C_{ij} は最大で $p \times q$ 個得られる．

クラスタの重ね合わせのアルゴリズムは以下の通りである．

（1） $C_{ij} = \emptyset$ とする，$i = 1, .., p, j = 1, .., q$

（2） 各 $a_i \in N$ に対して (3)-(5) を行う

（3） $a_i \in A_k$ となる k を求める

（4） $a_i \in B_{k'}$ となる k' を求める

（5） a_i にクラスタ $C_{kk'}$ を割り当てる

［例 3］ 重ね合わせたクラスタの例を示す．図 3 は例 1 の Link クラスタを A_1, A_2 を円形で, 例 2 の VSM クラスタ B_1, B_2 を矩形で表している．閾値 $\theta = 1$ としたとき, Link クラスタと VSM クラスタを重ね合わせると, クラスタ $C_{11} = \{a_1, a_4\}$ と, $C_{22}\{a_3, a_6\}$ に分割される．クラスタ $C_{12} = \{a_2\}$ と $C_{02} = \{a_5\}$

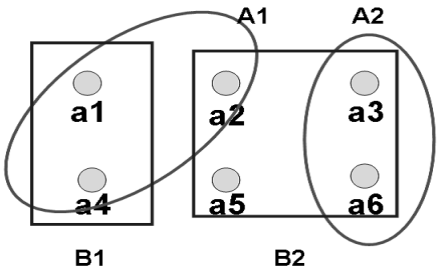

図 3　例 3:重ねたクラスタ

は閾値以下の頂点数のため破棄される.

□

4. 実　　験

4.1 実験環境

本研究では,　実験データとして,NTCIR-3 Web 文書データ(注 3) を使用する．NTCIR-3 は 2001 年 8 月 29 日から 2001 年 11 月 12 日の間に収集した.jp ドメインの拡張子 html と text データを集めたテスト・コレクションである．100Gbyte を越えるデータを含む NW100G-01,10Gbyte-01 のデータを含む NW10G-01 の 2 つがある．このテスト・コレクションはとくに Web 文書を対象とした検索, 分類, 情報抽出などの情報活用システムの比較評価, ならびに,Web テストコレクションの構築を目的としている．NW100G-01 中から 2001 年 9 月 29 日から 2001 年 10 月 5 日までに収集した 9929 件の日本語の Web サイトのデータを用いる.

本稿では，Zipf の法則に基づき頂点となる 3234 件を抽出する. またその頂点が持つ 2,429,984 個の辺と 3,825,293 個の単語に Zipf の法則を適用し 1285 個の辺,449 個の単語を抽出する. これより Link クラスタ及び VSM クラスタを求め, 2 つのクラスタから重なり合うクラスタをつくる. またクラスタの要素数が閾値 θ が 5 以下ならば破棄する.

4.2 実験 (1):VSM クラスタ 結果と考察

VSM クラスタを用いて得られた代表的なものを表 1 に示す. 表 1 より，VSM1,3,4 クラスタは個人のページが多く, VSM2,5 クラスタは企業, 大学等の公式ページが多く集まっていることがわかる．VSM によるクラスタリングでは類似した単語集合を持つページ同士がクラスタになりやすいため, ページ内で口語体, あるいは文語体であるという差が結果に現れている.

4.3 実験 (2):Link クラスタ 結果と考察

表 2 に Link クラスタリングの結果を示す。ここで参照しているハイパーリンクからそれぞれのクラスタは以下のようなトピックが含まれていると考えられる.

Link1 クラスタは「大学, 地域の話題, 書籍」に関するページへの参照が多く, 頂点は「個人サイト (日記, 旅行)43 個, 大学 7 個, 企業 10 個」を持つ.

Link2 クラスタは「OS, セキュリティ」で頂点は「個人サイト (日記, 無料掲示板)30 個, 大学の研究室 13 個・企業 23 個」を持つ.

Link3 クラスタは「プロバイダ」で, 頂点は「個人サイト (日記)24 個, 大学 2 個」. いずれのクラスタも参照がトピックと対応しているとするならば, トピックと頂点は対応している見ることができる.

(注 3)：http://research.nii.ac.jp/ntcir/

クラスタ名	
VSM1	個人サイト 203 個 大学, 企業　63 個
VSM2	個人サイト 40 個 大学, 企業　113 個
VSM3	個人サイト 127 個 大学, 企業　111 個 (図書館, 書籍に関するもの 25 個)
VSM4	個人サイト 109 個 (写真, イラストに関するもの 51 個) 大学, 企業　54 個
VSM5	個人サイト 72 個 大学, 企業 165 個

表 1　VSM の代表的なクラスタの内容

しかしトピックが複数にわたるクラスタでは全体としてまとまりが悪く，一見して集約することは容易ではない.

クラスタ名	Link1	Link2	Link3
クラスタ内の頻出辺	jp.freebsd.org	sun.com	try-net.or.jp
	aozora.gr.jp	linux.org	freeweb.ne.jp
	washingtonpost.com	sgi.com	so-net.ne.jp/postpet
	un.org	netbsd.org	webring.ne.jp
	tsukuba.ac.jp	linuxhq.com	ixla.com
	suzuki.co.jp	hp.com	alles.or.jp/ queen
	shogakukan.co.jp	freebsd.org	geocities.co.jp
	sanyo.co.jp	w3.org/Daemon	altan.hr/snow
	pref.niigata.jp	tripod.com	6.big.or.jp/ neon
	pref.fukuoka.jp	specbench.org	ushikai.com
	pref.aomori.jp	sleepycat.com	azaq.net
	pref.akita.jp	sequent.com	hp.bird.to
	nytimes.com	sco.com	7.big.or.jp/ jawa
	nikkeibp.co.jp	rsa.com	eva.hi-ho.ne.jp/takeuchi
	nasda.go.jp	openbsd.org	w3.org
	mycom.co.jp/career	nikkansports.com	odn.ne.jp
	monbu.go.jp	netcraft.com/Survey	nifty.com
	mext.go.jp	ncr.com	jra.go.jp
	maruzen.co.jp	my.host.com	zakzak.co.jp
	kyushu-u.ac.jp	multihost.com	winamp.com

表 2　Link の代表的なクラスタの内容

4.4 実験 (3):重ね合わせたクラスタ 結果と考察

重ね合わせたクラスタの結果を表 3 に示す。ここでは　7 つのクラスタを得た.

Link1 クラスタは 3 つのクラスタに分割されている. すなわち，(VSM1 クラスタと重なった) 個人のページ（日記・旅行）のクラスタ，(VSM4 クラスタと重なった) イラストをメインとする個人ページのクラスタ，(VSM5 クラスタと重なった) 大学・NTT や公務員に関するページのクラスタである.

各々は各トピックと対応するクラスタに分割されている．実際 (Link1 クラスタの) トピック「地域の話題, 書籍, 大学」は VSM1 では個人サイト,VSM4 ではイラストをメインとする個人サイト, VSM5 では大学のクラスタであった．重ね合わせたクラスタがそれぞれのトピックに対応しているクラスタに分割されている．

Link2 クラスタでは (VSM1 クラスタと重なった) 大学の研究室（電気, 土木）と個人サイトのクラスタ，(VSM2 クラスタと重なった) 大学の研究室（化学）, 個人サイトのクラスタ，(VSM3 クラスタと重なった) 大学の研究室（電気, 医）と無料掲示板のクラスタという 3 つのクラスタに分割された．Link2 クラスタのトピック「OS, セキュリティ」には，いずれの重ねたクラスタにおいても大学の研究室サイトが含まれる．VSM1,3 クラスタは個人サイトが多いクラスタであるため重ねたクラスタでも個人サイトをみることができている．

Link3 クラスタは，VSM1 クラスタと重なり個人サイトのみのクラスタが得られた．反面，大学の研究室などのページが除去された．Link3 も VSM1 も「個人サイト」というトピックのため，それ以外のページが除去されている．

	VSM1	VSM2	VSM3	VSM4	VSM5
Link1	個人サイト			個人サイト	企業
Link2	大学 個人サイト	大学 (化学)	大学 （電気, 医）		
Link3	個人サイト				

表 3　重ね合わせたクラスタの内容

4.5 議　　論

これまでの結果の考察から，クラスタ結果を重ね合わせることで，より明確で的確なクラスタへと分割することができたと言える．類似したクラスタへと分割できた．しかし 各クラスタの頂点の数とクラスタ数の表 4 によると, クラスタを重ね合わせによりクラスタ要素数が減少し，破棄クラスタが増加している．このうち要素数 1 のクラスタは 1121 ある．これらはリンクの保持数もしくは単語の保持数が突出している Web 文書が多いため, 雑音であると考えられる．

要素数 2 ないし 4 のクラスタは，Link クラスタリングにより少ない要素数であったクラスタを更に分割したものが大半である．実験では Link1 クラスタと VSM2 クラスタは共に大学というトピックであり, 重ねたクラスタでは要素数が 3 であった．この 3 つにクラスタはいずれも大学に関するページである．同様に Link2 クラスタと VSM3 クラスタを重ねたときも要素数が 3 であり, 大学の研究室のページが 3 つであった．これらは閾値以下だったため破棄した．しかしこれらは正しくクラスタに分割できており，真に雑音かどうかは即断できない．重ねたクラスタにおいては閾値を下げる等の判断が必要である．

5. 関 連 研 究

クラスタリングは統計, パターン認識, データベース, データマイニングといった分野で研究されている. 類似しているオブジェクト同士をまとめていきデータ集合を分割するのに利用される. 類似度を分析するのに通常ユークリッド距離などを用いるためオブジェクトは数値属性で表現されているのが一般的であるが, カテゴリカル属性を扱う手法も研究されている.Web サイトなどの半構造データにおいてはカテゴリ属性を扱う必要がある. そこでリンク概念を用いたカテゴリカルクラスタリング手法として ROCK(Robust Clustering using linKs) [6] と CACTUS(CAtegorical ClusTering Using Summaries) [3] がある．ROCK とは 2 つの対象で共起しているリンクが Jaccard 係数などを用いて閾値以上であるとき対象は類似しているとする手法である．2 つの対象だけでなく, それらの近隣の影響を考慮することで少数の例外的な対象の影響を受けにくい特徴がある．

頂点の数	Link	VSM	重ね
1-5	171	54	1747
6-10	118	19	23
11-30	84	27	6
31-50	6	2	6
51-70	2	0	1
71-90	-	2	-
91-110	-	0	-
111-130	-	0	-
131-150	-	0	-
151-170	-	3	-
171-190	-	0	-
191-210	-	2	-

表 4　頂点数とクラスタ数の対応表

$$\sum_{i=1}^{k} n_i \times \sum_{x_p,x_r \in C_i} \frac{link(x_p, x_r)}{n_i}$$

リンクの類似度の大きな対象同士を同じクラスタに分類する目的でこの評価関数を最大化する.link(x_p,x_r) は対象 x_p と x_r の間のリンク数を表す.

これに対して, CACTUS は類似性に基づく近隣関係ではなく対象集合中の属性値の共起性に基づいた連結関係を用いる．共起性の強い属性についての要約情報があればデータ全体の情報がなくてもクラスタを抽出できる性質を持つため記憶容量を削減できる．

6. 結　　論

本稿では, ハイパーリンクの共起性とベクトル空間モデルを用いたクラスタを重ね合わせる手法を提案した．2 つのクラスタリングの結果を重ね合わせることにより類似したクラスタを生成することができた．しかし，重ね合わせにより細分化されたクラスタについては，別途評価を行う必要があり，今後の問題として残されている．

謝　　辞

本研究の一部は文部科学省科学研究費補助金 (課題番号 16500070) の支援をいただいた．

本実験に対しては国立情報学研究所より NTCIR-3 Web 文書データの提供をいただきました. 関係各位に深く感謝します.

文　　献

[1] Chakrabat,S.: Mining the Web, Morgan Kaufmann, 2003
[2] Cutting, D., Karger, D., Pedersen, J. and Tukey,J.W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, SIGIR, 1992
[3] Ganti,V., Gehrke, J. and Ramakrishnan, R.: CACTUS Clustering Categorical Data Using Summaries, Knowledge Discovery and Data Mining (KDDM), 1999
[4] Gibson,D., Kleinberg, J. and Raghaven, P.: Clustering categorical Data, An Approach Based on Dynamic systems, VLDB, 1998
[5] Grossman,D. and Frieder,O.: Information Retrieval – Algorithms and Heuristics, Kluwer Academic Press, 1998
[6] Guha, S., Rastogi, R. and Shim, K.: ROCK: A Robust Clustering Algorithm for Categorical Attributes, ICDE, 1999
[7] Jain,A.K., Murty, M.N. and Flynn, P.J.: Data Clustering: A Review, ACM Computing Surveys 31-3, 1999
[8] Kleinberg, J.M. : Authoritative Sources in a Hyperlinked Environment, JACM 46-5, 1999
[9] Stefanowski, J., Weiss, D.: Carrot2 and Language Properties in Web Search Results Clustering, Atlantic Web Intelligence Conference, 2003
[10] Zamir, O. and Etzioni, O. Web Document Clustering: A feasibility Demonstration, SIGIR, 1998
[11] Zipf, G, K.: The human behavior and the principle of least effort, Addison Wesley, 1949