複数Webサイトからの共通側面の抽出と類似サイト検索

小谷　彬† 小山　聡† 田中　克己†

† 京都大学大学院情報学研究科社会情報学専攻　〒 606–0801 京都市左京区吉田本町
E-mail: †{kotani,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　大学の研究室の Web サイトなど，同種の類似した Web サイト群があり，それらは例えば論文に関するページなど，共通した属性の Web ページを持つ。本論文では、この属性を「側面」として，複数の類似 Web サイトを与えて，共通側面を抽出する手法を提案する．本手法は，各 Web ページ中から，HTML の構造を利用してキーワードを抽出し，さらには Web サイトごとに各側面に該当する Web ページを抽出するために，キーワードの site frequency という概念を導入して，側面抽出を行っている．本論文では他の側面抽出手法と比較するとともに，実験・評価を行う．本手法を用いることにより，類似 Web サイト集合を与えると，共通の枠組みでサイトマップが作成されることになり，異サイト間の共通項目の統一的・横断的な比較が可能になる．また，共通側面抽出によって得られた結果を元に，類似 Web サイト群に属しうる新たな Web サイトを検索する手法についても述べる．
キーワード　データマイニング，情報検索，Web とインターネット

Extracting Common Aspects from Multiple Web Sites and Search for Similar Web Sites

Akira KOTANI†, Satoshi OYAMA†, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University　Yoshidahonmati, Sakyou-ku, Kyoto 606-8501 Japan
E-mail: †{kotani,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　There is the similar Web site group such as Web sites of the laboratory of the university, and, for instance, they have the Web page of a common attribute like the page about the thesis etc. In this paper, we call this attribute "Aspect" We propose the technique for extracting a common aspect from multiple similar Web sites. We first extract the keyword by using the structure of HTML on each Web page. Then we introduce the concept named "site frequency" of the keyword to extract the Web page that corresponds to each aspect in each Web site. We compare with other aspcet extraciton techiniques and experiment and evaluate them. If the similar Web site set is given, the sitemap will be made with a common frame, and it becomes possible to compare the Web pages which have common aspect between different Web sites. More over we propose the technique for retrieving a new Web site in which it can belong to the similar Web site group using the result of obtaining by the common aspect extraction.
Key words　Data Mining, Information Retrieval, Web and Internet

1. はじめに

Web サイトには類似した内容や構成を持つものが存在する．たとえば，大学・研究室・企業・政治家の Web サイトなどである．大学の研究室のサイトの場合では，多くのそれらのサイトには，研究概要・発表論文・メンバー紹介といったページが含まれている，このような類似した Web サイト間において，共通の項目に関するページを比較して閲覧するためには，ユーザがそれらに該当するページを探して，比較する必要がある．この作業は手間がかかり，またその発見が容易でない場合もある．

我々はこの問題を解消するために，複数類似 Web サイト間における共通した属性を側面とし，複数の類似 Web サイトを与えて，その共通属性（側面）と，各 Web サイトにおける各属性に該当する Web ページを抽出する手法を提案する．例えば大学の研究室の場合，側面として考えられるのは，“ メンバー ”,“ 研究紹介 ”,“ 発表論文 ”,“ 教授 ”,“ アクセス ”などであり，

多くの大学の研究室の Web サイトにはこれらに該当する Web ページを持っていると考えられ，それらに該当する Web ページを抽出する．これが図 1 における側面抽出の部分に該当し，3 節で詳しく述べている．側面抽出の結果，複数 Web サイト中の Web ページが A,B,C,D.. とクラスタ分けされ，各クラスタには，もとの Web サイトそれぞれから，共通した側面の Web ページが含まれている．例えばクラスタ A は“ メンバー ”という共通側面であり，含まれる Web ページは○○研・××研・△△研それぞれのメンバー紹介の Web ページであるということである．

また側面抽出によって得られる結果は，与えられた類似 Web サイト集合に対する，共通のフレームのサイトマップとも言え，異サイト間の共通項目の統一的・横断的な比較が可能になる．例えば複数の研究室の Web サイトを与えると，発表論文に関する Web ページをそれぞれから抽出が可能であるので，発表論文に関する Web ページを複数の研究室の Web サイト間で比較閲覧することが可能である．

さらに，そのフレーム構成を基に，類似した Web サイトの検索なども考えられる．そこで本論文では，4 節で共通側面抽出の結果を用いた類似 Web サイト検索手法の提案・実験・評価を行っている．これが図 1 における類似 Web サイト検索の部分に当たる．例えば複数の研究室の Web サイトを与え側面抽出を行い，その結果から，類似した研究室の Web サイトを検索することが可能になる．応用として，多くの Web サイトが与えられたとき，その内の少数の Web サイトを手動で分類しそれらを基に類似 Web サイトを検索するということを繰り返せば，Web サイトの自動分類が可能になり，Yahoo!カテゴリ [1] のようなカテゴリ分けされたリンク集を作ることも可能になる．

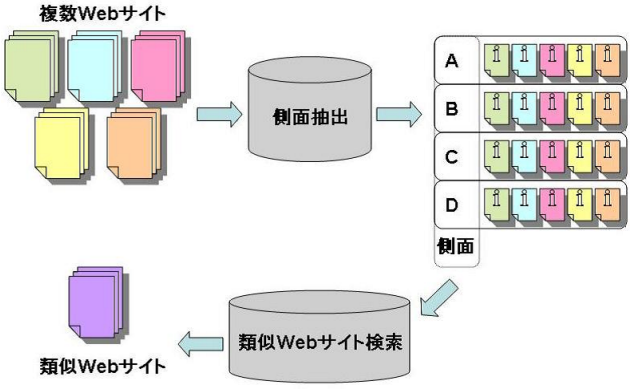

図 1　システムの流れ

2. 関 連 研 究

Web ページをクラスタリングする例として，例えば検索エンジン Clusty [2] がある．検索結果をクラスタリングしてその結果を提示している．検索エンジンでは返されるものは Web ページの集合であり，その複数 Web ページをクラスタリングしている．“ 研究室 ”で検索した場合，結果のクラスタは“ ○○大学 ”,“ ××大学 ”などがあり，“ ○○大学 ”には○○大学の研究室の Web ページが分類される．しかし本論文で提案している手法では，結果のクラスタは“ メンバー ”,“ 研究紹介 ”などであり，“ 研究紹介 ”には○○大学や××大学の研究室の研究紹介の Web ページが分類され，ある側面に沿った Web ページの比較閲覧が容易になる．

また，複数の Web サイトにおける共通属性およびそのインスタンスを抽出する研究は，Web からの情報抽出 (information extraction) 技術に欠かせないものであり，多くの研究がなされている．

河合ら [6] は複数サイトから収集した Web ページを個人の興味に基づいて分類統合する My Portal Viewer を提案している．My Portal Viewer では，あるひとつの Web サイトを与え属性辞書を作成し，それを用いて他の Web サイトから各属性に対応するインスタンス (文字列) を抽出し統合提示している．

灘本ら [7] は Comparative Web Browser(CWB) で二つのニュースサイトを比較閲覧する方法を提案している．基準サイトと比較サイトと呼ばれる Web サイトを指定し，基準サイト内の閲覧したい Web ページを選択すると，比較サイト中の類似している Web ページを自動で提示するものである．

3. 側面抽出手法

3.1 Web サイトの側面

はじめに Web サイトの定義を行う．ひとつの Web サイトは複数の Web ページからなる．それらはあるひとつの Web ページの URI が与えられたとき，そのページからリンクをたどることで移動可能な Web ページのうち，その URI が最初に与えられた Web ページが含まれるディレクトリの URI と前方一致するもの $\{p_0, p_1, ..., p_m\}$ である．これと，最初に与えられた Web ページ p_{top} を合わせた集合を,「Web サイト」とする．

次に,「側面」という語の定義を行う．まず複数の Web サイト集合が与えられているとする．これらは同種の実体（entity）を表現した類似 Web サイト群である．このような複数類似 Web サイトに共通する属性（attribute）を，Web サイトの属性を「側面」と呼ぶ．そして，各 Web サイトに対して各側面に該当する Web ページの決定を行う．

側面抽出の手法として，Web ページの HTML 構造を利用する方法が考えられる．これは，各 Web ページを単なる文書として扱うのではなく，タグ付けされている文書ということを利用したり，あるいは，リンク構造や被リンクアンカー文字列を利用するなどということである．

また，この側面抽出という作業は，Web ページ集合をクラスタリングすることとも見れる．その際，各クラスタが各側面に相当することになるが，ある Web サイトの Web ページが少数のクラスタにのみ含まれるのではなく，できるだけ多くのクラスタに散らばることが望ましい．しかし，単純にクラスタリングした場合，あるサイトに固有の情報が含まれている場合など，前者のようなケースに陥ることが多いと考えられる．これを防ぐために，サイトに依存してクラスタリングするなど工夫が必要である．例えば，小山ら [10] は情報理論における相互情報量 [5] の概念を用いてキーワードのサイト依存性の除去を提

案している．

以上を踏まえて，次節では

- サイト非依存かつ HTML 構造を利用しない場合
- サイト非依存かつ HTML 構造を利用する場合
- サイト依存かつ HTML 構造を利用しない場合
- サイト依存かつ HTML 構造を利用する場合

の場合に分けて側面抽出の手法について具体的に述べていく．

3.2 側面抽出手法

3.2.1 サイト非依存かつ HTML 構造を利用しない場合

この手法は単純な文書クラスタリングとなる．複数の Web サイト集合を $S = \{s_0, s_1, ..., s_n\}$ とする．さらに S の各要素 $s_k \in S$ に対して，そのサイトのページ群集合を $P_k = \{p_{k0}, p_{k1}, ..., p_{km}\}$ とする．全 Web サイトの全 Web ページ集合のひとつひとつの Web ページ $p_{ij}(0 \leqq i \leqq n, 0 \leqq j)$ を一文書として，クラスタリングすることになる．各文書を特徴付けるものは文書全文である．

基本的なクラスタリングである，階層的クラスタリングとして一般的な最短距離法を用いて，大学の情報系の研究室 Web サイト 30 に含まれる Web ページの 6,877 に対してクラスタリングを行った．各 Web ページに対して TFIDF 法で特徴ベクトルを作成し，その類似度をコサイン類似度で定め，クラスタに含まれる Web ページの数が 30 以上のクラスタが 30 個できるまでクラスタリングを行った．残りのクラスタは“ その他 ”として併合した．

このとき，例として“ メンバー ”のページがどのように各クラスタに分布しているかを示したのが，図 2 である．ある側面のページはひとつのクラスタに分布することが望ましいが，図のように各クラスタに少数ずつ分布しており，クラスタの文書数が 30 以下のその他のクラスタに半数が分布しており，複数 Web サイトの共通側面をひとつのクラスタに分布させることが困難であることが分かった．

また，各クラスタに含まれる Web ページをその Web サイトごとに分類したものが図 3 である．各クラスタとも少数の Web サイトで大半を占めており，各側面ごとではなく各 Web サイトごとにクラスタリングされてしまっている．図 4 は，各クラスタが，30 の Web サイトのうちどれだけの Web サイトの Web ページが含んでいるかを，割合で表したものである．この図からも各側面ごとではなく各 Web サイトごとにクラスタリングされてしまっていることが分かる．

以下の節では，単純なクラスタリングでは，各側面ごとではなく各 Web サイトごとにクラスタリングされてしまうという問題を解消するための手法を提案する．

3.2.2 サイト非依存かつ HTML 構造を利用する場合

3.2.1 の手法では，各文書 (Web ページ) を特徴付けるのに文書の内容すべてを用いていた．しかし，側面に着目して分類する場合，全文の内容を用いることは必ずしも適当ではない．たとえば，各文書を特徴付けるのに HTML 構造を用いて，次のようなその Web ページを表す抽象的な名詞が含まれると考えられる要素のみで，各文書を特徴づけるほうが適当であると考えられる．

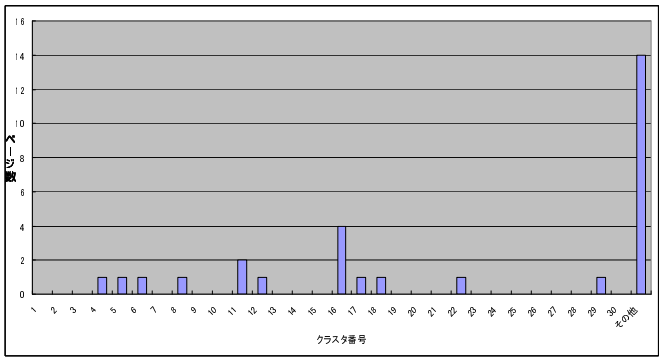

図 2 “ メンバー ”のページのクラスタ分布

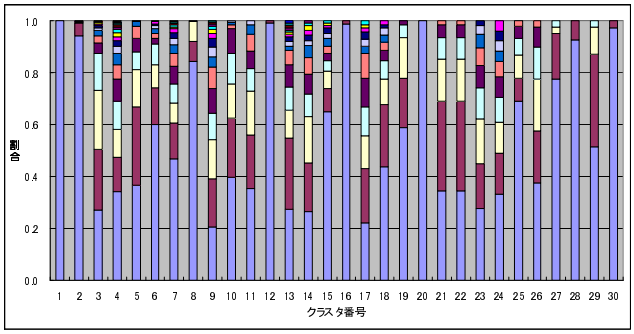

図 3 各クラスタの Web サイト別分類

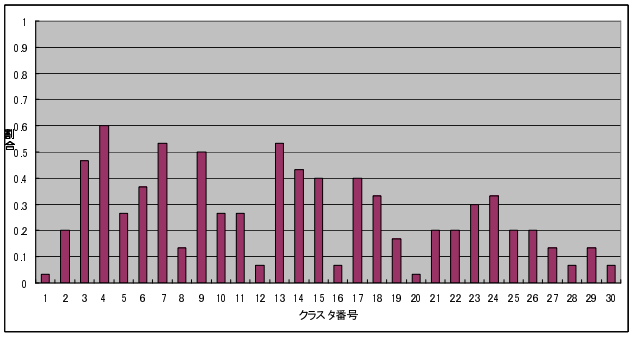

図 4 各クラスタの Web サイト数

- タイトルタグ
- 見出しタグ
- 強調タグ
- ページの名前 (***.html など)
- そのページへのリンクアンカー文字列

3.2.3 サイト依存かつ HTML 構造を利用しない場合

各クラスタが各側面に相当することになるが，ある Web サイトの Web ページが少数のクラスタにのみ含まれるのではなく，できるだけ多くのクラスタに散らばることが望ましい．しかし，上記のような方法で各 Web ページがどの Web サイトに属するものであるかという情報を含まずにクラスタリングした場合，あるサイトに固有の情報が含まれている場合など，Web サイト群の共通属性に基づいて分類されるのではなく，各 Web サイトごとに分類されてしまうなど，前者のようなケースに陥ることが多いと考えられる．これを防ぐために，サイトに依存してクラスタリングするなど工夫が必要である．

そこで，この問題を解消するために次のような手順で側面抽出を行う．この手法では，各側面に対してキーワードを割り当てるものとする．つまり，側面と呼ばれるクラスタごとにそのクラスタを表すキーワードひとつを割り当てるということであり，側面 $A = \{a_0, a_1, ..., a_m\}$ およびキーワード集合 $K = \{k_0, k_1, ..., k_m\}$ とするとき，

$$A \longrightarrow K$$

という 1 対 1 の写像を与えるということである．

（1） キーワード候補の取得

各 Web サイト各ページ毎に全文の内容を形態素解析し，名詞だけを抽出する．それらからストップワードを除いたものをキーワード候補とする．

（2） キーワードの選定

キーワード候補から側面を表す語として適当なものキーワードとして選定する．ここで，*site frequency* という概念を導入する．*site frequency* とは，ある語 t を含むページを持つサイトの数のことであり，$sf(t)$ で表すこととする．

たとえば，次の図 5 のような場合，語 t の *site frequency* は，$sf(t) = 2$ となる．

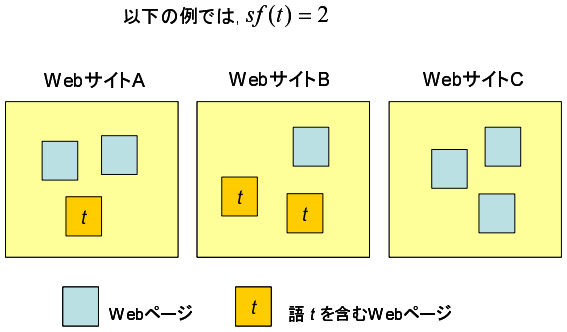

図 5　*site frequency* の例

$sf(t)$ が高いということは，対象とする類似複数 Web サイトそれぞれに，t に関するページを持つ割合が高いということであり，これは t は対象類似複数 Web サイト集合に含まれることが多い抽象的な名詞であると言うことができ，側面をあらわす語として適当だと言えるのである．

さらに各キーワード候補の *site frequency* を正規化するために，*aspect degree* を (1) 式で定める．

$$ad(t) = \frac{sf(t)}{n} \qquad (1)$$
$$(n = |S| : Web \text{ サイトの数}), 0 \leqq ad(t) \leqq 1$$

そして，$ad(t)$ がある一定の閾値以上の語 t をキーワードとする．

（3） 側面に該当するページの決定

上記の方法で側面に相当するキーワードが得られた．最後に，その側面に該当する Web ページを各サイトごとに決定する方法について述べる．

該当する Web ページをひとつに絞って提示する場合を考える．つまり，各 Web サイト s_i にキーワード k_j に対応する Web ページ p_i^j をひとつ定めるということである．

図 6 のような Web サイト A の場合，つまりある側面に対応するキーワードを含む Web ページが，その Web サイトにひとつしかない場合は，その Web ページを選択する．Web サイト A s_a のキーワード $k_j = t$ に対応する Web ページ p_a^j は，$p_a^j = p_{a3}$ ということである．

図 6 のような Web サイト B の場合，つまりある側面に対応するキーワードを含む Web ページが，その Web サイトに複数存在する場合，それらのうちからひとつを選ぶことになる．このときは，PageRank [8] などの手法を用いて Web サイト内における各 Web ページに対してスコア付けを行う．この値を $score(p)$ とする．そしてこのスコアが最も大きかったものを選択する．Web サイト B s_b のキーワード $k_j = t$ に対応する Web ページ p_b^j は，p_{b2}, p_{b3} のうち $score(p_{b2})$,$score(p_{b3})$ を比較して大きい方ということである．後の説明のため，$score(p_{b2}) > score(p_{b3})$ で p_{b2} が選択されたものとする．

図 6 のような Web サイト C の場合，つまりある側面に対応するキーワードを含む Web ページが，その Web サイトに含まれない場合，ほかの Web サイトにおける各該当ページと現在の Web サイトの各 Web ページとの類似度を求め，その和が最も大きいものを選択する．

つまり，ある Web ページ p の特徴ベクトルを$\boldsymbol{p}$，$\boldsymbol{p_0}$と$\boldsymbol{p_1}$の類似度を $sim(\boldsymbol{p_0}, \boldsymbol{p_1})$ で表す場合，図 6 の例では

$$f(p_{ci}) = sim(\boldsymbol{p_{a3}}, \boldsymbol{p_{ci}}) + sim(\boldsymbol{p_{b2}}, \boldsymbol{p_{ci}}) \qquad (2)$$

の値が大きいものを，Web サイト C s_c のキーワード $k_j = t$ に対応する Web ページ p_c^j とする．

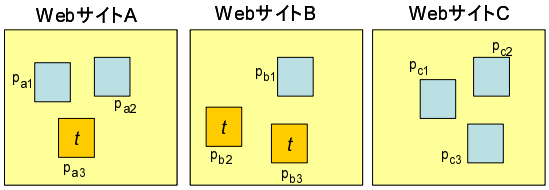

図 6　側面に該当するページの決定例

3.2.4　サイト依存かつ HTML 構造を利用する場合

3.2.3 に述べた手法では，キーワードの取得において各 Web ページの全文中の名詞から取得している．しかし，側面に着目して分類する場合，全文の内容を用いることは必ずしも適当ではない．そこで，3.2.2 で述べたように HTML 構造を用いて，その Web ページを表す抽象的な名詞が含まれると考えられる要素のみで，各文書を特徴づけるほうが適当であると考えられる．

3.2.3 の手法のキーワード候補の取得の段階を次のように変更する．

- キーワード候補の取得

各 Web サイト各ページ毎に，次のようなその Web ページを表す抽象的な名詞が含まれると考えられる要素である

  - タイトルタグ
  - 見出しタグ

– 強調タグ
– ページの名前 (***.html など)
– そのページへのリンクアンカー文字列

の内容を形態素解析し，名詞だけを抽出する．それらからストップワードを除いたものをキーワード候補とする．

この手法で大学の情報系の研究室 Web サイト 30 に対して，キーワード抽出を行なうと，表 1 のような結果になった．キーワード候補の取得における形態素解析には形態素解析システム茶筌 [3] を用いた．

また評価例として，得られた側面に該当する Web ページの選択を行い，その側面に対応する Web ページが取得できていると考えられる割合である適合率は表 2 のような結果になった．側面に該当するページの決定において，候補が複数ある場合には，各 Web ページに対して Mukherjea ら [9] が提案した，Web サイト内のリンク構造やトップページからの深さから算出される *importance* を用いてスコア付けをしてひとつを選択した．ただし，表中の拡大適合率とは，選択した Web ページが適合していなくても，候補の中に適当だと思われるページが含まれていた場合も適合とするときの割合である．また，$/ad(t)$ は，ある側面に対応するキーワードを含む Web ページが，その Web サイトに含まれない場合を除いた場合の各適合率である．

つまり，ある Web サイト A のキーワード k_j に関する正解ページを $p_a(k_j)$ とすると，適合率 P は，側面に該当するページの決定で選ばれた k_j に対応する Web ページ p_a^j が，$p_a^j = p_a(k_j)$ ならば適合として，適合した Web サイトの数を M，複数 Web サイトの総数 N とし，

$$P = \frac{M}{N} \tag{3}$$

である．

拡大適合率 P' は，Web サイト A の Web ページの部分集合

$$P_A(k_j) = \{p_{ai} \mid p_{ai}\text{は } k_j\text{を含むページ}\} \tag{4}$$

としたとき，$p_a(k_j) \in P_A(k_j)$ ならば適合として，適合した Web サイトの数 M' とし，

$$P' = \frac{M'}{N} \tag{5}$$

である．

また，適合率$/ad(t)$，拡大適合率$/ad(t)$ は，

$$P_X(t) = \emptyset \tag{6}$$

なる Web サイト X の数 N' とし，

$$P/ad(t) = \frac{M}{N - N'} \quad \text{および} \quad P'/ad(t) = \frac{M'}{N - N'} \tag{7}$$

である．

C 言語によるプログラミングの解説の Web サイト 18 に対しても同様の実験を行った結果が，表 3，表 4 である．

以上の結果を見ての考察を述べる．まずキーワードの抽出であるが，おおむね研究室，プログラミングともそれぞれの Web サイト群に対して側面と呼べる語が抽出できていると言える．しかし，一般的すぎる語が含まれていることもあり，ストップワードの検討を行う必要があると考えられる．また，“ 情報 ”と“ Information ”や“ 概要 ”と“ 紹介 ”と“ テーマ ”など，同義語や類義語がそれぞれ別のキーワードとなっている．これを解消するために，ひとつの側面をひとつの語と対応付けるのではなく，ひとつの側面と複数の語と対応付けるといった工夫も考えられる．つまり，各 Web サイト s_i に複数キーワード

表 1　キーワード抽出の結果 (研究室)

キーワード	$ad(t)$		キーワード	$ad(t)$
研究	0.97		教授	0.60
リンク	0.90		概要	0.60
メンバー	0.83		関連	0.57
情報	0.83		研	0.53
紹介	0.83		Information	0.50
論文	0.70		学会	0.50
テーマ	0.70		博士	0.50
年度	0.67		Publications	0.50
活動	0.60		修士	0.47
システム	0.60		大学院	0.47
アクセス	0.60		学生	0.47

表 2　適合率による評価例 (研究室)

キーワード	適合率	拡大適合率	適合率 $/ad(t)$	拡大適合率 $/ad(t)$
研究	0.267	0.733	0.276	0.759
リンク	0.433	0.467	0.481	0.519
メンバー	0.733	0.733	0.880	0.880
教授	0.267	0.400	0.444	0.667
紹介	0.200	0.367	0.286	0.524
論文	0.400	0.433	0.571	0.619
アクセス	0.367	0.400	0.611	0.667
発表	0.333	0.433	0.500	0.650
概要	0.333	0.400	0.556	0.667

表 3　キーワード抽出の結果 (プログラム)

キーワード	$ad(t)$		キーワード	$ad(t)$
条件	1.00		型	0.78
言語	0.89		変数	0.78
関数	0.89		プログラム	0.78
プログラミング	0.89		ポインタ	0.78
時	0.78		ファイル	0.78
場合	0.78		基礎	0.78
配列	0.78		整数	0.78
説明	0.78		入力	0.78

表 4　適合率による評価例 (プログラム)

キーワード	適合率	拡大適合率	適合率 $/ad(t)$	拡大適合率 $/ad(t)$
条件	0.78	0.78	0.78	0.78
関数	0.33	0.56	0.38	0.63
配列	0.56	0.67	0.86	1.00
型	0.67	0.78	0.71	0.86
変数	0.67	0.78	0.85	1.00
ポインタ	0.56	0.78	0.71	1.00
入力	0.56	0.67	0.86	1.00

$k_{j0}, k_{j1}, k_{j2}, ... \in K$ に対応する Web ページ $p_i^{j0,j1,j2,\cdots}$ をひとつ定めるということである．

次に適合率による評価例であるが，語によってばらつきがあるということが分かる．拡大適合率では当然適合率より値が高くなるが，ひとつの側面にひとつの Web ページに割り当てるのではなく，たとえば Web ページのスコア値が高いもの複数を割り当てたり，あるいはそれらを統合してひとつの Web ページとして抽出・提示するといったことを行えば，より適合率が高まると言えることを示している．つまり，各 Web サイト s_i にキーワード k_j に対応する複数 Web ページ $P_i^j = p_{i0}^j, p_{i1}^j, p_{i2}^j, \cdots$ をひとつ定めるということである．

また，$/ad(t)$ の場合も 0.5 以上と値が高くなっており，前述したキーワードの抽出で，$ad(t)$ が高くなるように複数の語とひとつの側面を対応付けるなどの工夫を行えば，より適合率が高まると言えることを示している．

4. 類似 Web サイト検索

本節では，前節で述べた共通側面抽出の結果を用いた類似 Web サイト検索手法の提案を行なう．3.2.4 で述べたサイト依存かつ HTML 構造の方法を利用して，側面抽出を行った場合での類似 Web サイト検索の手法を提案する．

4.1 *aspect degree* の高い語を用いた類似 Web サイト検索

aspect degree の高い語は，クエリとして与えられた類似 Web サイト群の多くに含まれる側面を表す語であり，これらを用いて類似 Web サイト検索を行う方法が考えられる．具体的には，*aspect degree* の高い上位たとえば 3 語 t_1, t_2, t_3 を，Google [4] などの検索エンジンで $t_1 \wedge t_2 \wedge t_3$ と AND 検索した結果得られる Web ページの属する Web サイトが，最初のクエリとして与えられた Web サイト集合と類似していると考えられる．

3.2.4 で示した実験結果では，*aspect degree* の高い上位 5 語は，“ 研究 ”,“ リンク ”,“ メンバー ”,“ 情報 ”,“ 紹介 ”であった．Google で上位数語を AND 検索して得られる結果上位 50 件の Web ページが属する Web サイトが，研究室のサイトであるかを判定して評価を行う．検索結果の文書の総数を M，そのうち適合する文書数を N とし，適合率

$$P = \frac{M}{N} \tag{8}$$

で算出する．その結果を表 5 および図 7 に示す．50 件の結果でもおおむね適合率 0.5 前後であることが分かり，図 8 における“ 研究室 ”というクエリと同等以上の適合率を示しており，研究室の Web サイトを表すクエリとして，適当であることが分かる．

また，“ 研究室 ”というクエリで検索した場合と，“ 研究室 ” $\wedge$ クエリ 1 で検索した場合の比較を図 8 に示す．“ 研究室 ”だけで検索した結果に比べ，適合率は大幅に上がり 0.9 前後である．クエリ 1 は“ 研究室 ”という語を補完する語であると言うことができ，キーワードのうち *aspect degree* の高いものは，検索エンジンにおける質問修正としても利用できると考えられる．

表 5　キーワードの AND 検索による類似 Web サイト検索の結果

	クエリ	適合率
1	研究 ∧ リンク ∧ メンバー	0.56
2	研究 ∧ リンク ∧ メンバー ∧ 情報	0.52
3	研究 ∧ リンク ∧ メンバー ∧ 情報 ∧ 紹介	0.33

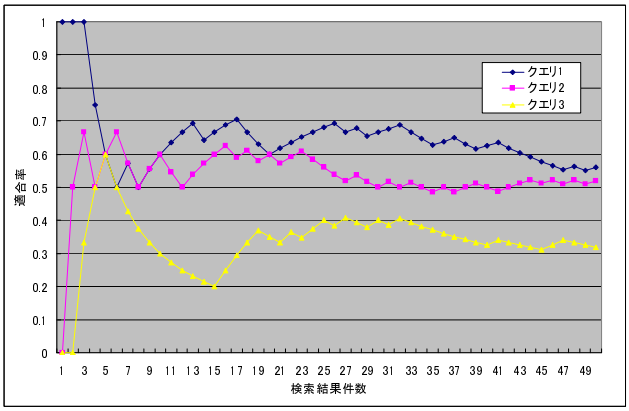

図 7　キーワードの AND 検索による類似 Web サイト検索の結果

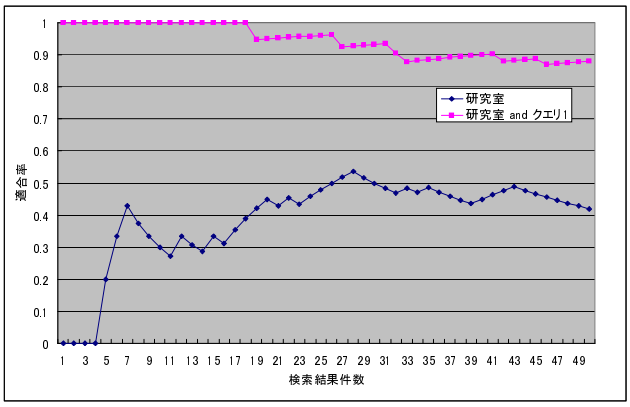

図 8　キーワードの AND 検索による類似 Web サイト検索の結果

検索エンジンに投げるクエリは AND 検索だけでなく，たとえば $t_1 \wedge t_2 \wedge (t_3 \vee t_4)$ などと，OR 検索も織り交ぜることにより，精度を高めることができるとも考えられる．また，別の評価方法として同様にして得られる Web サイトが，他のキーワード $t_4, t_5, t_6, ...$ を含む Web ページを含んでいる割合を見る方法も考えられる．

4.2 類似 Web サイトの判定

ある Web サイト t が，類似 Web サイト集合 S に属し得るかを判定する方法として，S から抽出された側面を表すキーワード集合 K を用いる方法が考えられる．つまり，各キーワード $k_j \in K$ を含む Web ページが Web サイト t に存在する割合が高いとき，t は S に属するというものである．ここで，Web ページ p がキーワード k を含むとは，3.2.4 に述べた Web ページを表す各要素中に k が出現するということである．

例として，3.2.4 の研究室の Web サイト 30 を S，S から抽出されるキーワード集合のうち表 2 に示したキーワードを K として，S に含まれていない研究室 10 の Web サイト集合 T に対して，それらに各キーワードに対応する側面の Web ページを抽出した結果の適合率・拡大適合率を表 6 に示した．ただ

し，括弧内の数字は表 2 で示した S に対する適合率である．サンプル数が少ないため一概に言うことはできないが，類似していると考えられる手動で指定した Web サイト群 T には，S から抽出された K に対する側面のページを持っていることがその適合率から判断でき，K を用いて $t \in T$ が類似 Web サイト集合 S に属し得ると判定することが可能であると言える．

表 6　T の適合率

キーワード	適合率	拡大適合率
研究	0.2 (0.267)	0.7 (0.733)
リンク	0.3 (0.433)	0.3 (0.467)
メンバー	0.9 (0.733)	0.9 (0.733)
教授	0.3 (0.267)	0.3 (0.400)
紹介	0.3 (0.200)	0.4 (0.367)
論文	0.2 (0.400)	0.2 (0.433)
アクセス	0.5 (0.367)	0.5 (0.400)
発表	0.1 (0.333)	0.1 (0.433)
概要	0.1 (0.333)	0.1 (0.400)

また，S を 10 ずつ部分集合の $S = S_0 \cap S_1 \cap S_2$ に分け，$S_0 \cap S_1$ で側面抽出を行ない得られたキーワード集合 K_0 を用いて，S_2 の各 Web サイトが $S_0 \cap S_1$ に属するかを判定し，同様に $S_1 \cap S_2$,$S_2 \cap S_0$ で側面抽出を行ない得られたキーワード集合 K_1,K_2 を用いて，S_0,S_1 の各 Web サイトが $S_1 \cap S_2$,$S_2 \cap S_0$ に属するかを判定し，それらの結果を評価する方法も考えられる．

さらに，類似 Web サイト集合 S_0, S_1, S_2 とそれぞれから抽出したキーワード集合 K_0, K_1, K_2 があるとき，任意の Web サイト t が，S_0, S_1, S_2 のうちどれに属するか，どれにも属さないかを判定することにより，Web サイト集合の分類が可能になる．つまり，任意の Web サイト集合 S のうち少数を手動で分類し，残りを上述の方法で分類するということが可能になる．また，個人のブックマークがカテゴリ分けされている場合，あるカテゴリに入りうるような Web サイトの検索を行ったり，現在閲覧中の Web サイトがどのカテゴリに分類されるかをシステムが自動で判断することが考えられる．

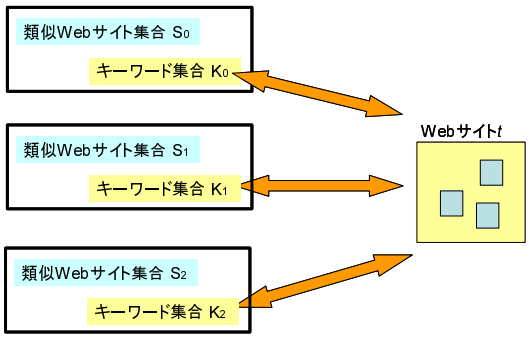

図 9　類似 Web サイトの分類

5.　比較閲覧支援としてのアプリケーション

前節まででは，複数 Web サイトからの共通側面抽出とその結果を用いた類似 Web サイト検索について述べた．本節では，それらの結果を提示して異サイト間の共通項目の統一的・横断的な比較が可能にするための比較閲覧支援アプリケーションの提案を行う．

5.1　システムの例

一例として図 10 のような比較閲覧ブラウザが考えられる．このブラウザは，複数 Web サイトおよび抽出された側面の選択ボックスと，ブラウザペインを持つ．1 で示された Web サイト選択エリアに複数の Web サイトをクエリとして入力し実行すると，2 で示された側面選択エリアに抽出された側面をあらわす語が列挙される．そして，Web サイトと側面それぞれ任意のものを選択すると，3 で示されたエリアに該当する Web ページの候補が列挙され，スコアの最も高い Web ページが自動的に 4 で示されたブラウザペインに表示される．図 10 では，Web サイトは “ dl.kuis.kyoto-u.ac.jp ”，側面は“ 論文 ”を選択しており，ブラウザペインに該当研究室の論文リストのページが提示されている．この状態で，Web サイトを別のものを選択すると，自動的にブラウザペインに別の研究室の論文に関するページが提示される．このようにして，複数の Web サイト間の共通側面の比較閲覧が容易に行える．

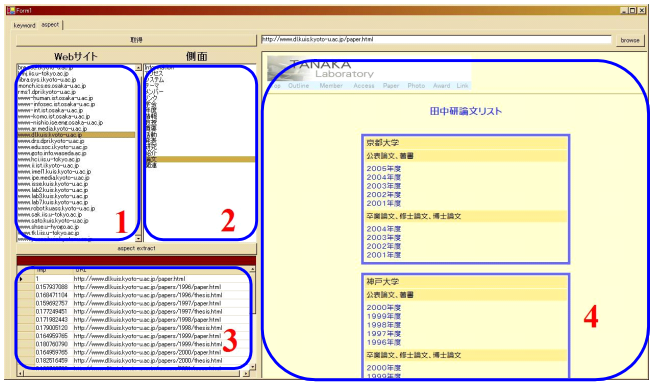

図 10　比較閲覧ブラウザ

また我々は街メタファを用いたインタフェースの提案も行っている [11]（図 11）．このインタフェースでは，Web サイトをひとつの建物に見立て，建物の各側壁に各側面の Web ページを貼り付ける．それら複数の建物を格子状に配置し，街メタファを形成する．このとき，共通の側面が格子状に配置された街の通り沿いに連続して並ぶように配置する．ユーザはウォークスルーすることによって，連続的比較閲覧の可能な Web ブラウンジングを行うことが可能になる．

6.　結　　論

本論文では，複数類似 Web サイト間における共通した属性を側面とし，複数の類似 Web サイトを与えて，その側面と各 Web サイトにおける各属性に該当する Web ページを抽出する手法およびその結果を用いて類似 Web サイトを検索する方法を提案し，それぞれに対して実験・評価を行い，前者では通常のクラスタリングに比べ各クラスタに各 Web サイトから文書が分類され，かつそのクラスタにおいては共通の側面をもつ文書が集まることを確認した．後者では，側面を表すキーワード集合を and で結合して検索エンジンで検索した結果，適合率

図 11　街メタファインタフェース

0.5 以上で類似した Web サイトが検索できることがあることも確認した．

今後は，研究室の例だけではなく，例えば企業や学会の Web サイトなどでもそれぞれの段階で実験を行い，有効性を検証したい．また側面抽出において，3.2.4 に述べたようにある側面を複数の語で表すように拡張し，更なる適合率の向上を図ったり，抽出されたキーワードに階層性・半順序を持たせ，“ 論文 ”の“ 2005 年度 ”という側面を抽出できるようするなどといったことを検討したい．

また，今回の例では研究室という括りで，類似 Web サイトとした．しかし，研究内容が似ている研究室の Web サイトを類似 Web サイトとして側面抽出や類似 Web サイト検索ができることが望ましい．例えば，データベースやコンテンツ処理関連について研究している○○研と××研の，各々の Web サイトの「論文」の Web ページを見ているような場合，これらの論文のページは，

- 構造的にある程度類似

つまり，論文といった語がどちらのページにも出現している．

- 内容的にも「ある程度」類似

どちらも，データベースやコンテンツ処理関連の論文が多い．という特徴を持っている．

このような状況で，

- 構造的には「論文」ページを含んでいるサイト

かつ

- 内容的には、データベースやコンテンツ処理などの分野の研究室サイト

の論文のページが欲しい．つまり，構造的に似ていて，内容的にも「ある程度」似ている，別の研究室サイトを検索する手法についても検討したいと考えている．本論文で提案した手法では，ここでいうところの構造的にある程度類似したものを抽出・検索にとどまっており，内容的にある程度類似したものを抽出・検索できるような拡張を行いたい．

謝　　辞

本研究の一部は，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」，文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検 索・統合ソフトウェア開発」（代表：田中克己），および，平成 17 年度科研費特定領 域研究 (2)「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247，代表：田中克己）および，平成 17 年度科研費若手研究（B）「参照の同一性判定に基づく複数 Web ページの検索閲覧方式の研究」（課題番号：16700097，代表：小山聡）によるものです．ここに記して謝意を表すものとします．

文　　献

[1] Yahoo!カテゴリ
`http://dir.yahoo.co.jp/`.

[2] Clusty
`http://clusty.jp/`.

[3] 形態素解析システム茶筌
`http://chasen.naist.jp/hiki/ChaSen/`.

[4] Google
`http://www.google.co.jp/`.

[5] Norman Abramson. *Information Theory and Codinf.* McGraw-Hill, 1963.

[6] Yukiko Kawai, Daisuke Kanjo, and Katsumi Tanaka. My portal viewer for information integration based on page layout and content. In *DEWS2005*, 2005.

[7] Akiyo Nadamoto and Katsumi Tanaka. A comparative web browser (cwb) for browsing and comparing web pages. In *WWW*, pp. 727–735, 2003.

[8] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1998.

[9] S.Mukherjea and Y.Hara. Focus + context views of world-wide web nodes. In *UK Conference on Hypertext*, pp. 187–196, 1997.

[10] 小山聡, 吉住貴幸. Web からの検索知識発見を利用した専門検索エンジンの構築. 第 46 回人工知能基礎論研究会, 第 54 回知識ベースシステム研究会合同研究会, November 2001.

[11] 小谷彬, 小山聡, 田中克己. 複数 web コンテンツの多面的閲覧のための空間インタフェース. 日本データベース学会 Letters, Vol. 4, No. 1, pp. 161–164, 2005.