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

2006－DBS－140（II）（36）
2006／7／13

複数 Web サイトからの共通属性抽出による共通サイトマップの生成

小谷　彬[†]　大島裕明[†]
小山　聡[†]　田中克己[†]

Web サイトには効率よく必要な情報を得るために，サイトマップが存在し，そのサイトの構造や内容に基づいて情報が整理され提示されている．ユーザにとっては，それが複数の Web サイト間で同様の形式で整理されていることが望ましい．なぜなら類似した 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ことは，ユーザにとって負担であり困難でもあるからである．そこで我々は複数の Web サイト間における共通属性を抽出し，その共通属性の各属性に該当する Web ページを抽出する手法を提案する．その結果，複数の Web サイトに共通のサイトマップが生成できることになる．共通属性抽出においては，属性を一語で表すための手法について述べ，さらに属性間の階層化や類似属性の統合のために，複数の語で属性を現す属性拡張の手法についても述べる．

Generation of Common Site Maps by Extracting Common Attributes from Multiple Web Sites

Akira Kotani,[†] Hiroaki Ohshima,[†] Satoshi Oyama[†]
and Katsumi Tanaka[†]

For getting information of necessity efficiently in a Web site, a Web site has the site map which arranges inforamtions based on the structure and contents of the site. For a user, it is desirable for site maps to be arranged in a similar form between multiple Web sites. Because, it is burden for a user to compare and browse the Web pages describing common subject between multiple similar Web sites. Therefore we propose the technique extracting common attributes between multiple Web sites and finding the Web page falling under each attribute of the common attributes. As a result, we generate the common site map for multiple Web sites. In common attribute extraction, we show the technique to express an attribute by a single word and the technique of attribute expansion to show an attribute in multiple words for hierarchization between attributes and unification of a similar attribute more.

1. はじめに

ある Web サイトから効率よく必要な情報を得るために，Web サイトにはサイトマップが存在し，その Web サイトの構造や内容に基づいて整理され提示されている．しかし，ユーザにとっては，複数の Web サイト間で同様の形式で整理されていることが望ましい．類似した内容や構成を持つもの複数の Web サイトとして，たとえば大学・研究室・学会の Web サイトなどがある．大学の研究室の Web サイトの場合では，多くのそれらの Web サイトには，研究概要・発表論文・メンバー紹介といった Web ページが含まれており，学会の Web サイトの場合では，会場案内やプログラムなどといった Web 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 Web 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 Web 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．

そこで，複数の Web サイト間に共通のサイトマップが存在すれば，ユーザは効率よく情報を得ることが可能になる．本論文では，複数の Web サイト間における共通属性を抽出し，その共通属性の各属性に該当する Web ページを抽出する手法を提案する．その結果，複数の Web サイトに共通のサイトマップが生成できることになる．例えば，大学の研究室の場合，共通属性は“メンバー”,“研究紹介”,“アクセス”,“発表論文”などであり，それらは共通サイトマップにおけるラベルに相当するものである．

処理の流れとしては，図 2 のように共通属性の抽出と各 Web サイトごとに各属性に該当するページの発

† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

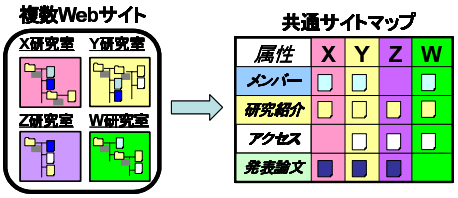

図 1 複数 Web サイトからの共通サイトマップの生成

見の 2 段階に分かれる．最初に行なわれる共通属性の抽出では，複数の Web サイトを解析することによって，属性語として名詞によって各属性を表す方法を提案している．また，各 Web サイトごとの各属性に該当する Web ページの発見では，検索エンジンの Google Web APIs[1] を用いて，発見を行なう方法を提案している．

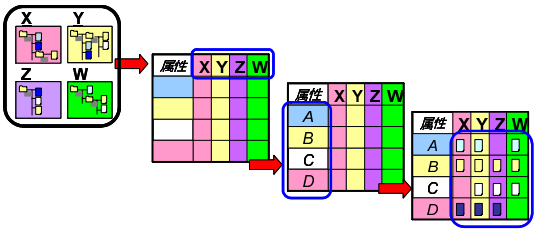

図 2 処理の流れ

本論文では，2 節では関連研究を紹介し，3 節では共通属性の抽出，4 節では各属性に該当するページの発見について提案手法を述べ，それらの実験・評価結果を示し，5 節ではサイトマップの生成，6 節で結論を述べる．

2. 関 連 研 究

小島ら[8] は，ある Web サイト上の意味的に関連した Web ページ群を，ページ間のリンク構造に着目してグループ化することにより，Web サイト内のページを整理する手法を提案している．本論文では，複数の Web サイトのページを，共通属性に注目してサイトマップとして整理しており，この研究とは異なる．

鈴木ら[11)12)] は，複数の Web サイトの文書をディレクトリ構造として階層的に分類する手法を提案している．アンカーテキストによって Web ページを特徴づけ，とハイパーリンクに着目してディレクトリの上位-下位関係を作り上げることを試みている．本論文では 2 節で述べるように，アンカーテキストのみではなく，Web ページのタイトルなどにもよって Web ページを特徴づけ，また階層構造は語の共起関係に着目している．

河合ら[5] 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[6] 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. 共通属性の抽出

本節では複数 Web サイトから共通属性を抽出する手法を説明し，実験結果を示す．

3.1 Web サイト

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 $s = \{p_{top}, p_0, p_1, ..., p_m\}$ とする．さらに，共通属性を抽出する複数の Web サイトを $S = \{s_0, s_1, ..., s_n\}$ とする．

3.2 属 性 抽 出

クエリとして与えられた複数の Web サイトから共通の属性を抽出するために，ひとつの属性をひとつの語で表す手法について述べる．また，この属性を表す語を属性語と呼ぶこととする．属性語を $T = \{t_0, t_1, ..., t_l\}$ で表す．

抽出の流れは，複数の Web サイトに含まれるすべての Web ページから属性語候補 T' を抽出し，そこから属性語としてふさわしいものの絞り込みを行なう，というものである．

3.2.1 属性語候補の取得

属性語候補の取得は，クエリとなる複数 Web サイトに含まれるすべての Web ページから行なう．Web ページ p に対して，以下に挙げるものを p の抽象要素と呼び，$A(p)$ で表すこととする．

- 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

この抽象要素に対して形態素解析を行い，名詞のみを抽出する．それらからストップワードを除いたものを属性語候補 T' とする．

Web ページの全文から候補を抽出するのではなく，抽象要素から抽出したのは，抽象要素には，その Web ページを良く表す抽象的な名詞が含まれていると考えられるからである．

3.2.2 属性語候補の絞り込み

属性語候補は非常に数が多いため，そこから属性語として適切なものを絞り込む手法を説明する．各サイトに共通する属性を表すものとして属性語を抽出するため，属性語を含むような Web ページはより多くの Web サイトに含まれていることが望ましい．そこで，ある語 t を含む Web ページを持つ Web サイトの数を数え，それを評価値として値の大きいものを属性語とする方法が考えられる．ここでページ p がある語 t を含む とは，p の抽象要素 $A(p)$ に t が含まれるということである．

このサイトの数を t の Site Frequency と呼び，$SF(t)$ で表すこととする．

$SF(t) = t \in A(p)$ なる $p \in s_i$が存在するサイト s_iの数

さらに Site Frequency を正規化するために，Attribute Degree $ad(t)$ を (1) 式で定める．

$$ad(t) = \frac{SF(t)}{n} \quad (1)$$

$(n = |S| : Web$ サイトの数$), 0 \leq ad(t) \leq 1$

そして，$ad(t)$ がある一定の閾値以上の語 t を属性語とする．

3.2.3 実　　験

以上を踏まえて，複数の Web サイト集合から共通属性を抽出した実験結果を表 1 から表 4 に示す．表 1 は研究室の Web サイト 30，表 2 は C 言語の解説 Web サイト 18，表 3 はデータベース関連の会議の Web サイト 11，表 4 は京都大学情報学研究科の各専攻の Web サイト 5，をそれぞれクエリとして属性語を抽出した結果である．なお，形態素解析には形態素解析システム茶筌[2] を用いた．

結果についての考察を述べる．いずれの場合においても，おおむね共通属性と呼べるものが抽出できていると考えられる．問題としては，まず一般的すぎる語が含まれていることであり，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．この問題については，属性語の拡張を用いた解決策をのちに述べる．また，京都大学情報学研究科の各専攻の Web サイトにおいては，全てが京都大学のサイトであり，共通属性として“京都大学”が属性語として抽出されている．しかし，これは適切ではない．そこで，クエリとなる Web サイトの全ての Web ページを母集団とした Document Frequnecy を求めて，Site Frequency に Inverse Document Frequnecy を掛け合わせることによって，このような問題が解消できると考えられる．

属性語候補からの絞込みで評価値として $SF(t)$ を導入したが，対象とする Web ページ全てを解析する必要があり，またインデックスを作成するにしても時

表 1 属性語抽出の結果 (研究室)

属性語	$ad(t)$	属性語	$ad(t)$
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 属性語抽出の結果 (C 言語)

属性語	$ad(t)$	属性語	$ad(t)$
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 3 属性語抽出の結果 (会議)

属性語	$ad(t)$	属性語	$ad(t)$
参加	0.55	運営	0.36
プログラム	0.45	組織	0.36
論文	0.45	情報	0.36
発表	0.45	方法	0.36
投稿	0.45	マイニング	0.36
登録	0.45	アクセス	0.36
委員	0.36	データ	0.36
募集	0.36	セッション	0.36
申し込み	0.36	リンク	0.36
申込	0.36	システム	0.36

表 4 属性語抽出の結果 (専攻)

属性語	$ad(t)$	属性語	$ad(t)$
科目	1.00	教員	0.80
入試	1.00	分野	0.80
専攻	1.00	過程	0.80
大学院	1.00	紹介	0.60
情報	1.00	入学	0.60
修士	1.00	応用	0.60
研究	1.00	論文	0.60
問題	0.80	受験	0.60
修了	0.80	専門	0.60
京都大学	0.80	基礎	0.60
講座	0.80	Department	0.60

間がかかった．そこで Google Web APIs を用いてより高速に各属性語候補に対して評価値を定める方法も考えられる．$SF(t)$ は，抽象要素 $A(p)$ に t が含まれる Web ページ p が存在するサイトの数であったが，近似的にタイトルに t が含まれる Web ページ p が存在するサイトの数を Google Web APIs を用いて求めることが可能である．Google Search の“ intitle: ”と“ site: ”オプションを利用して，(2) 式で近似的 Site Frequency $SF'(t)$ を求める．また近似的 Attribute Degree $ad'(t)$ を (3) 式で定める．ただし，$R(Q)$ は Google Search でのクエリ Q に対する検索結果数とする．

$$SF'(t) = \sum E(t, s_i) \tag{2}$$

$$E(t,s) = \begin{cases} 0 & (R(\text{“}intitle:tsite:s\text{”}) = 0 \text{ のとき}) \\ 1 & (R(\text{“}intitle:tsite:s\text{”}) > 0 \text{ のとき}) \end{cases}$$

$$ad'(t) = \frac{SF(t')}{n} \tag{3}$$

$$(n = |S| : Web \text{ サイトの数}), 0 \leq ad'(t) \leq 1$$

この近似的 Attribute Degree $ad'(t)$ と $ad(t)$ の値の比較を表 5 に示す．なお，$SF'(t) \geq SF(t)$ となっているのは，Google Search では PDF 文書や Word 文書なども検索できるためである．

結果に対する考察を述べる．おおむね $ad(t')$ は $ad(t)$ に近似できていると考えられる．しかし大きく値が減少するものもあり，例えば“ 教授 ”という属性語もそうであり望ましくない．これは，Web ページのタイトルからのみ属性語候補を抽出するだけでは不十分であるという可能性を示唆しており，より適切に属性語候補を取得するために，抽象要素の各項目に対して，それぞれ属性語候補として抽出するか否かの比較実験などを行う必要があると考えられる．

表 5 属性語抽出の結果 (研究室)

属性語	$ad(t)$	$ad'(t)$	属性語	$ad(t)$	$ad'(t)$
研究	0.97	0.89	教授	0.60	0.37
リンク	0.90	0.74	概要	0.60	0.37
メンバー	0.83	0.85	関連	0.57	0.44
情報	0.83	0.85	研	0.53	0.44
紹介	0.83	0.59	Information	0.50	0.74
論文	0.70	0.74	学会	0.50	0.52
テーマ	0.70	0.33	博士	0.50	0.19
年度	0.67	0.70	Publications	0.50	0.63
活動	0.60	0.44	修士	0.47	0.37
システム	0.60	0.70	大学院	0.47	0.30
アクセス	0.60	0.48	学生	0.47	0.30

3.3 属性表現の拡張

以上に述べた方法はひとつの属性を現すのに，ひとつの名詞を用いていた．しかし，先にも述べたように類義語が別々の属性として扱われたり，また，属性に階層構造がないことなどから，共通サイトマップを生成する上で属性表現の拡張が必要であると考えられる．そこで，以下でひとつの属性を名詞集合で表現する手法について述べる．基本的なアイデアは，ある属性語 t を抽象要素に含むページ群の本文中によく共起する語を抽出し，t とそれら共起語を合わせた名詞集合で表すというものである．

属性語 t で現される属性の拡張を具体的に述べる．t を抽象要素に含む Web ページの集合を

$$P_t = \{p \mid t \in A(p)\} \tag{4}$$

とする．さらに，P_t の各 Web ページの全文中に含まれる名詞を抽出し，C_{P_t} とする．それらは抽象要素に含まれる t と共起している語の集合である．

C_{P_t} の各共起語 c に対して，(5) 式で評価値 $v_t(c)$ を定める．

$$v_t(c) = DF_{P_t}(c) * IDF_S(c) \tag{5}$$

$DF_{P_t}(c)$: P_tのうち，c が含まれる Web ページの数

$DF_S(c)$: S の全ページ中，c が含まれる Web ページの数

$IDF_S(c)$: $log\frac{N}{DF_S(c)}$

N : S の全 Web ページ数

$v_t(c)$ が閾値 α 以上の c を

$$C_t = \{c \mid v_t(c) \geq \alpha\} \tag{6}$$

とする．そして，従来 t のみで表していた属性を $t \cap C_t$ で表し，

$$D_t = t \cap C_t \tag{7}$$

とする．また t を属性 D_t の**代表語**と呼ぶことにする．

実験として，先の研究室の例で“ メンバー ”という語に対して，上記の方法を適用した結果得られた語を

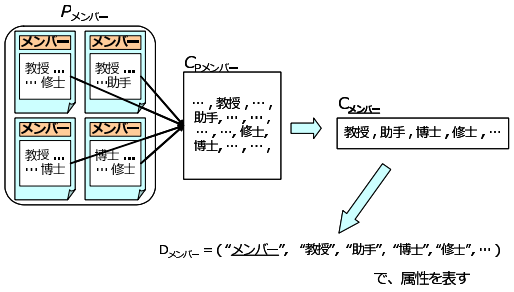

図 3 属性の拡張

表 6 に示す．ただし各 DF 値は 1 で正規化している．

表 6 属性表現拡張の結果 (代表語：メンバー)

共起語 c	$DF_{P_{メンバー}}(c)$	$v_{メンバー}(c)$
助教授	0.61	3.90
助手	0.51	3.90
教授	0.58	3.24
博士	0.58	3.17
修士	0.48	3.16
学部	0.39	2.79
卒業生	0.26	2.31
秘書	0.29	2.18
研究	0.80	2.03

結果の考察を述べる．おおむねメンバーを紹介するページに含まれる内容を表す語が抽出されていると考えられる．また，$DF_{P_{メンバー}}(c)$ だけでなく，$IDF_S(c)$ を掛け合わせることによって，“研究”のような多くのページに出現する語の評価が不当に高くなることを防いでいる．また，“秘書”や“卒業生”のような語は，メンバーを紹介するようなページに含まれる内容としては適切と考えられるが，$DF_{P_{メンバー}}(c)$ だけでは値が低い．しかしながら，$P_{メンバー}$ 以外の Web ページには出現数が非常に少ないと考えられるため，$IDF_S(c)$ を掛け合わせることによって，このような語を抽出することに成功している．

応用として，これら複数の語によって表される属性をベクトル化することによって，類似属性の統合や，属性の階層化が考えられる．前者は属性ベクトル間の類似度をコサイン類似度などで算出し，類似度が閾値以上の 2 属性を統合するものである．後者は，代表語 t_1 が，D_{t_2} に含まれており，かつ $DF_{P_{t_2}}(t_1)$ が閾値以上の場合，$t_2 \rightarrow t_1$ という上位-下位関係を構築し，親が D_{t_2}，子が D_{t_1} という属性の階層化を行なうことも考えられる．

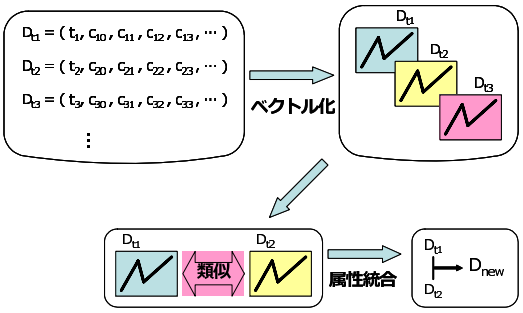

図 4 類似属性の統合

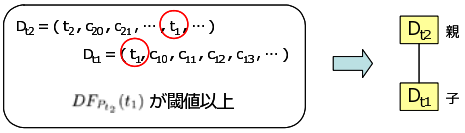

図 5 属性の階層化

4. 属性に該当するページの発見

本節では各属性に該当するページの発見する手法を説明し，実験結果を示す．3.2 で述べた属性を 1 語で表す場合の方法を示す．属性語が含まれる Web ページを各サイトごとに取得するという方法が考えられる．しかしそのような Web ページは複数存在することが一般的である．そこで，そのような Web ページにスコア付けを行いランキングを行なうことにより，上位のページを提示するという方法が考えられる．我々は過去の研究[9)10)] でリンク構造のみに着目したスコア付けとランキングを行なった．しかし，その手法は計算に時間がかかる上，発見の精度も改善の余地があるものであった．

そこで，Web サイト s における属性語 t で表される属性に該当する Web ページの発見を，Google Search の“site:”オプションを利用して，“t site:s”というクエリに対する検索結果を用いる手法を提案し，実験を行なった．表 7 にその結果を示す．表中の**適合率**は，上記のクエリに対する検索結果の 1 位のページがその属性を表すページとしてもっともらしいものであるような Web サイトの割合で，**拡大適合率**はそのようなページが検索結果の上位 10 位以内に存在するような Web サイトの割合である．また，**適合率'** および**拡大適合率'** は，我々が過去の研究での手法で行なった場合の結果である．

結果に対する考察を述べる．従来の手法と比べて，各属性語について適合率・拡大適合率が上昇しているものが多く，減少しているものは少なくその数も上昇

表 7　属性に該当するページの発見 (研究室)

属性語	適合率	拡大適合率	適合率’	拡大適合率’
研究	0.464	0.679	0.267	0.733
リンク	0.607	0.679	0.433	0.467
メンバー	0.714	0.857	0.733	0.733
教授	0.464	0.679	0.267	0.400
紹介	0.292	0.500	0.200	0.367
論文	0.393	0.607	0.400	0.433
アクセス	0.321	0.357	0.367	0.400
発表	0.250	0.429	0.333	0.433

表 8　属性に該当するページの発見 (C 言語)

属性語	適合率	拡大適合率	適合率’	拡大適合率’
条件	0.833	0.833	0.780	0.780
関数	0.500	0.722	0.330	0.560
配列	0.500	0.833	0.560	0.670
型	0.389	0.833	0.670	0.780
変数	0.670	0.722	0.670	0.780
ポインタ	0.611	0.833	0.560	0.780
入力	0.500	0.722	0.560	0.670

しているものより少ないという結果であり，改善されていると言える．

さらなる改善方法としては，現在の手法では PageRank[7] などの Google のランキングアルゴリズムによってランキングがなされているが，これはここでの問題に対する適切なランキングとは限らない．そのため，より適切なリランキング手法を考える必要がある．また，単一の Web ページでひとつの属性の属性値たりうるとは限らないので，Web ページ集合を属性値として提示することも考えられる．

他の問題として，属性を 1 語で表しているため，その属性語を含むような Web ページを持たない Web サイトに対しては，その属性に該当する Web ページが存在しないことになってしまう．そのため，属性語 t に対する適合率・拡大適合率は，$SF(t)$ 以上にならない．これを改善するために 3.3 で述べた属性表現の拡張の利用を考える必要がある．

例えば，属性語 t を含まないような Web ページでも，代表語が t である属性 D_t の各要素を多く含むようなページならば，属性 D_t の属性値たりうると考えられる．そのようなページも属性値発見の対象とすれば，適合率が上昇する可能性がある．さらに“メンバー”という代表語で表される語に該当するページを発見し，かつそのページから“教授”や“助教授”など共起語のインスタンスに相当するもの（例えば“教授”なら教授の氏名）を抽出・提示することも考えられ，重要である．

5. サイトマップ生成

本研究の目的は，複数の Web サイトに共通するサイトマップの生成である．一例として，図 6 のようなインタフェースでの提示例を作成した．1 で示された部分にクエリとなる複数の Web サイトが，2 で示された部分に属性選択エリアに抽出された共通属性を表す属性語が列挙され，Web サイトと属性語それぞれ任意のものを選択すると，3 で示された部分に該当する Web ページの候補が列挙され，ランキングが 1 位の Web ページが自動的に 4 で示された部分に表示される．属性を固定して Web サイトを適宜変更すれば，共通の項目に関するページを比較して閲覧することが可能である．

図 6　提示例

6. 結　　論

本論文では，複数の Web サイトに共通するサイトマップを生成するために，複数 Web サイト間における共通した属性を抽出し，各 Web サイトごとに各属性に該当する Web ページを抽出する手法を提案した．各々の段階で実験を行い，一定の有効性を確認した．またそれらに対して問題点を提起し改善策も考案した．また，語の共起関係に着目して複数の語による属性表現の拡張についても述べた．今後は，3.3 節で述べたように属性表現の拡張結果を用いて，類似属性の統合や属性の階層化について検討していくとともに，実験対象をより増やして有効性の確認を行ないたい．

また，最終アウトプットであるサイトマップの生成についてもそのユーザビリティ[3] について考慮して，デザインや構成を考える必要がある．

さらに結果として得られたサイトマップを元にクエリとして与えた複数 Web サイトに類似した Web サイトを検索することも重要である．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[4] 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り，そのカテゴリに共通するサイトマップを保持しており，比較閲覧が非常に容易になる．今後はこの問題についても検討していきたい．

謝辞 本研究の一部は，文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」（リーダー：田中克己，平成 14〜18 年度）および文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，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（研究代表者：田中克己）および文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者：田中克己，A01-00-02, 課題番号：18049041）および文部科学省科学研究費補助金若手研究 (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」（研究代表者：小山聡，課題番号：16700097）によるものです．ここに記して謝意を表すものとします．

参 考 文 献

1) Google Web APIs
`http://www.google.com/apis/`.
2) 形態素解析システム茶筌
`http://chasen.naist.jp/hiki/ChaSen/`.
3) Site Map Usability (Alertbox Jan. 2002)
`http://www.useit.com/alertbox/20020106.html`.
4) Yahoo!カテゴリ
`http://dir.yahoo.co.jp/`.
5) Yukiko Kawai, Daisuke Kanjo, and Katsumi Tanaka. My portal viewer for information integration based on page layout and content. In *DEWS2005*, 2005.
6) Akiyo Nadamoto and Katsumi Tanaka. A comparative web browser (cwb) for browsing and comparing web pages. In *WWW*, pp. 727–735, 2003.
7) Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1998.
8) 小島秀一, 高須淳宏, 安達淳. Web ページ群の構造解析とグループ化. *NII Journal*, No.4, pp. 23–35, March 2002.
9) 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.
10) 小谷彬, 小山聡, 田中克己. 複数 Web サイトからの共通属性の抽出と類似 Web サイト検索. 電子情報通信学会 第 17 回データ工学ワークショップ, May 2006.
11) 鈴木祐介, 松原茂樹, 吉川正俊. アンカーテキストとハイパーリンクに基づく web 文書の階層的分類. 第 19 回 人工知能学会 全国大会, June 2005.
12) 鈴木祐介, 松原茂樹, 吉川正俊. アンカーテキストを用いた web ディレクトリの構築. 第 168 回 自然言語処理研究会, July 2005.