DEWS2008 E4-3

コミュニティ構築型辞書グラフを用いた異種メディアデータ連結機構の実現

佐藤　篤† 倉林　修一†† 清木　康†

† 慶應義塾大学環境情報学部
†† 慶應義塾大学大学院政策・メディア研究科
E-mail: †{t04561as,kurabaya,kiyoki}@sfc.keio.ac.jp

あらまし 本稿では，Web 2.0 環境において，コミュニティにより構築された知識である単語辞書を辞書グラフとして構築し，これを用いて，画像データやブログ記事など，異種メディア間の意味的関連性を計量することにより，相関の高いメディアデータ間の結合を行う方式，および，その実現システムを示す．本方式の特徴は，コミュニティにより構築された単語辞書におけるハイパーリンク構造を再帰的にトラバースし，単語間の意味的な関連性を，ホップ数を用いた距離算出により計量する点にある．メディアデータに関連付けられたすべての単語を対象として本方式による意味的距離計算を適用することにより，強い文脈的関連性を有するメディアデータの組み合わせを抽出することが可能となる．また，本方式を実装したシステムにより，次の応用システムの実現例を示す．1) 文書データに対する抽象アイコンの付与による文書データの一覧性の向上．2) 関連異種メディアデータの同時的な表示による新たなマルチメディア表現の実現．本稿では，評価実験により，実現方式による異種メディア統合機構の実現可能性，有効性を示す．

Heterogeneous Media Data Linkage Mechanisms Using Online Dictionaries Created by Community

Atsushi SATO†, Shuichi KURABAYASHI††, and Yasushi KIYOKI†

† Keio University, Faculty of Environmental Information
†† Keio University, Graduate School of Media and Governance
E-mail: †{t04561as,kurabaya,kiyoki}@sfc.keio.ac.jp

Abstract The emergence of Web2.0 has created a large crowd of multimedia data which are dynamic and ever-changing. We present a novel media-integration mechanism that calculates implicit relationships between heterogeneous media data. This mechanism includes a function to calculate implicit relationships between two metadata words by recursively traversing hyper-link structure of onlie dictionary generated by online-community. We also present two application systems of our method. 1) A document viewer that associates icon images, which represent the content of associated document, with document. 2) A new multimedia infrastructure that embeds heterogeneous media data into a single multimedia document. We show several experimental results to clarify the feasibility and effectiveness of our method.

1. はじめに

インターネットの広帯域化や Web 2.0 と呼ばれるユーザ主導によるメディアデータ構築の普及とともに，ネットワークを介して多数のユーザがオンラインコミュニティを形成し，協調して共有可能なメディアデータを構築している．ブログ記事，SNS，そして Wikipedia に代表されるこれらのメディアデータ群は，任意のユーザによるデータのアップロード機能，ソーシャルタギング機能，トラックバック送信機能等のユーザ参加型機能を有するシステム上に構築されており，ユーザが容易にコミュニティに参加しメディアデータを構築することを可能にしている．散在するこれらのメディアデータ群を対象として，メディアデータ間に内在する意味的関連性を計量し，それらを動的に関連付けてユーザに提示することができれば，コンテンツの活用範囲は拡大する．一方，Web 2.0 サイト群には，メディアデータだけでなく，それらメディアデータを説明するた

めの語彙を，オンラインコミュニティにより定義する機能を有するものがある．例えば，コミュニティにより構築された日本語の辞書として有名な「はてなダイナリーキーワード」[3] は，2007 年 1 月 15 日現在において，214,294 語が登録されており，ひとつの単語を記述するページは平均して，おおよそ 18 単語へのハイパーリンクを有することが確認されている．これらオンラインコミュニティにより定義された単語辞書（以下 “Web 単語辞書”）は，一般的な辞書と異なりオンラインコミュニティにおける最新の語彙や文脈を反映しており，時々刻々と変化を続けている．このため，Web2.0 環境におけるメディアデータ間の意味的関連性の計量をするためには，絶え間なく変化するこれら Web 単語辞書を知識として用いることが重要と言える．

そこで，本稿では，Web 単語辞書を用いて，同じくオンラインコミュニティにより構築されたメディアデータ間の関連性を計量することにより，異種メディアデータの動的な関連付けを行う異種メディアデータ連結機構を示す．異種メディアデータ連結機構は，Web 単語辞書を辞書グラフとして構築し，これを用いて，アイコン画像やブログ記事といった異種メディア間の意味的関連性を計量することにより，相関の高いメディアデータ間の結合を行う．本方式の特徴は，Web 単語辞書における単語間のリンク構造を再帰的にトラバースし，単語間の意味的な関連性を，リンクホップ数を用いた距離算出により計量する点にある．本システムは，次の 4 機能により特徴づけられる．

- **機能１ 辞書グラフ構築機能：** Web 単語辞書に蓄積された断片的な単語間の関連を表すリンク構造を再帰的にトラバースすることにより，辞書グラフとして形成する機能．
- **機能２ メディアデータを対象としたメタデータ生成機能：** Web 上のメディアデータを対象として，文書内の単語をメタデータとして抽出し付与する機能．本機能は Web 単語辞書内に存在する単語をメタデータとして付与することにより実現する．
- **機能３ 異種メディア関連性計量機能：** 機能 1 により構築された辞書グラフを用いて，機能 2 により生成されたメタデータを利用し、異種メディア間の関連性を計量する機能．
- **機能４ 異種メディアデータを対象とした文脈的連結機能：** 異種メディアデータ間の関連性を機能 3 により算出し，ユーザが指定した閾値以上の相関量を有するメディアの対を「連結メディア」として出力することによって，異種メディアの文脈的連結を行う機能．

図 1　ブログ記事とアイコン画像の連結アプリケーション例

本システムを既存のブログなどのメディアデータに適用することで，自動的にアイコン画像や背景画像を関連付けるアプリケーションを実現することが可能である（図 1）．我々は，本システムの応用として，ブログ記事などの文書メディアを対象としてアイコン画像を付与するアプリケーションを実際に構築した．このアプリケーションは，ユーザが記述する記事を対象に，テキスト内容と意味的に関連性の強いアイコン画像を抽出し，ブログ記事一覧を装飾するアイコンとして自動的に付与するものである．本方式は，関連異種メディアデータの同時的な表示による新たなマルチメディア表現を実現するものである．本稿では，評価実験により，本方式の実現可能性，および，有効性を明らかにする．

2. 提 案 方 式

本方式におけるメタデータは，１メディアデータについて，1 つ以上の単語郡で表現されるものを対象とする．単語郡をメタデータとして扱う本方式は，機能 2 による文書データを対象とした自動メタデータ抽出生成のほか，メタデータを単語郡として表現する既存自動抽出方式や人手によるタギングといった，Web2.0 環境における包括的なメタデータ付与方式を対象としている．

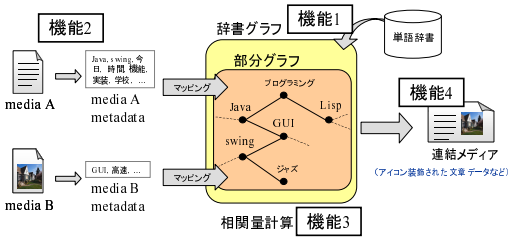

図 2　システム概要図

本方式を実現するシステムの概要図を図 2 に示す．本システムは，メディアデータに関連付けられたメタデータである単語郡間の関連を，Web 単語辞書を用いて計量する．本節では，異種メディアデータ連結機構における，Web 単語辞書を意味的関連性を計量する空間として形成する手法，および，形成した計量空間上において，上記のメディアデータ間の関連性を計量する手法について述べる．本方式では，Web 単語辞書における単語定義文書から別の単語へのリンク構造を，“単語ノード”，“意味の関連エッジ” とする重み無し無向グラフとして形成する．本稿では，このグラフを **“辞書グラフ”** と呼ぶ．なお，グラフ上での路数の単位を “ホップ” と呼ぶものとする．

本方式は，辞書グラフから，単語間の意味的な関連性を計量するための**単語距離マトリクス**を生成する．単語距離マトリクスは，二つのメディアデータ間のグラフ上の距離を各単語メタデータについて計量し，テーブルとして表現したものである．本方式は，絶え間なく追記され，無限に近い広がりを有する辞書グラフを用いて単語間の意味的な距離を計量可能とするために，２つのメディアデータが有するメタデータ単語から n ホップで到達可能なノードにより構成される部分グラフを抽出し，二つのメディアデータ間の関連性を，部分グラフ上の単語間の

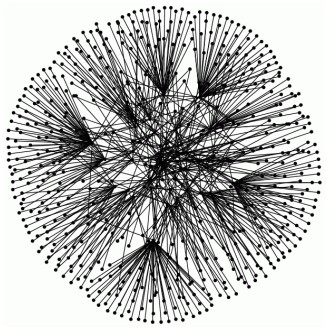

図 3　単語ノード “Java”（図中央）から 2 ホップ分探索した場合の部分グラフ

距離の総計として計量する．すなわち，単語距離マトリクスは，辞書グラフにおける非線形的な意味の広がりに対し，辞書グラフ内における 2 単語間の距離を線形に扱うことの可能な相関量に変換するためのデータ構造として位置付けることができる．本方式における単語間の意味的相関量は，この単語距離マトリクスを対象とした演算として実現される．すなわち，単語距離マトリクス上の列，および，行に対して集約演算を適用することにより，二つのメディア間の意味的関連性を計量する．

本方式の特徴は，ユーザの目的に応じて，マトリクス上の集約演算を選択することにより，それぞれの単語間の距離の影響度を制御可能にする点にある．これにより，メディアが有するすべての単語間の意味が近い場合，あるいは，メディアが有する単語のうち，一つでも意味が近い場合，など，異種メディア連結の目的に応じて，様々な意味的距離を算出することが可能である．本方式は，辞書グラフという膨大な広がりを持つグラフ空間を用いるが，異種メディア間の距離を，各メディアが有するメタデータが辞書グラフ上で有する距離（路数）の集約として計量するため，グラフにおける意味の発散が問題とならない．例えば，Web 単語辞書である “はてなダイアリーキーワード” [3] から本方式における辞書グラフを抽出する場合を考えると，2008 年 1 月 15 日現在において “Java” いう単語から 2 ホップ分のみの探索で，約 700 のノードが検出される（図 3）． 図 3 は “Java” という単語ノードから 2 ホップ分，辞書グラフ内を探索した際の，部分グラフのノードとエッジを図示したものである．1 ホップの距離にあるノードは “Swing”，“Sun Microsystems” といった “Java” と意味的に近いものが多いが，2 ホップの距離にあるノードは “コンピュータ” といった比較的意味の近いものはあるものの，“衛星放送” といった意味の遠いものが多く存在する．本方式は，意味的に関連の低いノイズ単語と適切な単語が混在した 2 ホップ距離の単語ノードを用いて意味を計量した場合においても，マトリクス上の距離演算による行・列の集約に和演算を用いることにより，ノイズ単語の影響を低減することが可能である．また，辞書グラフを 1 ホップのみで構築した場合のマトリクス上の距離演算においても，行・列の集約に積演算を適用することで，全ての単語間の距離が近いこと判定できる演算を行うことが可能である．

2.1　データ構造

ここでは，辞書グラフのデータ構造，および，辞書グラフより自動的に生成される単語距離マトリクスのデータ構造について述べる．

2.1.1　辞書グラフ

単語をノード，単語間の意味の関連をエッジとする無向重みなしグラフである．Web 単語辞書に登録されている 1 単語を単語ノード，単語定義文書に登場する別単語へのリンクを，別単語ノードと繋ぐエッジとして解析することにより形成される．2 つのノード間の距離 L は，2 つのノードを結ぶ最短経路長を用いる．なお，同じノード間の距離を 0，到達不可能なノード間の距離を無限大とする．

2.1.2　単語距離マトリクス

2 つのメディアデータに付与されているメタデータを行・列の属性としたテーブルデータ構造である（表.1）．

表 1　単語距離マトリクス

		Media M_B metadata		
		m_1	m_2	... m_m
Media M_A metadata	t_1			
	t_2			
	t_3			
	$...t_n$			

2.2　基 本 機 能

2.2.1　機能 1. 辞書グラフ構築機能

本機能は，Web 単語辞書を解析し，辞書に登録されている単語をノード，その単語と単語の定義文章に登場する単語との関係を，エッジとした無向重みなしグラフである “辞書グラフ” を形成する機能である．

本方式では，この辞書グラフを計量空間として用い，異種メディアの意味的関連性を計量する．任意のメディアデータ 2 つの意味的関連性の計量は，この辞書グラフの部分グラフのみを計量毎に形成して用いることにより実現する．これにより，時々刻々と変化を続ける Web 単語辞書を常に最新の状態で計量空間として利用することが可能である．

以下に，二つのメディアデータ M_A, M_B の関連を計量するための部分グラフ G' を生成する関数 $f_{subgraph}$ を示す．$f_{subgraph}$ は，M_A, M_B に関連付けられたメタデータである単語の和集合をノードとし、それらの各ノードから有限路数分エッジを持つ部分グラフ G' を全体グラフ G から抽出する．

$$f_{subgraph}(M_A, M_B, G) \rightarrow G'$$

2.2.2　機能 2. メディアデータを対象としたメタデータ生成機能

本機能はメディアデータに応じた方式によりメタデータ生成を行う機能である．1 メディアデータについて，メタデータは，1 つ以上の単語で表現され，各単語にはメディアの重要度を重みとして付与したものを生成する．メタデータは文章中から Web 単語辞書に掲載されている単語を抽出し，出現頻度を重要度とすることにより生成する．ブログ記事などの文書メディアは内容文章を対象として抽出し，アイコン画像などの文章を含まないメディアは関連文章を対象として抽出を行う．この方式

により生成された単語メタデータは，辞書グラフにおいて単語ノードとして存在しているため，辞書グラフをメディアの関連性を計量する空間として利用する本方式において，メタデータを計量空間へのマッピングにすることに適した方式と言える．

上記以外のメタデータ抽出方式として，“del.icio.us(http://del.icio.us/)” に代表されるソーシャルブックマーク・サービスにおける「タグ」情報のように，人手によりメタデータを付与する方法や，文書メディアを対象として形態素解析と tf-idf アルゴリズム [2] を用いて求めた重要語をメタデータとする方法，静止画像の色彩情報を対象としたメタデータ自動抽出方式 [1] により，映像の色彩情報を印象語に関連づける手法などがある．本方式では，Web 単語辞書を意味的相関量の計量空間として用いているため，メタデータが人手により付与されたもか自動生成されたかにかかわらず，Web 単語辞書中に登録されている単語であれば意味計量の対象とすることができるため，メタデータの抽出方式に依らず適用可能である．

2.2.3 機能 3. 異種メディア関連性計量機能

本機能はメタデータが付与された，互いに異種であることを認める，メディア M_A と関連性計量対象メディア M_B 間の関連性を部分グラフ G' 上において計量する機能である．本機能 $f_{distance}(M_A, M_B, G')$ を次のように定義する．

$$f_{distance}(M_A, M_B, G') \rightarrow corelation(M_A, M_B)$$

ただし，M_A: メディア A，M_B: 関連性計量対象メディア B，G : 辞書グラフ，t_i: メディア A に付与されている i 番目のメタデータ，m_j: メディア B に付与されている j 番目のメタデータとする．関連性計量処理の各ステップを次に示す．

Step-1．距離計量関数： メディアデータに付与されている各メタデータと，対象メディアデータに付与されている各メタデータ間について，辞書グラフ上での距離を求める．ノード間の距離は，同じノードの場合を 0，経路無しを無限大として扱う．ここで，辞書グラフ G におけるメタデータ t_i, m_j の距離を $L(t_i, m_j, G')$ とする．表 2 に，各単語間の距離を計量した結果を示す．

表 2　路数計測結果例

		Media M_B metadata		
		m_1	m_2	... m_m
Media M_A metadata	t_1	2	1	∞
	t_2	1	0	∞
	t_3	2	1	∞
	$...t_n$	∞	∞	1

Step-2．メタデータ間の距離を関連値へ変換： Step1 で計測したグラフ上の単語間の距離を単語間の関連値として変換する．Step1 で求めた値はグラフ上における単語ノード間の路数であり，単語間の意味の関連を示す値ではない．また，辞書グラフにおける路数は，始点とする単語ノードから離れるほど意味の関連性が 0 に収束するように減衰すると考えられる．そこで，この Step では，書辞書グラフの特性に応じて距離変換関数を提供し，単語ノード間の路数を単語間の関連値として変換する．距離を x としたときの関連値 $f(x)$ を求める関数として次の 2 関数を提供する．なお，関数は本方式の実装が対象とする Web 単語辞書の特性に応じて設定するものとする．

- 関数 1 $f(x) = 1/(x+1)$
- 関数 2 $f(x) = e^{-x}$

変換関数 1 を用いた場合，関連値 $L'(t_i, m_j)$ は次の式で表される．

$$L'(t_i, m_j) = 1/(L(t_i, m_j) + 1)$$

表 3　関数 1 を用いた関連値への変換結果例

		Media M_B metadata		
		m_1	m_2	... m_m
Media M_A metadata	t_1	1/3	1/2	0
	t_2	1/2	1	0
	t_3	1/3	1/2	0
	$...t_n$	0	0	1/2

Step-3．重み付け： Step-2 で求めた関連値に，メディアに付与されている各メタデータの特徴量を表す重みを加える．各メディアデータについて，メタデータの特徴量の合計が 1 になるように 1 ノルムで正規化したものを重みとし，Step-2 での関連値に乗算する．メタデータ t_i, m_i の特徴量を $M(t_i)$， $M(m_i)$ としてもつ場合の，各メディアデータの重み付けされた相関量 $W(t_i, m_j)$ は次の式で求められる．(表 4 の各メタデータの重要度ディアデータ数は $M(t_1) = 3, M(t_2) = 3, M(t_3) = 1, M(t_n) = 1$ である)

$$W(t_i, m_j) = L'(t_i, m_j) * |M(t_i)| * |M(m_j)|$$

表 4　重み付け結果例

		Media M_B metadata		
		m_1	m_2	... m_m
Media M_A metadata	t_1	0.15	0.22	0
	t_2	0.15	0.30	0
	t_3	0.05	0.07	0
	$...t_n$	0	0	0.07

Step-4．集約： Step-3 で求められた各メタデータ間の関連値をメディアデータ間の相関量として集約する．ここで，集約方式として次の 4 つの演算を用意し，本方式の応用システムで扱うメディアデータに応じて演算を提供することを可能とする．行の集約に和演算を適用する場合，メディア M_A の 1 メタデータが対象メディア M_B のすべてのメタデータに対して有する距離の和を計量するため，著しく距離の遠いメタデータ単語がある場合でも，その影響を受けない．すなわち，OR 条件のような特徴を持つ．一方，行の集約に積演算を適用する場合，メディア M_A の 1 メタデータがメディア M_B のすべてのメタデータに対して有する距離の積を計量するため，すべての単語間の距離が近い場合にもに高い相関量を示す．すなわち，AND 条件のような特徴を持つ．同様に，列の集約では，行の集約におけるメディア M_A とメディア M_B の関係の逆のことが言える．集約方式と特性を以下に示す．

（1） 行・列和演算：メディア M_A のメタデータとメディア M_B のメタデータの関連を OR 条件で集約する．すなわち，高い相関量を有するメタデータ・ペアの相関量が累積する．

（2） 行・列積演算：メディア M_A のメタデータの全てがそれぞれ，メディア M_B のメタデータ全て関連があるときにのみ高い相関量を示す．すなわち，AND 条件を意味する．

（3） 列積演算・行和演算：メディア M_A のメタデータ全てが，メディア M_B のメタデータと関連があるものの，メディア M_B のすべての特徴がメディア M_A と相関量を持つ必要はない場合に用いる．これは，メディア M_B のメタデータの数が多く，散漫である場合に有効である．

（4） 行積演算・列和演算：メディア M_B のメタデータ全てが，メディア M_A のメタデータと関連があるものの，メディア M_A のすべての特徴がメディア M_B と相関量を持つ必要はない場合に用いる．これは，メディア M_A のメタデータ数が多く，散漫である場合に有効である．

2.2.4 機能 4. 異種メディアデータを対象とした文脈的連結機能

機能 3 によりもとめられた関連性の強い異種メディアを，メディアに応じた形式で，同一のメディアに埋め込むことによって，異種メディアの文脈的連結を行う．連結を行う際，メディアデータのジャンルを自動的に判定することで，同一ジャンルに属するメディアデータのみを連結する．メディア M_A が属するジャンル t を判定する関数 $f_{genre}(M_A, \mathcal{M}_\mathcal{B}, G')$ を次のように定義する．

$$f_{genre}(M_A, \mathcal{M}_\mathcal{B}, G') \rightarrow t$$

ここで，$\mathcal{M}_\mathcal{B}$ は，メディアデータの全集合を意味する．関数 f_{genre} は，メディアデータ M_A と全ての対象メディアデータ $\forall\mathcal{M}_\mathcal{B}$ との相関量を計量し，各ジャンルにおける相関量の平均を求め，最も平均相関量の高いジャンルを該当対象メディアデータのジャンルとして採用する．これは異種メディア連結の際，同一ジャンルに属する対象メディアデータのみを連結することで，精度の高い連結が可能となると考えたためである．表 5 はあるブログサイト “miyu’s diary”(注1)の記事を主観によりジャンル分類したものと，本方式を実装したシステムにより自動的に判定したジャンルを比較した表である．このように，システムによる自動ジャンル分類が，ある程度の精度をもって実現されていることがわかる．このように，対象メディアデータに関連付けられたジャンル，および，メタデータを用いた自動ジャンル分類機能を適用することで，人手によるメディアのジャンル分類を行わずともメディアデータの連結が行えることが分かる．

表 5 miyu’s diary システムによる自動ジャンル分類評価表

タイトル	主観による ジャンル分類	システムによる ジャンル分類
涙…	pet,entertainment	movie
valentineday ☆	event,school	sports
お知らせ♪	**book**	**book**
オスだけど…	**fashion**,event	**fashion**
久しぶりに☆	event,**groumet**	**groumet**
あかね空	**movie**	**movie**
ドラマ撮影アップ！	movie,**school**	**school**
ミニミニ♪♪	**entertainment**	**entertainment**
ロケジャン★	**movie**	**movie**
ぽかぽか (嬉’V｀*) ⌒★	entertainment	music
冷え冷え	season	event
おはよッ	season	fashion
ドラマ撮影ッ☆	entertainment	event
ピチレモン撮影☆	**entertainment**	**entertainment**
衣装合わせッ♪	entertainment	sports

3. プロトタイプシステムの実装

本節では本方式の有用性を示すプロトタイプの実現として，ブログ記事とアイコン画像といった異種メディア間の関連性を計量し，関連性の高いメディアどうしの連結を行うことによる，異種メディアの連結を装飾として応用したシステムを実現する．本システムは，RSS により取得されたブログテキストを対象に，予め用意されたアイコン群との相関量を計量することで，ブログサイトの記事一覧の自動装飾を行うシステムである．本システムは，Java SE6.0u2 を用いて実装した．システムの出力は，異種メディア連結を行った HTML ファイルを出力する．本システムの実現により，Web 上に散在するブログ記事を一覧する際，これら 1 件毎に意味的関連性のあるアイコン画像を付与し可視性を持たせることによって，多量のデータの一覧性を向上させることが可能となる．

3.1 対象 Web 単語辞書

本方式のプロトタイプ実装において，機能 1．辞書グラフ構築，機能 2. メディアデータからのメタデータ抽出では，コミュニティサイト “はてな” の提供する Web 単語辞書である “はてなダイアリーキーワード” を対象とする．なお，この Web 単語辞書の利用は，“はてなダイアリーキーワード連想語 API”，および，“はてなダイアリーキーワード自動リンク API” を用いる．これら 2 つの API は，XML-RPC [4] プロトコルによりアクセス可能となっている．本システムにおける実装では，オープンソースの XML-RPC プロトコルハンドラである，Apache XML-RPC(注2)を用いて，これら Web サービスへのアクセス部の実装を行った．

- はてなダイアリーキーワード連想語 API

はてなダイアリーキーワード連想語 API は XML-RPC を用いて利用することが可能であり，はてなの持つ Web 単語辞書内のある単語から，その単語の定義文章に登場する別単語へのリンク構造を調べることが可能な Web サービスである．

- はてなダイアリーキーワード自動リンク API

入力された文書から，文書中に含まれている単語辞書登録単語を抽出し，自動的に Web 単語辞書へのハイパーリンクを付与する Web サービスである．これを利用し，ブログ文書内に

(注1)：http://yaplog.jp/miyuyagyu/

(注2)：http://ws.apache.org/xmlrpc/

含まれる単語辞書登録単語を抽出する．

上記 API を利用した，単語辞書へのアクセスは次の手順の繰り返しにより実行する．

（1） API を実装している Web サービスと接続

（2） XML-RPC リクエストを送信

（3） XML-RPC レスポンスを受信

3.2 クラス構成

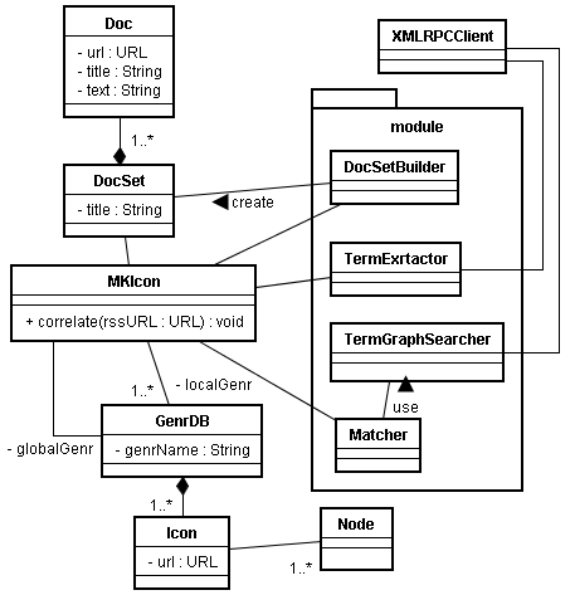

図 4　プロトタイプシステムのクラス構成

プロトタイプシステムのクラス構成を図 4 に示す．本システムは主要な 11 クラスから構成され，実行プロセスは，MKIcon クラスの correlate(URL) がエントリポイントとなっている．本システムの応用システムを実装する場合，MKIcon クラスの API のみを用いるだけでよい．すなわち，本システムは，典型的な Facade パターンに従った実装であるといえる．また，Module パッケージ内のクラスは，インタフェイスクラスとして定義してある．第 2.2.3 節で示した集約関数の切り替えは，Matcher インタフェイスを実装する具象クラスにより実現される．

- MKIcon: メインクラス． 入力されたブログ RSS から，RSS に含まれるブログエントリー文書に関する，アイコンのジャンルを判定し，相関量によるランキングを用いてアイコンを付与した HTML ドキュメントを出力する．
- Doc:文書メディアを表すクラスである．タイトル，本文，URL の 3 属性を有する．本システムにおいて，装飾対象を表す，最も基本的なデータ構造である．
- DocSet: 文書メディアの一覧を表す．Doc クラスの集約クラスであり，Doc クラスのインスタンス集合を管理する．
- Icon: アイコン画像を表すクラスである．本システムでは，この Icon クラスのインスタンスと，Doc クラスのインスタンスを関連付けることにより，ブログの自動装飾を行う．
- GenrDB: アイコン画像の集約クラス．ジャンルごとに管理された画像アイコンを持つ．
- Node: 辞書グラフ上でのノードを表す．すなわち，辞書グラフ上の単語文字列を表す，シンプルなデータ構造である．
- DocSetBuilder: 入力された RSS URL から，RSS 内容を解析し，DocSet クラスを作成するモジュール．RSS により取得されたブログテキストを対象に，次の TermExtractor により，文書メタデータを抽出する．
- TermExtractor: 文書から単語辞書登録単語を抽出するモジュール．第 2.2.2 節で示した機能 2 を実装するクラスである．
- TermGraphSearcher: 任意の単語ノードに隣接する単語ノードを辞書グラフ内から探索するモジュール．
- Matcher: 異種メディアの関連性を計量するモジュール．第 2.2.3 節で示した機能 3 を実装するクラスである．
- XMLRPCClient: はてなの API を利用する際の，XML-RPC 通信を行うモジュール．本クラスの実装には，Apache XML-RPC ライブラリを使用している．

3.3 基本機能の実装

3.3.1 機能 1. 辞書グラフ生成

はてなダイアリーキーワード連想語 API(注3)を利用し，辞書グラフを構築する．構築された辞書グラフを再帰的にトラバースするアクセスの度に，はてなダイアリーキーワード連想語 API を利用する．はてなダイアリーキーワード連想語 API は，はてなダイアリーという日記サイト中で用いられるキーワード配列をパラメタとして送信すると，はてなダイアリーのキーワードデータベースと照合し，関連するキーワードを返信する．具体的には，検索対象キーワードとして “Java” を送信すると，連想語として次の単語群を得ることができる．

> 末松千尋，JVM，ネットワーク，ジャワ，JavaScript，タイプ，再生，Java EE，サーバー，欠点，J2EE，クリティカル，環境，B2B，Sun Microsystems，マシン，KAFFE，バイトコード，JDK，テレビ，選択肢，Java Virtual Machine，サイト，言語，James Gosling，コンパイラ，技術，Enterprise，Swing，ミッション，1991 年，Oak，フリー，JIT，デスクトップ，存在，Java Web Start，セキュリティ，限界，Java Development Kit，コンパイル，携帯，GUI，Time，家電，Applet，Servlet，プログラミング言語

なお，この例の実行に要する時間は 5 回の試行を平均して 3000ms であった．これらの単語は，はてなダイアリーキーワードという日記サービス（ブログ）において，ユーザが関連を記述することによって得られたものである．

3.3.2 機能 2. メディアデータを対象としたメタデータ生成機能

メディアデータを対象としたメタデータ生成機能として，ブログ記事からメタデータを自動生成する機能を実装した．はてなダイアリーキーワード自動リンク API(注4)を利用し，ブログ記事内に含まれる Web 単語辞書登録単語を抽出する．アイコ

(注3)：http://d.hatena.ne.jp/xmlrpc の hatena.getSimilarWord

(注4)：http://d.hatena.ne.jp/xmlrpc の hatena.hatena.setKeywordLink

ン画像データは，予め任意のメタデータを付与した画像データ群を用意し，これを対象に実験を行った．はてなダイアリーキーワード自動リンク API は，任意のテキストを送信すると，はてなダイアリーのキーワード，すなわち，本システムにおける辞書グラフのノードとなる単語を抽出し，単語部分を自動的にハイパーリンクに変換し，返信する API である．この API の返却値のアンカータグを解析することにより，任意の文書における，辞書グラフ単語の出現を分析することができる．具体的には，メタデータ生成対象文書として本稿のあらまし部分のテキストを送信すると，当該文書のメタデータとして次の単語群を得ることができる．

> Web 2.0, 環境, コミュニティ, 辞書, 辞書, グラフ, 画像, データ, ブログ, メディア, 意味, メディア, データ, システム, コミュニティ, 辞書, ハイパーリンク, 再帰, トラバース, 意味, ホップ, メディア, データ, 意味, メディア, データ, システム, システム, データ, アイコン, データ, メディア, データ, マルチメディア, 実験, メディア

なお，この例の実行に要する時間は 5 回の試行を平均して 600ms であった．これらの単語は，Web 単語辞書はてなダイアリーキーワードにユーザが登録したことにより得られたものである．

3.3.3 機能 3. 異種メディア関連性計量機能

ブログ記事とアイコン画像の関連性を計量する機能として，Step-4. 集約 では方式 (3)「行積演算・列和演算」を実装した．これは，ブログ記事データが幅広く散漫なメタデータを有しており，かつアイコン画像データがメディアの特徴を捉える正確なメタデータを有しているため，アイコン画像の各メタデータの距離集約に積演算を用いることが，二つのメディアの距離を計量する上で望ましいと判断したためである．

3.3.4 機能 4. 異種メディアデータを対象とした文脈的連結機構

ブログ記事 1 件に対し，機能 3 で求められたアイコン画像との相関量を利用し，ブログ記事一覧の一覧性を向上させる HTML を出力する．なお，アイコン画像はジャンル分類されており，平均相関量の最も高いジャンル中からのみアイコンを抽出し，相関量順に付与する．これにより，意味的に関係が強く，なおかつ，同一ジャンルのアイコンのみが付与されることになり，高い精度の出力結果を得ることができた．また，ジャンルに関わらないアイコン画像群を用意し，相関量順に付与する．これにより，ジャンルによらない記事の一般的な特徴をアイコン画像として連結することができた．アイコン画像データをジャンル別に用意するためには，人手による分類が必要となるが，現在，Flickr(注5) などの写真共有サイトでは，ユーザによる画像の協調的な分類，所謂 folksonomy が一般的となっており，人手による分類を前提としても，本システムの実用性に影響を及ぼさないと判断した．次に，本機能の実行結果を示す．図に示すように，ブログ記事中のテキスト内容に応じて，ブログと同一ジャンルのアイコンを自動的に付与できていることが分かる．

4. 実　　験

本方式の有用性を示すため，前述のプロトタイプシステムを使用し，日記系ブログサイトのエントリー記事数件とアイコン画像との関連性を計量し，記事一覧に，記事との関連性が高いアイコン画像を付与する実験を行った．

4.1 実験データ

- アイコン画像: メタデータ抽出済みのアイコン画像 263 件をあらかじめ用意した．アイコン画像は book，entertainment，fashion，gourmet... といったブログ記事の分類に適した 14 のジャンルに分類されているものを使用している．
- 辞書グラフ：はてなダイアリーキーワード連想語 API により取得された，2008 年 1 月 16 日現在の “はてなダイアリーキーワード登録単語” を対象に構築された辞書グラフを用いた．なお，辞書グラフ構築のために当該 API へのアクセス対象単語ノードの規模は，延べ 483,563 ノードであり，辞書グラフへのアクセス回数は，延べ 26,099 回に及んだ．
- 対象ドキュメントデータ: 3 つのブログ記事を対象に，実験を行った．3 ブログ記事の詳細は，次のとおりである．1．女性アイドルによる日記系ブログサイト，“miyu’s diary”(2007 年 2 月 17 日現在)，2．お菓子に関するブログサイト “Tea Spoon Cafe”(注6)(2008 年 1 月 17 日現在)，および，3．技術ニュースを掲載しているブログサイト “GIGAZINE”(注7)(2008 年 1 月 17 日現在)．各ブログ RSS に記述されている記事エントリー 5 件について，記事エントリー本文を取得し，それら記事ドキュメント群を対象に本方式を適用した．

4.2 実 験 結 果

実験結果を以下に示し，アイコン付与の精度，および，実用性について考察する．なお，実験結果に含まれるアイコン画像は本稿の公開のため，著作権を考慮した画像に差し替えたものである．

- **ブログサイト miyu’s diary：** 解析結果の HTML ドキュメントを図 5 に示す．エントリ 1，エントリ 2 では文章中に含まれる単語が，アイコン画像のメタデータと直接マッチしたことが確認できる． エントリ 1 は “学校帰りに TSUTAYA に寄り，DVD を借りた” という内容の文章であり，文章中に含まれている，“TSUTAYA”，“DVD” といった単語と，それぞれの単語をメタデータとするアイコン画像との相関量が上位にランキングされている．エントリ 2 は学校でサッカーをして遊んでいたら怪我をした，という内容の文章であり，エントリ 1 と同様の理由で，“サッカー” アイコン画像が高くランキングされている．エントリ 3 からは，文書中に含まれる暗黙的な意味を解析した相関量計算が確認できる． エントリ 3 は，ヤングチャンピオンへの自分の写真が掲載されているという内容の記事であり，姉妹雑誌である “少年チャンピオン” を示すアイコン画像との相関量が 2 件目にランキングされ，1 件目には “漫

(注5)：http://www.flickr.com/

(注6)：http://blogs.yahoo.co.jp/vanille524/

(注7)：http://gigazine.net/

画” を意味するアイコン画像がランキングされている．ここで，文章中には “ヤングチャンピオン” という単語に含まれてはいるが，文章中には，“チャンピオン” が “ヤングチャンピオン” と関連があることや，それらが雑誌の名称であり，また，漫画雑誌であることは記述されていない．しかし，本方式による辞書グラフ上での関連性計量では，“ヤングチャンピオン” と “漫画” という単語に意味的な関連があることが計量され，文章中に含まれている暗黙的な意味を解析することが可能となったのである．

- **ブログサイト Tea Spoon Cafe：** 解析結果の HTML ドキュメントを図 6 に示す．エントリ 1〜6 は年末年始のことについて文章中で触れており，ジャンル “event” の，クリスマス画像や，年賀状画像などの，年末年始関係のアイコン画像が高くランキングされている．かつ，エントリ 1〜2 とエントリ 3〜6 を比較すると，エントリ 4〜5 に関しては，クリスマスについて触れており，“クリスマス”，“クリスマスケーキ” とった単語を文章中で使用しているため，年賀状アイコンに比べて，クリスマスアイコンとの相関量が高くランキングしている．エントリ 1〜3 に関しては，正月について触れており，“おせち”，“初詣”，“正月” といった単語を使用しているため，クリスマスアイコンに比べ，年賀状アイコンとの相関量が高くランキングされている．
- **ブログサイト GIGAZINE：** 図 7 は，テクノロジに関するニュースを集めたブログサイトである GIZAGINE を対象としたアイコン修飾結果を示したものである．結果中のエントリ 2 の記事は，音楽に関する広告情報であり，音楽関連のアイコンが付与されている．これは，記事中の “iTunes” や “レコード” といった単語と強い相関を有するアイコンが付与されたためである．また，エントリ 3 の記事は “EeePC” という廉価なラップトップ PC について触れた記事であり，PC と相関の強いアイコンが付与されている．“EeePC” という単語は実験時において，新しい単語で話題性があり，Web 単語辞書に単語の意味が登録されていたため，本方式による意味的な関連性を計量することが可能となっている．3 つのブログサイト中，GIGAZINE は他の 2 サイトにとは異なるジャンルのブログサイトであるが，同程度の性能の連結を可能としていることが確認できる．これらの実験結果は，Web 単語辞書が話題性を持った新規の単語に素早く対応する性質と，コミュニティの持つ多様な語彙を，本方式により利用できている為である．

5. おわりに

本稿では，Web 2.0 環境において，コミュニティにより構築された知識である単語辞書を辞書グラフとして構築し，これを用いて，画像データやブログ記事など，異種メディア間の意味的関連性を計量することにより，相関の高いメディアデータ間の結合を行う方式，および，その実現システムを示した．本システムの応用として，ブログ記事文書メディアを対象としてアイコン画像データを付与するアプリケーションを実際に構築した．本方式は，Web2.0 環境における関連異種メディアデータの同時的な表示による新たなマルチメディア表現を実現するものである．また，評価実験により，本方式の実現可能性，および，有効性を示した．今後は，動画像などの時系列メディアデータに対する多メディアの関連付け，および，大規模ドキュメント群に対するスケーラブルな関連付けを実現するシステム構築を予定している．

図 5　miyu’s diary(http://yaplog.jp/miyuyagyu/) の記事一覧装飾結果

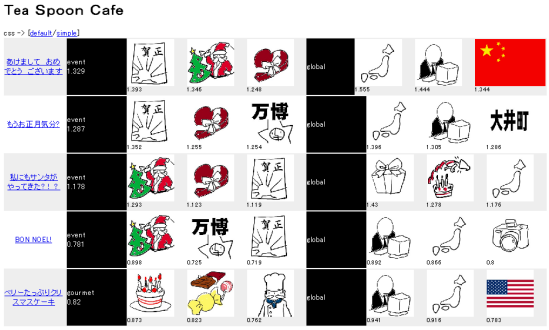

図 6　Tea Spoon Cafe(http://blogs.yahoo.co.jp/vanille524/) の記事一覧装飾結果

図 7　GIZAZINE(http://gigazine.net/) の記事一覧装飾結果

文　献

[1] 北川高嗣，中西崇文，清木康: “静止画像メディアデータを対象としたメタデータ自動抽出方式の実現とその意味的画像検索への適用”，情報処理学会論文誌：データベース，VOl.43,No.SIG12(TOD16), pp38-51, 2002.

[2] Salton, G. and McGill, M. J. “Introduction to modern information retrieval”, McGraw-Hill, 1983.

[3] 株式会社はてな: `http://www.hatena.ne.jp/`

[4] Winer, D. : XML-RPC Specification, `http://www.xmlrpc.com/spec`, 1999.