B-024

Web サイト群の構造分析による典型的構造の抽出法

An Extraction Method for Typical Structure by Analyzing Structure of Web Sites

東 祐太朗* 粕谷 英人† 大久保 弘崇 † 山本 晋一郎 †
Yutaro Higashi Hideto Kasuya Hirotaka Ohkubo Shinichiro Yamamoto

1 はじめに

インターネットの普及により，Web による情報発信は必要不可欠となった．それに伴い，Web サイトの粒度は細かくなり，企業につき 1 つの Web サイトという考え方から，プロジェクトにつき 1 つの Web サイトを用意し，企業のトップページはポータルと位置付ける考え方が主流になりつつある．

その為，新たに Web サイトを立ち上げる機会が増えるが Web サイトを設計するにあたり，

- どのような情報を提示すべきであるか
- 掲示する情報をどのような項目に分けるか
- どのような見た目でそれらを掲示するか

といった設計上の決定が必要となる．本論文ではこれらを総じて Web サイト構造と呼ぶことにする．Web の黎明期であればこれらは全て設計者の独自の判断で決められたが，Web が普及した現在では類似の情報を発信する Web サイトに共通して見られる構造を無視することはできない．

類似サイト群に存在する Web サイト構造の共通点は閲覧する側の利便性に影響を与えると考えられる．すなわち，これを大きく逸脱したサイトは必要な情報が得にくいという印象を与えることになる．また逆に，類似サイト群とまったく同じでは，没個性でつまらないサイトという印象を与えかねない．すなわち，今日の Web サイト設計者には，類似サイト群における Web サイト構造の共通点を理解し，それに従う部分と差別化する部分のバランスを考慮した設計が求められる．

しかし，Web サイト構造の共通点を取得するには，設計者自身が人手で読み取る方法しか今のところは存在しない．これには，設計者の主観が結果に影響すること，人手による方法では共通点を抽出する処理の対象となるサイトの数や規模が限定されてしまう問題がある．そのため，Web から機械的に Web サイト構造の共通点を抽出する手法が求められている．類似サイト群から抽出された Web サイト構造の共通点は Web サイト設計に関する客観的な指標を与える．

本論文では，上に述べた 3 項目のうち前 2 つについて共通点を抽出する手法を提案する．以降，この 2 点に限って「Web サイト構造」，またその中の共通点を「典型的構造」と呼ぶ．

2 Web サイト開発の現状と問題

Newman ら [1] は，Web サイトの作成の際の設計項目として以下が使われていると述べている．

Information Design(**情報設計**) 関係する内容のグループをとらえ，全体の構造をはっきりさせる設計

Navigation Design(**ナビゲーション設計**) 情報の構造を探索する手段についての設計

Graphic Design(**外観設計**) 色や画像，フォントやレイアウトといった見た目の設計

User Interface Design(**ユーザーインタフェース設計**) Navigation Design から主に構成され，Information Design と Graphic Design を含んだ，閲覧者と情報をやりとりするための設計

新たに Web サイトを作成する際に，作成者は閲覧者の利便性を考えなければならない．その際，上で挙げた設計項目が利便性を大きく左右する．

既存の Web サイトには上に挙げられた設計項目をある程度考慮した構造が既に存在しているため，既存の Web サイト群には一定の共通構造があると考えられる．新たに Web サイトを作成する際にこの構造から大きく逸脱してしまうと必要な情報が得にくい．また，それに添い過ぎたデザインも没個性的でありつまらないという印象を与えてしまう．

これらを解消するために，同業者の既存 Web サイトを調査し共通の構造を探すことが必要だと考えられる．

2.1 問題点

既存 Web サイトに共通する構造を発見する為には，多くの Web サイトを分析する必要がある．しかし，現状ではそのようなツールが無いため手作業で構造の分析

* 愛知県立大学大学院情報科学研究科
† 愛知県立大学情報科学部

と共通構造の抽出を行わなければならない．また現在，確立された分析方法もなく Web サイト作成者の主観でしか多く存在する構造を発見することができない．

そこで，本論文ではそれらの問題を解決する為に，同業者の既存 Web サイトから構造を分析し，その中に多く含まれる構造である典型的構造を抽出する手法を提案する．特に，見た目を構成するコンテンツは流行により共通構造の変動が大きい為，本論文では [1] で述べられた Information Design と Navigation Design で設計される，Web サイトを構成するコンテンツの構造を対象とする．

3 Web サイトの構造

2.1 節で述べたコンテンツの構造は実際の Web サイトにおいて，Web サイトの構成要素から作られる構造と作成者が明示的に与える構造の二つから得られる．

3.1 Web サイトを構成する要素から得られる構造

ハイパーリンクから導出される構造 Web サイトのページ間に存在するハイパーリンクから定まるグラフ構造を Web サイトの構造と捉えることができる．

ディレクトリから導出される構造 Web サイトの HTML ファイルを配置しているディレクトリ構造を Web サイトの構造として捉えることができる．

3.2 Web サイト作成者が明示的に与える構造

サイトマップ サイトマップとは，Web サイト内のページの構造を一覧できるようにしたものである．サイトマップはページのタイトルとそのページへのリンクで構成されており，Web サイトの製作者がサイトマップを用意している場合，このサイトマップから Web サイト内に存在する個々のページへアクセスすることができる．例として，図 1 に愛知県立大学の Web サイトに用意されているサイトマップを示す．

パンくずリスト パンくずリストとは，Web サイトの構造を木構造として見た際に親となるページへのリンクをリストにすることにより，サイト内においてそのページがどこに位置するかを簡潔に記述したものである．

例 (愛知県立大学　：情報科学研究科　：入試情報)

TOP > 大学院紹介 > 入試情報

3.3 本研究で用いる Web サイト構造

Web サイトによって，サイトマップやパンくずリストは用意されていない事がある．その為，本論文では Web サイトに必ず存在するハイパーリンクを用いて Web サイトの構造とする．ここでは Web ページをノードに，ハイパーリンクをエッジとしたグラフ構造を Web サイト構造と呼ぶ．

図 1　愛知県立大学 サイトマップ

4 典型的構造の抽出

複数の同一ジャンルの Web サイトを比較すると同様の内容を持つページがある．本論文ではページの内容を比較し，同様であると判別することを対応付けと呼ぶ．

また，複数の Web サイト構造において対応付けされたページが共通の構造で配置されている構造を典型的構造とする．典型的構造は次数を持ち，高次であるほど多くの Web サイトに共通して存在する構造である．

本研究で提案する典型的構造の抽出手法の流れは以下の通りである．

1. Web サイト構造の抽出 (4.1 節)
2. 複数の Web サイト間のページの対応付け (4.2 節)
3. 対応関係を用いた典型的構造の抽出 (4.3 節)

4.1 Web サイト構造の抽出

1 つの Web ページが表す情報を，そのページの URL，本文の内容，ページ内のリンクと抽象化する．1 つの Web サイトとは，そのサイトにあるページの集合である．ただしページ内のリンクは，サイト内のページへのリンクのみを考慮する．すなわち，存在しないページへのリンクおよび他サイトへのリンクは除外して以降の処理を行う．

ここで得られた Web サイト構造を，ページをノードに，リンクをエッジとする木構造とする．ここではトップページを根ノードとしリンク先を子ノードとする．また，子ノードから親ノードへのリンクのような，深さが小さくなる方向へのリンクは排除した．Web サイトは循環するリンクを持ちうることから，リンクの構造を正確にモデル化するためには制限のない有向グラフが必要

になるが，提案手法では上記のように木構造でモデル化する．

4.2 Web サイトに存在するページ間の対応付け

4.1 節で抽出した各 Web サイトの構造を対応付けする．

本論文で行った実証実験ではページ内容の同様さを比較する代わりに，<title>タグで宣言されるページのタイトルを用いて比較を行った．比較は形態素解析に基づいて行うが，類似サイトの比較という性質を考慮した補正を行った．

4.2.1 前処理

タイトル文字列の中の記号は形態素解析に悪影響を及ぼすため，前処理で取り除く．例えば，“愛知県立大学 情報科学部｜大学案内” に含まれる文字列を分ける “｜” や，“南山大学 数理情報学部・構内地図” に含まれる “・” のような，タイトル内の区切りを表すような記号がある．このように，内容とは関係の無い記号について全て除去した．

4.2.2 同義語およびサイト固有の語の辞書による置換

ページのタイトルは自然言語であるため，同義語や表現のゆらぎがあると，後述する対応付けの際に Web サイト間のタイトルの類似度が下がり，正しく対応付けされない．例えば，「入試」と「入学試験」は同じ意味を表すが，機械的に文字列の比較をすると一致しない．また，大学の Web サイトであれば大学名のような，Web サイト固有の文字列もページの内容という観点からは同等に扱われなければならない．そのため辞書を用意し文字列の置換を行うことにより同じ文字列にする．

4.2.3 形態素解析を用いたページの対応付け

ページタイトルに対して 4.2.1 項と 4.2.2 項で述べた前処理を行った後，ChaSen[2] を用いて形態素解析を行う．形態素解析の結果，タイトルは品詞ごとに分割される．この中の名詞だけを比較し類似度のスコアを算出する．次にこの類似度を用いて，二つのサイト間のページの対応を求める．

定義 1 (類似度) 二つのタイトル a, b からそれぞれ名詞 $a_1, a_2, \ldots, a_m$ 及び，$b_1, b_2, \ldots, b_n$ が取り出されたとする．また a_1〜a_m の中に b_1〜b_n に等しいものが k 個あるとき，a と b の類似度を $2k/(m+n)$ と定義する．b から見たときに a に等しいものがやはり k 個含まれるために倍にしている．

■例 類似度を求めるタイトルを t1=“愛知県立大学 情報科学部｜大学案内”, t2=“南山大学 数理情報学部・構内地図” とする．

1. タイトルそれぞれから記号を除去する．
   t1=“愛知県立大学情報科学部大学案内”
   t2=“南山大学数理情報学部構内地図”
2. 大学名や学部名等の固有名詞を同一視する為に “愛知県立大学” と “南山大学” を “XXX 大学” に，“情報科学部” と “数理情報学部” を “YYY 学部” に置き換える．
   t1=“XXX 大学 YYY 学部大学案内”
   t2=“XXX 大学 YYY 学部構内地図”
3. 形態素に分解する
   t1′=[XXX, 大学, YYY, 学部, 大学, 案内]
   t2′=[XXX, 大学, YYY, 学部, 構内, 地図]
4. ここで，[XXX, 大学, YYY, 学部] の 4 つが共通しているので t1 と t2 の類似度は 8/12 である．

4.2.4 ページの対応付け

前節により，全てのサイトのページ間に対して類似度が定義された．次にこの類似度を用いて，二つのサイトの間のページの対応を求める．

定義 2 (ページ間の対応付け) サイト A の各ページ a_i について，サイト B のページで最大の類似度を持つページが b_j であるとする．逆に，ページ b_j に対してサイト A のページで最大の類似度を持つページがやはり a_i であるとき，a_i と b_j は対応していると定める．

4.3 典型的構造の抽出

k 次の典型的構造は以下の二段階で抽出される．

1. 与えられた n 個の Web サイトから任意の k 個を取る全ての組み合わせについて，それぞれ構造の共通部分をページの対応関係に基づいて求める．
2. 上で得られた ${}_nC_k$ 個の共通構造について，対応する部分の重ねあわせを行う．

木は根のラベルと子の部分木の集合で表す．ここではラベルは URL の集合とする．子は集合なので，順序は考慮しない．

4.3.1 共通構造の抽出

二つの Web サイト構造に対して，その共通部分とは新たな Web サイト構造であり，双方の Web サイト構造の対応するページのうち，木構造の位置も等しいページだけを取り出したものである。定義を以下に示す．ここで ∼ は木の対応関係を表す．

定義 3 (Web サイト構造の共通構造) 根同士が対応した 2 つの *Web* サイト構造 $t1, t2$ に対して，その共通構造は *intersect* $t1$ $t2$ である．*intersect* の定義は図 2 に示す．

本論文では木の根ノードのページ間が定義 2 で定めたページの対応関係にあるとき木が対応しているとする．

$intersect\ t1\ t2\ =$ 出力木 *where*
　出力木のラベル $=\ t1$ のラベル $\cup\ t2$ のラベル
　出力木の子　　$=\ \{\ intersect\ s1\ s2 \mid s1\ \in\ t1$ の子, $s2\ \in\ t2$ の子, $s1\ \sim\ s2\ \}$
$merge\ t1\ t2\ =$ 出力木 *where*
　出力木のラベル $= t1$ のラベル $\cup\ t2$ のラベル
　出力木の子　　$=\ \{\ merge\ s1\ s2 \mid s1\ \in\ t1$ の子, $s2\ \in\ t2$ の子, $s1\ \sim\ s2\ \}$
　　$\cup\ \{\ s1 \mid s1\ \in\ t1$ の子, $\neg\ \exists s2\ \in\ t2$ の子 , $s1\ \sim\ s2\ \}$
　　$\cup\ \{\ s2 \mid s2\ \in\ t2$ の子, $\neg\ \exists s1\ \in\ t1$ の子 , $s1\ \sim\ s2\ \}$

図 2　木に関する演算 intersect, merge の定義

k 次の典型的構造は，k 個の Web サイト構造の間の共通構造と定める．三つ以上の木構造について共通部分を抽出する際は，二つの構造から共通構造を抽出し，その構造とと三つめの構造について，更に共通部分を抽出する．同様にして，k 個の共通構造を取る際には k-1 番目の構造から得られる共通構造と k 番目の構造を用いて共通構造を抽出する．

4.3.2　典型的構造の構築

前項で得た共通部分の集合を重ねあわせ，典型的構造を抽出する．二つの Web サイト構造の重ねあわせとは，対応するノードは一つのノードに結合し，対応しないノードはそのまま残して得られる Web サイト構造である。定義を以下に示す．

定義 4 (Web サイト構造の重ねあわせ) 二つの *Web* サイト構造 $t1, t2$ に対して，その重ねあわせは $merge\ t1\ t2$ である．*merge* の定義は図 2 に示す．

3 つ以上の共通構造の重ねあわせについては，共通構造のときと同様に，2 つの重ねあわせを順に適用した結果と定める．

4.3.3　2 次の典型的構造抽出例

図 3(a)， 図 3(b)， 図 3(c) の三つの Web サイト構造を考える．数字が同じノードは互いに対応関係にあるノードであることを表す．

■共通部分の抽出　4.3.1 項の定義よりサイト AB 間，AC 間，BC 間の共通部分として，それぞれ図 4(a), 4(b)， 4(c) が得られる．

■典型的構造の抽出　4.3.2 項の定義に従い上で得られた構造を重ねあわせると，図 4(d) で示す 2 次の典型的構造が得られる．

5　評価

提案手法を関数型言語 Haskell で実装し, 実際の Web サイトを対象として実験を行った．

表 1　典型的構造を抽出する Web サイト群

Web サイト名 (トップページ)	URL
北海道大学	www.hokudai.ac.jp
茨城大学	www.ibaraki.ac.jp
熊本大学	www.kumamoto-u.ac.jp
名古屋大学	www.nagoya-u.ac.jp
名古屋工業大学	www.nitech.ac.jp
滋賀大学	www.shiga-u.ac.jp
静岡大学	www.shizuoka.ac.jp
琉球大学	www.u-ryukyu.ac.jp
富山大学	www.u-toyama.ac.jp
愛知県立大学	www.aichi-pu.ac.jp

5.1　実際のサイトにおける典型的構造

表 1 に示す 10 個の国公立大学の公式 Web サイトの典型的構造の抽出を試みた．実験の結果抽出された典型的構造のノードの数を表 2 に示し，一例として 4 次の典型的構造を図 6 に示す．ただし，ノードの末尾に “*” が付いているものは対応付けが間違ったノードを表す．

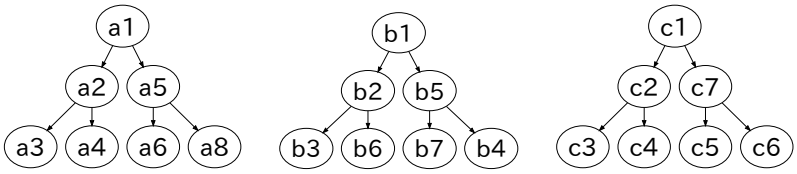
表 2 では，6 次以上のページはノード数が一つしかない．これはトップページ以外の共通構造が抽出されないことを意味する．

5.2　典型的構造の出現率

提案する典型的構造が，実際に同業者の Web サイトにおいてどのくらいあてはまるかを検討する．5.1 節で得られた典型的構造と，別の Web サイト構造間の共通構造を求め，この共通構造のノード数と元の典型的構造のノード数の比を典型的構造の出現率として計測を行った．この実験は表 3 に示す Web サイトについて調べた．これは以下の点を考慮して選定した．

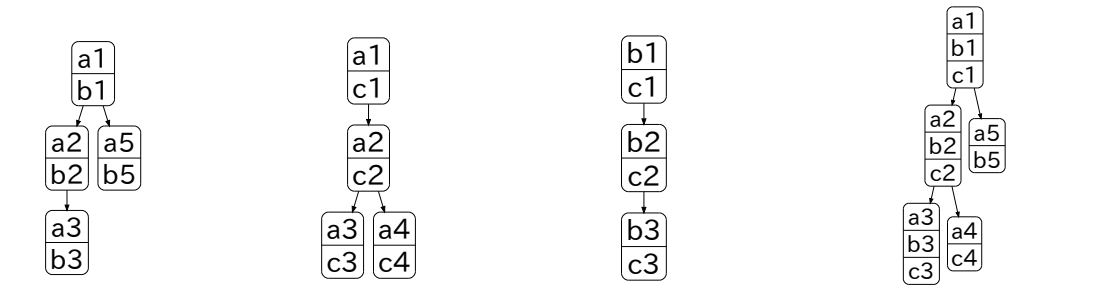
- 同業者である大学の Web サイト (1, 2, 3, 4)
- 類似業者である高校の Web サイト (5, 6)
- 他業種の Web サイト (7, 8)

2 次から 5 次の典型的構造に対して出現率を計測した結果を表 4 に示す．ただし，8 番の任天堂の Web サイ



(a) Web サイト A　(b) Web サイト B　(c) Web サイト C

図 3　入力 Web サイト



(a) AB 間の共通部分　(b) AC 間の共通部分　(c) BC 間の共通部分　(d) Web サイト A, B, C から抽出された 2 次の典型的構造

図 4　抽出される構造

表 2　典型的構造のノードの数

	2 次	3 次	4 次	5 次	6 次	7 次	8 次	9 次	10 次
ノード数	129	34	11	4	1	1	1	1	1

表 3　出現率を調査した Web サイト群

番号	Web サイト名 (トップページ)	URL
1	秋田大学	www.akita-u.ac.jp
2	神戸市外国語大学	www.kobe-cufs.ac.jp
3	名城大学	www.meijo-u.ac.jp
4	名古屋外国語大学	www.nufs.ac.jp
5	愛知中学校／愛知高等学校	www.aichi-h.ed.jp
6	名城大学附属高等学校 【名城高校】	www.meijo-h.ed.jp
7	首相官邸ホームページ	www.kantei.go.jp
8	任天堂	www.nintendo.co.jp

トについてはトップページ以外に典型的構造は出現しなかった．

5.3 考察

表 2 では，次数が低いときに多くのノードを含む典型的構造を得られた．これは，次数が上がるにつれ抽出に用いた Web サイトに共通して含まれるノードが残り，一方，次数が低いときは一部の Web サイトにのみ共通して含まれる構造も抽出されるからである．

表 4 では，同業者の Web サイト (1〜4) においては，次数が上がるにつれ出現率が上がっている．このことから，抽出した典型的構造が実際の Web サイトにも含まれていることがわかる．一方，他業者の Web サイト (7,8) については出現率が低いため，構造が殆ど含まれていないことがわかる．また類似業者の Web サイト (5,6) では出現率は低いが，他業者よりも大きくなった．

このことから，本実験で抽出した典型的構造が大学という業界に特有の構造であることを確認できた．

また，図 6 では，大学特有のコンテンツであるページを含む構造が実際に抽出された．4 次の典型的構造では与えられた Web サイトのうち 4 個以上の Web サイトに含まれるページのみが抽出されている．このことから，ある Web サイト固有の構造を除いた，典型的に含まれる構造を抽出したことがわかった．しかし，構造の深さは 2 しかない為，Web サイト全体の典型的構造の抽出はできなかった．また，図の右下にあるノードに卒業生向けのコンテンツを表すノードがあるが，その中に“滋賀大学　滋賀大学で学びたい方へ”という関係の無いノードが入っている．これは 4.2 節で述べた対応付け

表 4　典型的構造の出現率

番号	Web サイト名 (トップページ)	典型的構造出現率			
		2 次	3 次	4 次	5 次
1	秋田大学	5%	26%	72%	100%
2	神戸市外国語大学	16%	35%	36%	100%
3	名城大学	3%	11%	45%	100%
4	名古屋外国語大学	7%	17%	27%	50%
5	愛知中学校／愛知高等学校	5%	11%	18%	25%
6	名城大学附属高等学校 【名城高校】	3%	26%	36%	50%
7	首相官邸ホームページ	3%	8%	9%	25%
8	任天堂	0%	2%	9%	25%

での誤りが原因である．

また，抽出した典型的構造と実際の Web サイトの構造を実際に人が見たとき，多くのページが実際の Web サイトにも含まれていることがわかる．しかし，大学の Web サイトについて調査した際に，直感的に含まれているはずと感じるページや構造の一部が図 6 に欠けていることもわかった．

6　おわりに

6.1　まとめ

本論文では，同業者の Web サイト群から共通構造を機械的に抽出する手法を提案した．提案手法は，自然言語処理を用いることによって主観の影響を排除し，形式的に共通構造を得ることができる．機械化したため、人手による従来手法にあった規模の問題を克服することができた。Web サイト作成者は個性的かつ利便性を備えた Web サイトの作成に，この共通構造を役立てることができる．また，提案手法を実装し，10 個の Web サイトを用いた初期の評価実験を行い，大学という業界に特有の構造が存在することを確認した．

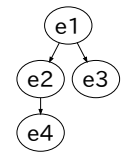
6.2　今後の課題

本論文の今後の課題として以下の二点を挙げる．

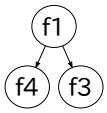
第一に，本論文ではページ間の内容を比較するために，ページタイトルを用いた．ページタイトルに含まれる情報は限られており，比較精度が悪い．そこで，今後はページ内に含まれる見出しや本文，パンくずリスト等を利用することにより，精度向上を検討したい．

第二に，現在の手法は完全に一致した構造から典型的構造を抽出しており，僅かな構造の差があるだけで典型的構造の抽出から除外されてしまう．その為，部分木の経路の差を許すことによって多くの構造を抽出できるようにすべきである．

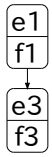
例として図 5(a)，図 5(b) の共通構造を考える．本論文の手法では図 5(c) となるが，ここで親ノードと子ノードの間に対応関係が無いノードが一つ入ることを許すと，図 5(d) となり，構造に含まれるノード数を増やすことができる．



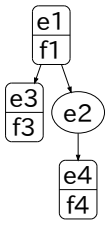
(a) Web サイト E



(b) Web サイト F



(c) EF 間の共通構造 (本論文の手法)



(d) EF 間の共通構造（今後の課題）

図 5　今後の課題

謝辞 本研究は科研費（22300011）の助成を受けたものである．

参考文献

[1] Newman, Mark W. and Landay, James A. : ‘Sitemaps, Storyboards, and Specifications: A Sketch of Web Site Design Practice.’ In: Proceedings of DIS00: Designing Interactive Systems: Processes, Practices, Methods, & Techniques 2000. pp. 263-274.

[2] ‘chasen legacy – an old morphological analyzer’ `http://chasen-legacy.sourceforge.jp/`

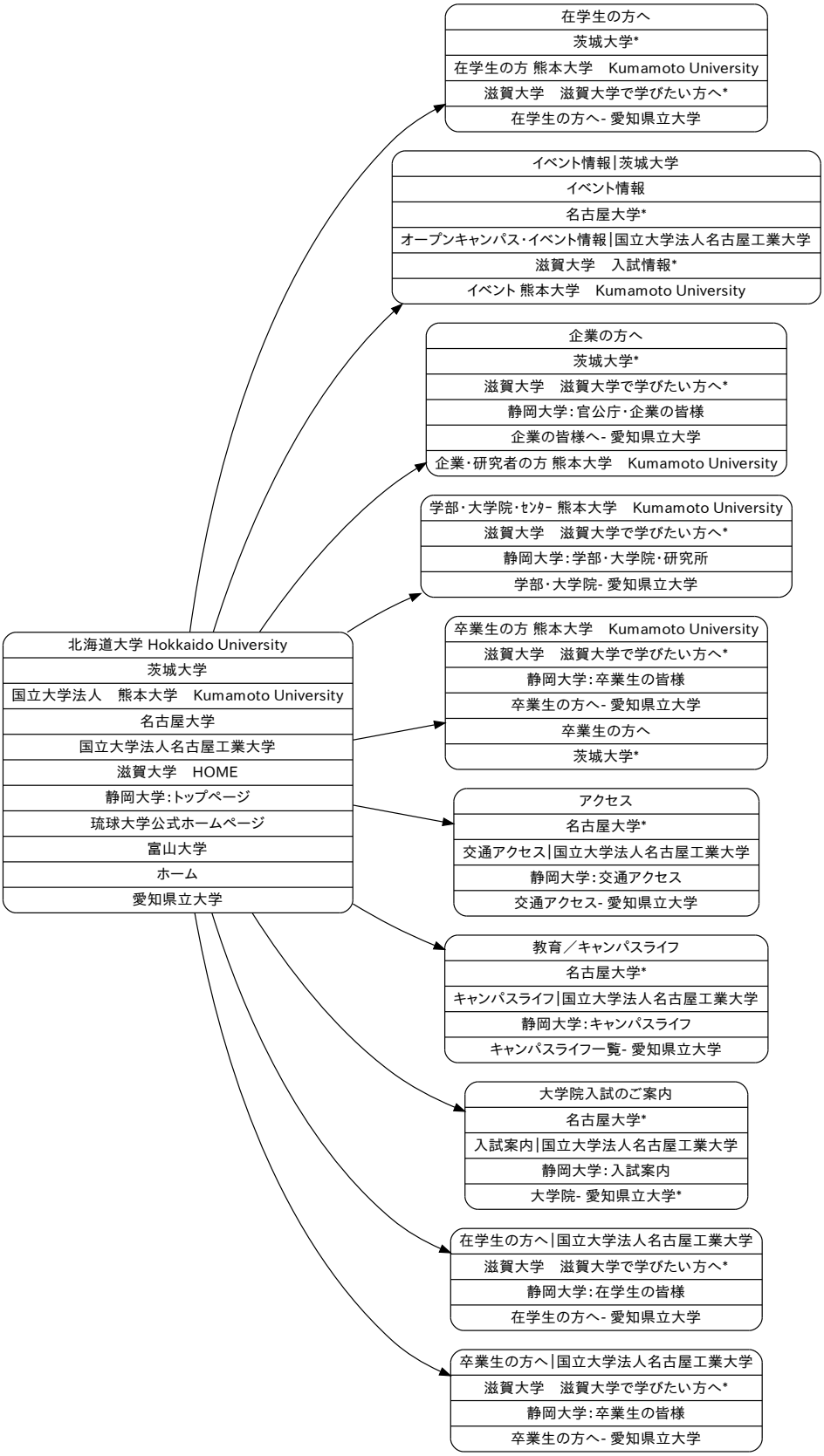


図 6　4 次典型的構造 (*が付いたノードは対応付け間違い)