

## Web 検索における情報統合化に関する研究

長尾 光悦 大内 東

北海道大学 工学研究科 複雑系工学講座

**あらまし:** 近年の Web サイトの急速な増加は、WWW 利用者が必要とする情報を簡便に収集することを困難にしている。本稿では、複数のサーチエンジンを利用することによって収集された情報を利用者が簡便に取捨選択することが可能な形式で提供する、メタサーチエンジンに基づく情報統合システムを提案する。提案システムにおいては、既存の 8 つのサーチエンジンを統合し、これらサーチエンジンが検索結果として出力する Web サイトの内容の分析を行う。この分析結果に基づき Web サイトを幾つかのクラスへ分類することによって効率的な情報の取捨選択を可能とする。また、WWW における実験を通して提案システムの有効性を検証する。

## A Study of Information Integration for Web Search

Mitsuyoshi NAGAO and Azuma OHUCHI

Research Group of Complex Systems Engineering,  
Graduate School of Engineering, Hokkaido University

**Abstract:** In this paper, we propose an information integration system for web search. The proposed system consists of a meta-search engine module and a clustering module. The eight conventional web search engines are combined for the meta-search engine module. The web information on WWW is collected by using this meta-search engine module. In the clustering module, the collected information is classified into some clusters on the basis of the web contents. The user can easily obtain the required information on WWW by using the proposed system. Moreover, we confirm the effectiveness of the proposed system through a computational experiment. The experimental results revealed that the effective information collection can be performed by using the proposed system.

### 1 はじめに

近年のインターネットの急速な普及に伴い、WWW(World Wide Web) 上の Web サイト数が急増している。これによって、WWW 利用者が膨大な Web サイト情報から必要とする情報を獲得することが困難になっている [1][2][3]。

WWW に蓄積されている情報の中から利用者が必要とする情報を獲得するための一般的な方

法としては、サーチエンジンを利用する方法が挙げられる。しかしながら、現在、WWW 上には多くのサーチエンジンが存在し、また各サーチエンジンが持つデータベースに蓄えられている Web サイト情報も異なることから、利用者が必要とする多種多様な情報を獲得するためには、複数のサーチエンジンを利用し情報収集を行わなければならない。この作業は利用者に対して非常に負担のかかる作業であり、多大な労

力と時間が必要とされる。

この問題を解決するために、複数のサーチエンジンにおいて同時に検索を実行し、一つの検索結果として出力するメタサーチエンジンが提案されている。既存のメタサーチエンジンでは各サーチエンジンが出力する URL(Uniform Resource Locator)リストを統合し利用者へ提供する。しかしながら、更新速度が速い現在の WWWにおいて効果的な情報提供を行うためには、実際の Web サイト内の情報に基づき、効率的な情報の取捨選択が可能となるような形式で情報提供を行う必要があることは明らかである。

本稿では、メタサーチエンジンに基づく情報統合システムを提案する。提案システムでは単に各サーチエンジンによって出力された URL リストを統合するのではなく、効果的な情報の取捨選択を行うことが可能となるよう、検索結果として得られた Web サイトの内容を分析し、幾つかのクラスタへ Web サイトを分類した形式で情報提供を行う。サーチエンジンの検索結果のみを用いるのではなく、実際の Web サイト内の情報を分析しクラスタリングを行うことによって、時々刻々と変化する WWWにおいて効果的に簡便な情報収集が可能となる。また、WWW における実験を通して提案システムの有効性を検証し、今後の課題を議論する。

## 2 Web 検索における情報統合

WWW 上の Web 情報に対して有効な情報統合システムを実現するためには以下の課題を解決する必要があると考えられる。

1. 分散的情報源: Web サイト情報は WWW 上に分散して存在する。また、サーチエンジンのデータベースに登録されている Web サイト情報も各エンジンで異なる。このような分散化した情報を効果的に収集することが必要とされる。
2. 情報の混在: Web サイトは情報提供者が HTML 言語を用いて自由に記述することができる。このため利用者が必要としない情報、例えば、広告等の情報が含まれる。

れる。すなわち、Web サイトは多大なノイズを含む情報源である。したがって、このようなノイズに対してロバストであり、かつ、各 Web サイトの内容を的確に把握することが可能な手法が必要とされる。

本稿では、これら課題に対して以下のアプローチを行う。

1. メタサーチエンジンに基づく Web サイト情報の収集: WWW 上の Web 情報を収集するために複数のサーチエンジンを統合したメタサーチエンジンの構築を行う。構築するメタサーチエンジンにおいてはディレクトリ型、ロボット型の二種類のサーチエンジンを採用し多種多様な情報を収集することを可能とする。
2. HTML タグに基づく Web サイト内容推定: HTML 言語によって記述される Web サイト情報は視覚的構造のみを与えられ、意味的構造を表現することは困難である。しかしながら、HTML 言語による視覚的構造から情報提供者の意図、すなわち、意味的構造が推測可能であると考える。よって、本研究ではノイズを含む Web サイト情報において的確な内容把握を行うために、Web サイトにおいて視覚的構造を付与するための HTML タグを利用した Web サイト内容推定を行う。
3. 形態素解析技術の利用: Web サイトから抽出される情報を全て用いる場合には多大な計算コストが必要とされる。よって、本研究では形態素解析技術を用い Web サイトの内容に関係する情報のみを抽出することによって、計算コストを抑え、かつ的確な内容推定を行うことを可能とする。

## 3 メタサーチエンジンに基づく情報統合システム

以下、本稿で提案するメタサーチエンジンに基づく情報統合システムの詳細について述べる。

### 3.1 システム構成

本システムの構成とデータの流れを図1に、インターフェースを図2に示す。図1に示されるように、本システムはWWW上の情報を収集するためのメタサーチエンジンモジュールと収集された情報を分析し利用者へ提供するクラスタリングモジュール及びWebサイト情報を表示するためのブラウザから構成されている。利用者はメタサーチエンジンモジュールに対して必要とする情報と関連があるキーワードを与える。メタサーチエンジンモジュールはWWW上に存在する複数のサーチエンジンに対してこのキーワードをクエリとして送信し、検索結果のURLリストを受け取る。これらURLを比較し、重複がある場合には除去を行う。重複するURLの除去後、各URLに基づきWebサイトの内容、すなわち、HTMLテキスト情報の獲得が行われる。メタサーチエンジンモジュールが獲得したHTMLテキスト情報はクラスタリングモジュールへ送られ、Webサイトの内容に基づくクラスタリングが行われる。

クラスタリングモジュールでは、HTMLテキスト情報からHTMLタグに基づきテキスト情報の抽出が行われる。抽出された情報に対して形態素解析が適用され、この結果に基づき各Webサイト間の類似度が算出される。算出された類似度に基づきWebサイトのクラスタリングが行われ、この結果が利用者へ提供される。以下、各モジュールの詳細について述べる。

### 3.2 メタサーチエンジンモジュール

本システムでは、分散化されたWWW上のWebサイト情報を収集するために既存のサーチエンジンを統合したメタサーチエンジンを用いる。本システムにおけるメタサーチエンジンモジュールでは、国内の代表的な8つのサーチエンジンを統合している。これらは、DOKO DA<sup>1</sup>、Excite<sup>2</sup>、goo<sup>3</sup>、Google<sup>4</sup>、Infoseek<sup>5</sup>、Ly-

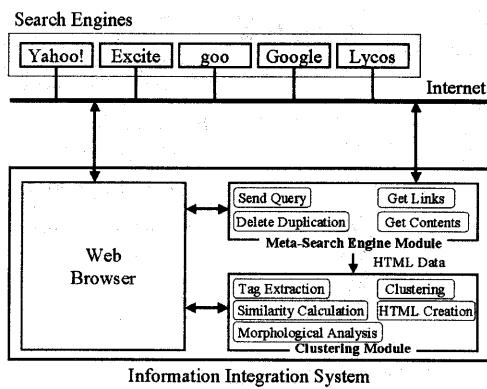


図1: システム構成

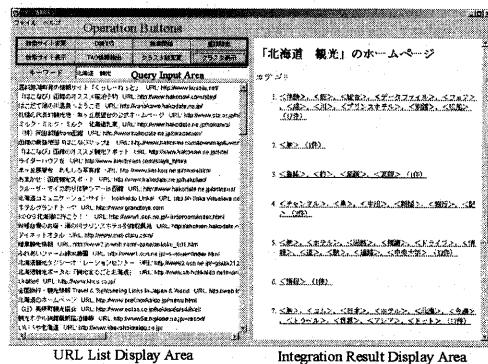


図2: インターフェース

cos<sup>6</sup>、Yahoo<sup>7</sup>、Freshey<sup>8</sup>の計8サーチエンジンである。これら8つのサーチエンジンの内6つがロボット型のサーチエンジン、2つがディレクトリ型のサーチエンジンである。また、検索方法も各サーチエンジンで異なることから、同一のキーワードをクエリとして与えた場合にも多種多様な情報が収集可能となる。

本システムにおけるメタサーチエンジンモジュールでは図3に示されるように、各サーチエンジンに対してURLアドレスとしてクエリを送信する。メタサーチエンジンモジュールは送信

<sup>1</sup><http://www.dokoda.com>

<sup>2</sup><http://www.excite.co.jp>

<sup>3</sup><http://www.goo.ne.jp>

<sup>4</sup><http://www.google.com>

<sup>5</sup><http://www.infoseek.co.jp>

<sup>6</sup><http://www.lycos.co.jp>

<sup>7</sup><http://www.yahoo.co.jp>

<sup>8</sup><http://www.fresheyeye.co.jp>

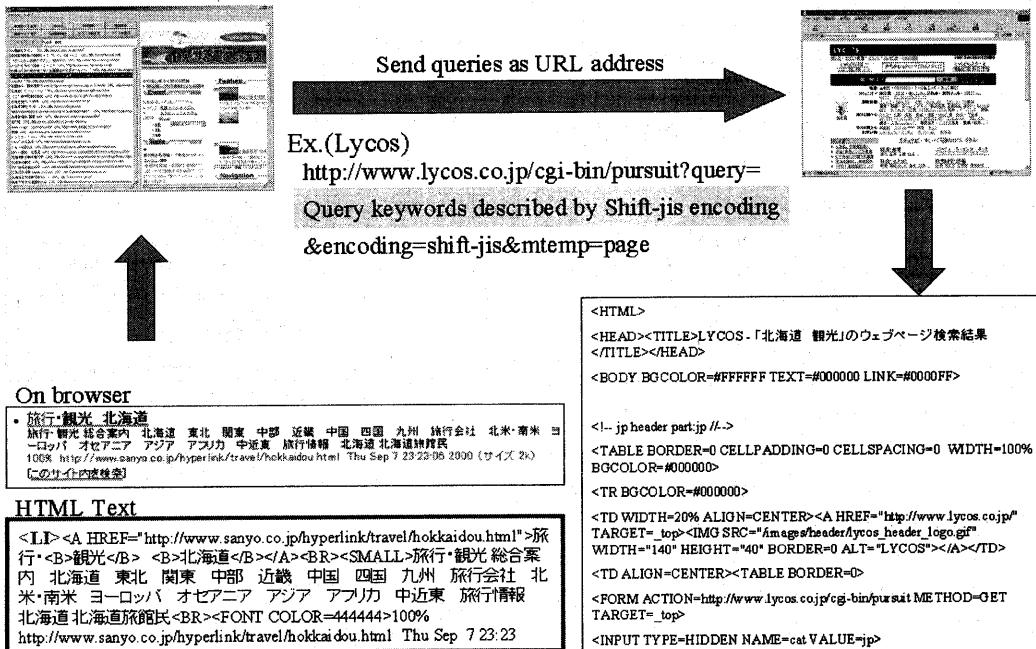


図 3: URL アドレスに基づく Web 検索

したクエリに対する検索結果を HTML テキスト形式で受け取る。この HTML テキスト情報から検索結果としての URL をタグ情報や画像等の各サーチエンジンに対応した特徴を利用して抽出する。抽出された URL はインターフェースにおける URL アドレス表示領域に表示される。また、URL によって示される各 Web サイトのコンテンツが HTML テキスト形式で獲得される。これら各 Web サイトの HTML テキスト情報をクラスタリングモジュールへと送信する。

### 3.3 クラスタリングモジュール

クラスタリングモジュールでは、メタサーチエンジンモジュールから送られた HTML テキスト情報に基づき、Web サイトのクラスタリングが行われる。以下に詳細を示す。

#### 3.3.1 クラスタリングプロセス

以下に提案システムにおけるクラスタリングプロセスを示す。

1. HTML テキスト情報からの HTML タグに基づくテキスト情報抽出
2. 抽出された情報に対する形態素解析の適用
3. 名詞頻度ベクトルの生成
4. Web サイトの名詞頻度ベクトル間の類似度算出
5. 類似度に基づくクラスタリング

クラスタリングモジュールでは、第一にメタサーチエンジンモジュールから送られた HTML テキスト情報から HTML タグに基づきテキスト情報の抽出が行われる。抽出されたテキスト情報に対して形態素解析が適用され、HTML タグに基づく名詞句情報が獲得される。獲得された名詞句情報を用いて名詞頻度ベクトルを生成する。この名詞頻度ベクトルが Web サイトの内容を表す情報として扱われる。次に、名詞頻度ベクトル間の類似度が算出され、算出された類似度に基づき Web サイトのクラスタリング

が行われる。以下に各ステップの詳細について述べる。

### 3.3.2 HTML タグに基づく情報抽出

Web サイト情報は、提供者が自由に情報発信を行えるために広告等の利用者が必要しない情報も含まれる。すなわち、Web サイト情報は多量のノイズ含む情報源である。しかしながら、Web サイトの内容に基づく情報統合を行うためには、このようなノイズを含む情報において的確に Web サイトが表す内容を把握する必要がある。このため、提案システムでは、Web サイトの内容を的確に把握するために HTML 言語において用いられる HTML タグに基づく Web サイトの内容推定を行う。HTML タグは Web サイト情報を視覚的かつ効果的に利用者へ提供するための言語記述方式である。この記述方式では、開始タグと終了タグと呼ばれるタグ間にテキストを包含することによってテキストに対して視覚的な構造を付与する。情報提供者は利用者へ効果的な情報伝達を行うために HTML タグを用いる。すなわち、HTML タグを利用するによって HTML 言語では本来扱うことが困難なテキスト情報における意味的構造を推定することが可能であると考えられる。

本研究では、この HTML タグに基づく Web サイト内容推定において図 4 に示す二種類のタグを採用する。一つはタイトルタグ（<TITLE>）と呼ばれるタグであり、Web サイトをブラウザにおいて表示した際にブラウザのフレーム部分に表示されるテキストを決定するためのタグである。更に、リンクタグ（<HREF>）と呼ばれるタグを採用する。リンクタグは指定された URL への移動を行うために用いられる。タイトルタグは Web サイト内で一つ出現し、Web サイトを表示した際のフレーム部分に表示されるテキストを指定するタグであることから、Web サイトの概略を表すテキスト情報が含まれると考えられる。また、リンクタグはその Web サイトから新たな情報が記述されるサイトへの移動を意図していることから Web サイトに含まれる詳細情報に関するテキストが含まれると考えられる。

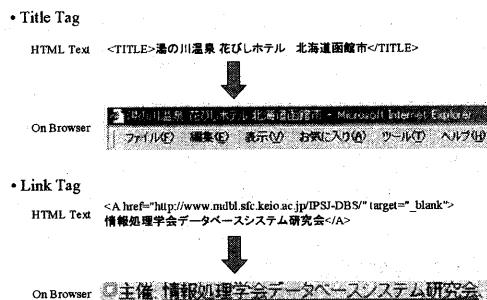


図 4: HTML タグ

本システムでは、この二つのタグ内に含まれるテキスト情報を個別に抽出し、以下の類似度算出及びクラスタリングにおいて用いる。

### 3.3.3 形態素解析

類似度の算出において、HTML タグに基づき抽出されたテキスト情報を全てを利用した場合には計算コストが多大であり、更に、テキストにノイズが含まれると考えられるため、適切な類似度の算出が困難になると考える。よって、本研究では HTML タグに基づき抽出したテキスト情報を対して形態素解析を適用することによってこれら問題の解決を行う。形態素解析は、近年、自然言語処理や情報検索などの分野で頻繁に用いられており、その有効性も確認されている [4]。本研究では形態素解析のために「茶筌」<sup>9</sup>を用いる。茶筌によって分解されたテキスト情報の中から名詞句のみを抽出する。次に、各名詞句の出現頻度を算出し、名詞頻度ベクトルを生成する。但し、名詞頻度ベクトルにおいてはメタサーチエンジンモジュールに与えたキーワードは含めないものとする。これは、サーチエンジンによって検索される Web サイトに検索キーワードが含まれる可能性は高く、このキーワードを類似度の算出に用いた場合、クラスタリングに悪影響を及ぼす可能性が考えられるためである。また、この名詞頻度ベクトルは出現頻度順にソートされる。類似度の算出においては、名詞頻度ベクトルにおいて高頻度の名詞句が用いられる。

<sup>9</sup><http://chasen.aist-nara.ac.jp>

高頻度の名詞句のみを類似度算出において利用することによってノイズに対してロバストであり、かつ計算コストを抑えた類似度算出が可能となる。

この操作によって、各 Web サイトにおけるタイトルタグに基づく名詞頻度ベクトル及びリンクタグに基づく名詞頻度ベクトルが生成される。

### 3.3.4 類似度算出方法

形態素解析に基づき算出された名詞頻度ベクトルを用いて各サイト間の類似度を算出する。この類似度の算出においてはリンクタグの名詞頻度ベクトルを用いる。リンクタグにおける名詞頻度ベクトルを用いることによって、Web サイト内に含まれる詳細レベルでの内容の類似度を算出する。これによって、的確な Web サイトの内容に基づくクラスタリングを実現可能にする。以下に名詞頻度ベクトルに基づく類似度の算出方法を示す。

$$S(N_x, N_y) = \frac{N_x \cdot N_y}{|N_x||N_y|} \quad (1)$$

$S(N_x, N_y)$  : 名詞頻度ベクトル  $N_x$  と  $N_y$  の類似度

$N_x$  : Web サイト  $x$  におけるリンクタグ内テキストの名詞頻度ベクトル

図 5 に式 (1) に基づく名詞頻度ベクトル間の類似度算出例を示す。本研究で用いる類似度算出式では、Web サイトが持つ情報量、すなわち、名詞句数が考慮される。これは、少ない情報量の Web サイト間で一致する名詞句が多い場合には両 Web サイト間にはノイズが少なく、Web サイトの内容が限定された中で一致している可能性が高いという仮定に基づくものである。また、本システムでは類似度算出に用いる名詞句数の上限を指定することが可能となっている。

### 3.3.5 クラスタリング方法

算出された類似度に基づき各 Web サイトのクラスタリングが行われる。クラスタリング方法としては類似度が最大の Web サイトを結合する最短距離法を採用する。但し、本研究にお

名詞頻度ベクトル  $N_x$ : 函館-4, データ-4, 工学-2, 研究会-2  
名詞頻度ベクトル  $N_y$ : 電子-3, 工学-2, 通信-2, 情報-1, 研究会-1

	函館	データ	工学	研究会	電子	通信	情報
$N_x$	4	4	2	2	0	0	0
$N_y$	0	0	2	1	3	2	1

$$S(N_x, N_y) = \frac{N_x \cdot N_y}{|N_x||N_y|} = \frac{4+2}{\sqrt{40}\sqrt{19}} = 0.218$$

図 5: 類似度算出例

いては、ある名詞頻度ベクトル  $N_x$  と  $N_y$  が結合されたクラスターにおける名詞句ベクトルは、 $N_x$  と  $N_y$  における名詞句を結合させ、この中で高頻度の名詞句を用いて生成されたものを用いる。また、この操作によって生成されたクラスターの名前がクラスター内の Web サイトが持つタイトルタグのための名詞頻度ベクトルにおける高頻度名詞句を利用して付与される。更に、本システムではクラスタ数は利用者が指定することが可能となっている。

## 4 計算機実験

提案システムを評価するため WWW において実験を行った。以下に実験方法及び実験結果を示す。

実験では、検索キーワードとして「北海道」及び「観光」の 2 つを用いたアンド検索を行った。各サーチエンジンの検索結果の中から上位 20 サイトを採用し、重複除去後の計 142 サイトを用いてクラスタリングを行った。類似度の算出においては、出現頻度の高い名詞句の上位 10 個を用いた。また、クラスタ数としては 10 クラスタを生成するものとした。

実験結果を図 6、図 7、図 8 に示す。図 6 は提案システムにより分類された各クラスター内に含まれた Web サイト例である。図 6 では、各クラスター内に含まれた Web サイトをブラウザ上で表示している。図 6 において、クラスター 5 に含まれた Web サイトは、デッドリンクとなっている Web サイト（図 6 中）や画像のみによって構成される Web サイト（図 6 右）であり、検索キーワードに関連した情報の収集が困難な Web

サイトであった。しかしながら、クラスタ5には「北海道」及び「観光」の検索キーワードに対して適切であると考えられるWebサイト(図6左)も含まれた。クラスタ5におけるこのようなWebサイトはフレーム構造を用いたWebサイトであった。フレーム構造を用いたWebサイトの場合には、サーチエンジンの検索結果としてリンクされているWebサイトにはフレームの構造の情報のみが含まれるため検索キーワードと関連する情報は含まれない。したがって、このようなクラスタリング結果になったと考えられる。これは、図7に示される各クラスタにおけるリンクタグに基づく名詞頻度ベクトル情報からも確認された。また、クラスタ9では「北海道」及び「観光」に関係するリンクが含まれるもの全体としては検索キーワードと関連が薄いと考えられるWebサイトが含まれた。これは図7からも示された。一方、クラスタ10に含まれたWebサイトでは検索キーワードに関連したリンクが含まれ、かつ、利用者が求める情報が多く含まれていた。これは図6におけるブラウザ上のWebサイト、及び図7における各クラスタにおけるリンクタグに基づく高頻度名詞句からも明らかである。

また、図8に各クラスタ内のWebサイト間の類似度を示す。上記の実験結果とこの結果からクラスタ5にはリンク情報が存在しないWebサイトが集められ、クラスタ9には検索キーワードと関連が薄いWebサイトが存在していることがわかる。また、クラスタ10には検索キーワードと関連したWebサイトが集められていることが明らかとなった。

これらの結果を総合して、提案システムを用いることによってWWW上で利用者が簡便な情報の収集を行うことが可能であることが示された。

しかしながら、本研究ではクラスタ名としてタイトルタグに基づく名詞句を利用している。実験結果からタイトルタグにおける名詞句が検索キーワードと関連するが、リンクタグにおける名詞句が検索キーワードと関連しない場合があることが明らかとなった。このような場合には、利用者の取捨選択の効率を低減させると考えられるため改善を行わなければならない。こ

れは今後の課題である。

## 5 おわりに

本稿では、WWWにおいて効率的な情報収集が可能となるよう、メタサーチエンジンに基づく情報統合システムを提案した。提案システムはメタサーチエンジンモジュールとクラスタリングモジュールから構成されている。メタサーチエンジンモジュールでは、複数の既存のサーチエンジンを統合し、WWW上に存在する多種多様な情報を獲得することが可能となっている。また、クラスタリングモジュールでは形態素解析技術を用い、メタサーチエンジンモジュールが収集したWebサイトの内容に基づくクラスタリングを可能としている。実験結果から提案システムを利用することによってWWW上で有効な情報収集を実現可能であることが示された。しかしながら、実験結果から現在のシステムでは、クラスタ名の決定方法、フレーム構造を持つWebサイトへの対応等において改善が必要であることも示された。これら問題点の解決は今後の課題である。

## 参考文献

- [1] 河野浩之、川原稔：Web検索におけるテキストマイニング、人工知能学会誌、Vol.16, No.2, pp.212-218 (2001)
- [2] 渡部勇：ビジュアルテキストマイニング、人工知能学会誌、Vol.16, No.2, pp.226-232 (2001)
- [3] 坂本比呂志、有村博紀：Webマイニング、人工知能学会誌、Vol.16, No.2, pp.233-238 (2001)
- [4] 那須川哲哉、河野浩之、有村博紀：テキストマイニング基礎技術、人工知能学会誌、Vol.16, No.2, pp.201-211 (2001)



図 6: クラスタリング結果例

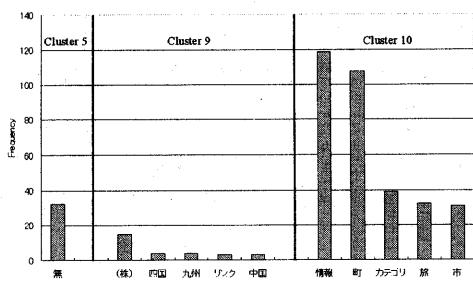


図 7: 各クラスタにおけるリンクタグに基づく名詞頻度ベクトル（高頻度上位 5 名詞）

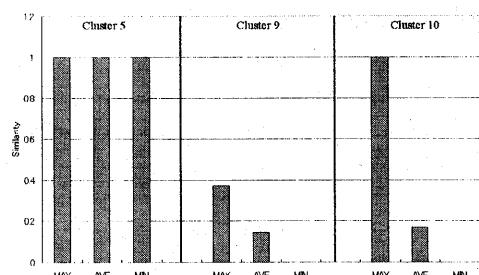


図 8: 各クラスタ内の類似度