リンク参照と文書構造に基づく Web ページのアスペクト抽出

荒木　良† 是津　耕司††,††† 角谷　和俊†† 田中　克己††

† 神戸大学大学院自然科学研究科情報知能工学専攻　〒 657–8501 神戸市灘区六甲台町 1–1
†† 京都大学大学院情報学研究科社会情報学専攻　〒 606–8501 京都市左京区吉田本町
††† 独立行政法人 通信総合研究所　〒 184 − 8795 東京都小金井市貫井北町 4–2–1
E-mail: †ryo@ai.cs.scitec.kobe-u.ac.jp, ††{zettsu,sumiya,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　Web 上から目的の情報を取り出す手段として，現在では，検索エンジンを用いたキーワード検索が広く利用されている．この検索手法は，検索者が目的とする Web ページの内容についてある程度の予備知識を持ち合わせており，適切な検索キーワードを選択することが可能である場合には非常に有効な手段といえる．しかしながら，そうでない状況では，所望する Web ページを検索することが非常に困難である．このような場合，検索者は，断片的な知識をもとに試行錯誤によって 検索キーワードを決定し，さらにその検索結果から自分の所望する情報に関連のありそうな Web ページへのリンクをたどり，内容を確認しつつ目的の Web ページを 探索しな ければならない．この問題を解決するための一手段として，検索結果として返された Web ページ群が Web 上においてどのような側面を持つのかという情報を与えることにより，検索者が実際に探索する Web ページを絞り込みやすくするというアプローチが考えられる．しかしながら，このような Web ページの周辺情報は既存の検索エンジンでは取り出すことができない．そこで本研究では，このような Web ページの周辺情報を表すコンテンツ集合のことを “Web ページのアスペクト” と呼び，その抽出手法についての提案およびアスペクトを用いた Web ページ検索についての考察を行う．

キーワード　アスペクト，リンク集

Extracting Aspects of Web Pages by Hyperlinks and Document Structures

Ryo ARAKI†, Koji ZETTSU††,†††, Kazutoshi SUMIYA††, and Katsumi TANAKA††

† Department of Computer and Systems Engineering, Graduate School of Science and Technology, Kobe University　Rokkodai 1–1, Nada-ku, Kobe City, 657–8501 Japan
†† Department of Social Informatics, Graduate School of Informatics, Kyoto University　Yoshidahon-machi, Sakyo-ku, Kyoto City, 606–8501 Japan
††† Communications Research Laboratory　Nukui-Kitamachi 4–2–1 , Koganei, Tokyo 184–8795 Japan
E-mail: †ryo@ai.cs.scitec.kobe-u.ac.jp, ††{zettsu,sumiya,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　As a mean which takes out information from the web database, we often use search engines. This search technique is a very effective means when it is possible for the search person who has a certain amount of preliminary knowledge with him about the contents of the target web page, and can choose suitable search keywords. However, when the user has only the fragmentary knowledge relevant to the target web page, it becomes very difficult to search web page which the user really wants with this technique. In this paper, we focus attention on ”how the web page is recognized by others”. We labeled the other’s recognition as *”the aspect of the web page”*. By using ”aspect” for information retrieval, a search candidate is scolded and it becomes easy to search.

Key words　Aspects, Link Collections

1. はじめに

近年， PC 等の普及，通信帯域のブロードバンド化が急速に進み， Web 上には莫大な数の Web ページが散在している．その結果 Web には様々な情報が蓄積され，いわば，巨大なデータベースとなりつつある．この巨大なデータベースから目的の情報を取り出す手段として，現在では，検索エンジンを用いたキーワード検索が広く利用されている．この手法は，目的の

Web ページの内容に関するキーワードを入力することにより，そのキーワードにマッチする Web ページのリストが検索結果として返され，検索者がこの検索結果リストに表示されているタイトル名や要約文を参考に，リンクをたどって Web ページを閲覧しながら目的の Web ページを発見するという手法である．この検索手法は，検索者が目的とする Web ページの内容についてある程度の予備知識を持ち合わせており，適切な検索キーワードを選択することが可能である場合には非常に有効な手段といえる．

しかしながら，このような手法では，目的とする Web ページに関連する断片的な知識しか持ち合わせていない状況では，所望する Web ページを検索することは非常に困難となる．このような場合検索者は，断片的な知識しか持ち合わせていないために，試行錯誤によって適切な検索キーワードを決定しなければならない．また，その検索結果として表示されている Web ページ群の中で，どの Web ページが自分の所望する情報と関連があるのかといった判断を下すことが困難であり，検索結果に表示された Web ページの最初から手当たり次第に閲覧してその内容を確認しなければならない状況に陥る可能性がある．この作業は非常に煩雑であり，膨大な数の検索結果が返された場合には，目的の Web ページにたどり着くまでに大変な時間を要することとなる．

この問題を解決するための一手段として，検索結果として返された Web ページ群が Web 上においてどのような側面を持っているのかという情報を与えることにより，検索者が実際に探索する Web ページを絞り込みやすくするというアプローチが考えられる．

Web ページの側面とは，図 1 のマクドナルドジャパンのホームページ [1] について考えてみると，マクドナルドジャパンのホームページは，グルメサイトや，生活情報発信サイトなどからは「ファーストフード」や「バーガー」として参照されているが，その一方で，株情報などを発信しているビジネス系サイトからは「有力企業」として参照されている．すなわち， 図 1 の例においては，マクドナルドジャパンのホームページには，「ファーストフード」と「有力企業」という二つの側面を持っていると考えられる．

Web ページの側面を表すような情報は，その Web ページ自身からではなく，その周辺から取り出す必要があるが，既存の検索エンジンではこのような情報を取り出すことができない．ここでいう，Web ページの周辺の情報とは，ある Web ページに対して直接リンクしている Web ページに含まれるコンテンツのうち，リンク先ページに関連しているコンテンツの集合のことである．このようなコンテンツ集合には，ある Web ページから見たときのリンク先の Web ページの内容が含まれており，リンク先の Web ページの側面の一つを表していると考えられる．

このような情報を検索者に与えることにより，検索結果として呈示されている Web ページの内容がより明確になり，検索者が実際に探索する Web ページを絞り込むことが容易になると考えられる．また，このような Web ページの側面を検索キーとするような検索が可能になると考えられる．

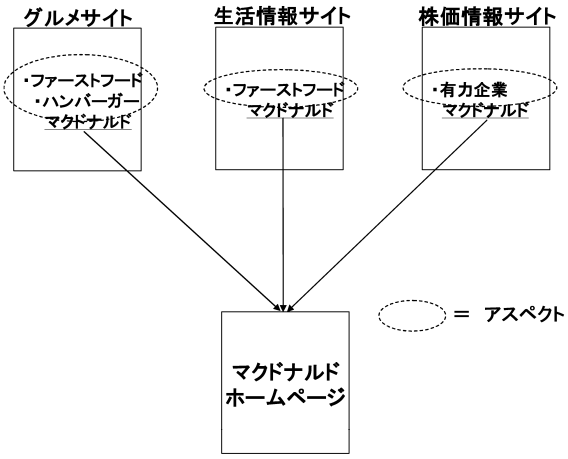

図 1　Web ページのアスペクト
Fig. 1　Aspects of a Web page.

そこで本研究では，このような Web ページの周辺情報を表すコンテンツ集合のことを “Web ページのアスペクト” と呼び [2]，その抽出手法についての提案を行う．また，提案手法に基づくプロトタイプシステムを実装し，実際の Web ページを対象としたアスペクト抽出実験を行い，その結果について考察する．

2. 本研究のアプローチ

現在の一般的な情報検索手法である，検索エンジンを用いたキーワード検索は以下のようなステップにより，目的の情報を検索する．

（1） 目的の情報に関連するキーワードを検索エンジンに入力

（2） 検索エンジンから返された検索結果のリストに表示されているタイトルや要約文を参考に，目的の情報に関連のありそうな Web ページを選択し，リンクをたどって閲覧

（3） たどった先の Web ページに目的の情報が存在しなかった場合，さらにその Web ページから張られているリンクをたどり，目的の情報を探索

（4） 目的の情報が見つからなかった場合，検索結果のリストに戻り，リストに表示されている他の Web ページに対して同様の操作を実行

（5） 検索結果のリストに表示されている Web ページをいくつか閲覧しても，目的の情報を得ることができない場合や，検索結果として返された Web ページ数が莫大で，閲覧すべき Web ページを決定することが困難である場合は，検索キーワードの追加・変更・削除を行い，その結果返された検索結果のリストに対して，同様の操作を実行

このような検索手法において，検索キーワードの決定・変更や検索結果から実際に閲覧する Web ページの決定を行うためには，目的の情報に関するある程度の予備知識が必要となる．しかしながら，検索者が予備知識を持っていない場合，適当な検索キーワードや閲覧するページの決定が困難であり，その結

果目的の情報にたどり着くまでに上記のステップを何度も踏むこととなり，大変な労力と時間を消費することになる．

このような問題を解決するための一手法として，検索結果に表示されている Web ページの内容に関する情報を同時に呈示するという手法が考えられる．このような情報は追加キーワードや，閲覧する Web ページの決定する際の判断材料となるため，目的の情報に関する予備知識を持たない場合でも，効率的な情報検索が可能となると考えられる．

このような情報として考えられる情報に， Web ページの周辺情報がある．これは，Web ページに対してリンクしている Web ページのアンカーテキストの周辺のコンテンツのことであり，このコンテンツには，リンク先の Web ページの内容がある程度含まれている．本研究では，ある Web ページに対してリンクしている Web ページ群から，このようなコンテンツを取り出し，情報検索に有用な情報を抽出する．

本研究では，このようなコンテンツ集合から抽出された情報を “アスペクト” と呼んでおり，本稿ではこのアスペクトの抽出手法について述べる．以降，アスペクト抽出の対象となる Web ページを “対象ページ p”，対象ページに対してリンクしている Web ページを “リンク元ページ q”，リンク元ページ q 中の対象ページ p へのアンカーを “キーアンカー a” と呼ぶことにする．

対象ページ p のアスペクト抽出のアプローチとしては，まず対象ページへリンクしているリンク元ページ群を収集し，その中から外部サイトへのリンクを数多く持つ “リンク集ページ” を特定する．ここで，リンク集ページとは，あらゆる Web ページ作成者によって作成された様々な Web ページへのリンクを，ある特定の目的に沿って収集・分類した Web ページのことを指す．一般にリンク集ページは，経験的にリンクアンカーの周辺，特にキーアンカー a からみて上部に対象ページの見出しが記述されていることが多い．本研究では，複数のリンク集ページからこのような見出しを複数集め，それらを分類することにより，対象ページ p のアスペクトを抽出する．

あるキーアンカー a を持つリンク集ページにおいて，対象ページ p の見出しの役割を果たすコンテンツのことをキーアンカー a の “見出しコンテンツ” と呼び， $T(a)$ で表す． $T(a)$ は，以下の条件を満たすキーワード k の集合として定義される：

【キーアンカー a の見出しコンテンツ $T(a)$ の定義】

- キーアンカー a はリンク集ページに含まれている．
- ツリー構造で表現されたリンク集ページにおいて，キーワード k はキーアンカー a の親ノードもしくは先祖ノードに含まれている．
- キーワード k は既定の “見出しタグ” に囲まれている．

ここで見出しタグとは，見出しを表すと考えられる特定のタグを指し，経験的に決められる．例えば， <Hn> $(n = 1, 2, \ldots)$ や <B> ， <I> などである． 図 2 に，キー・アンカーとその見出しコンテンツの関係の概要を示す．

さらに，対象ページに対する複数のリンク集ページにおける

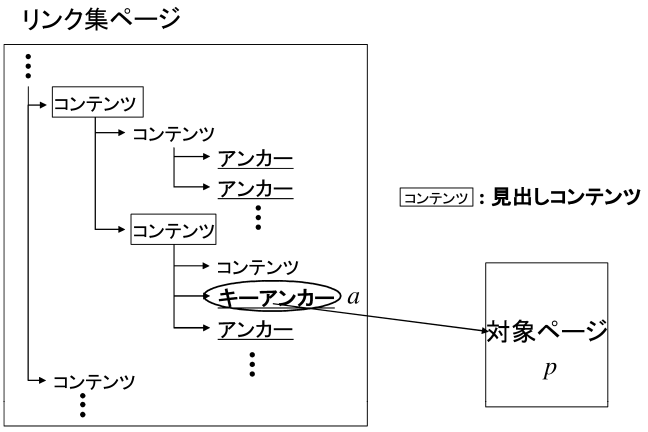

図 2　キーアンカーの見出しコンテンツ
Fig. 2　Header contents of a key anchor.

見出しコンテンツをクラスタリングする．クラスタリングの結果得られたクラスタが，対象ページのアスペクトに対応する．アスペクトを代表する語として，各クラスタの中心ベクトルを求め，そのベクトルの特徴量が高いキーワードを抽出する．

3. 関 連 研 究

リンク元のページの情報から Web ページの特徴を抽出しようとする研究としては，まず Glover ら [4] の Extended Anchor Text の研究がある．この研究では，従来のアンカーテキストだけでなく，その周辺の文字列も含めて Extended Anchor Text と定義し，語句レベルでの特徴を抽出して Web ページの特徴付け，クラスタリングおよびクラスタの名前付けなどを行っている．

しかしながら，周辺の文字列がアンカー文字列と必ずしも関連があるとは限らず，これだけでは十分とはいえない．本研究では，一般にリンク集と呼ばれている Web ページの見出しからキーワードを抽出することにより，Web 文書構造を考慮した周辺コンテンツからのアスペクト抽出を行っている点が異なる．

また，是津ら [5] は， Web から画像などのマルチメディア・オブジェクトの周辺コンテンツを Web 文書構造やハイパーリンクに基づいて抽出し，マルチメディア・オブジェクトの用例の生成を行っている．文書構造やリンクを用いて周辺コンテンツを抽出している点では本研究と関連しているが，本研究は Web ページを対象にしていることや，単に周辺を抽出するだけでなく対象ページと特定の関係（見出し）にある周辺コンテンツをアスペクトとして抽出し，対象の意味付けを行っている点が異なる．

4. アスペクトの抽出

本研究のアプローチに基づき，リンク元ページからアスペクトを抽出する手法について述べる．アスペクトを抽出するステップは以下の 3 段階に分かれる．

（1） リンク集ページの取得

リンク元ページ群の中からリンク集ページを特定し，取り出す．

（2） 見出しコンテンツの抽出

取り出したリンク集ページをツリー構造表現にし，これから見出しコンテンツを抽出する．

（3） 見出しコンテンツからのアスペクト抽出

リンク元ページごとに抽出した見出しコンテンツを分類し，各分類から対象ページの個々のアスペクトを抽出する．

以下，各ステップについて述べる．

4.1 リンク集ページの取得

アスペクトを抽出する第一段階として，まず，対象ページのリンク元ページ群を検索し，その中からリンク集ページを選択・取得する．アスペクトを抽出するためのリンク集ページは，対象ページの作者とは異なる作者・組織が作成したリンク集を対象とする．このようにすることにより，対象ページ作成者が持つ対象ページへの見解を排除し，対象ページに対する外部からの様々な見解をアスペクトとして取得できるようにする．対象ページ p に対するあるリンク元ページ q がリンク集ページであるかどうかは，以下の基準に基づいて判定される：

【リンク集ページの判定基準】

- リンク元ページ q 内に含まれるリンク数 $N_{lnk}(q)$ が既定の閾値 (θ_{lnk}) 以上である．
- リンク元ページ q 内に含まれるリンクのうち，外部サイトへのリンクの占める割合が既定の閾値 (θ_{exlnk}) 以上である．即ち，

$$\frac{N_{exlnk}(q)}{N_{lnk}(q)} \geqq \theta_{exlnk}$$

すなわち，数多くのリンクを持ち，かつそのうちの多くが外部に向けたリンクである Web ページをリンク集ページとする．ここで，$N_{exlnk}(q)$ はリンク元ページ q 内に含まれる外部サイトへのリンク数である．外部サイトへのリンクとは，リンク元ページ q が属している Web サイトとは異なる Web サイトに属する Web ページへのリンクを指す．同一 Web サイトに所属するか否かは，例えば URL のホスト名によって区別される．

4.2 見出しコンテンツの抽出

次に，4.1で取り出したキーアンカー a を含むリンク集ページの HTML 構造をツリー構造で表現する．このツリー構造を用いることにより，このリンク集ページにおける，キーアンカー a の見出しコンテンツ $T(a)$ を抽出する．見出しコンテンツとは，リンク集ページにおいて対象ページの内容をある程度記述しているコンテンツに含まれるキーワード集合のことである．

一般にリンク集と呼ばれる Web ページにおいては，各 Web ページへの複数のリンクを階層構造を持った見出しによって分類していることが多い．あるアンカー に関連する全ての見出しを抽出するためには， HTML 構造をツリー構造にし，その階層構造にキーアンカーから順に上へたどることにより，対象ページに関連するキーワード集合を収集していく必要がある．

あるリンク集ページにおいてキーアンカー a の見出しコンテンツ $T(a)$ を，ツリー構造を用いて抽出する手順は以下の通りである：

【見出しコンテンツ抽出手順】

（1） キーアンカー a を含むノード $o(a)$ を特定．

（2） ノード $o(a)$ の親ノードが見出しタグに囲まれたノードであれば，親ノードに含まれるキーワード集合 $\{k_i\}$, $(i = 1, 2, \ldots)$ を見出しコンテンツ $T(a)$ に追加．

（3） 現在のノードがルート・ノードに達すれば終了．そうでなければ 4. へ．

（4） ノード $o(a)$ の親ノードを次の $o(a)$ として 2. へ．

この結果，見出しコンテンツ $T(a)$ には，キーアンカー a を含むノードの先祖ノードに存在するキーワードの集合が得られることになる．この処理を 4.1 で取り出した全てのリンク集ページに対して行う．すなわち，見出しコンテンツ $T(a)$ は，リンク集ページと同じ数だけ抽出される．

4.3 見出しコンテンツからのアスペクト抽出

対象ページに対して様々なリンク集ページから取得された複数の見出しコンテンツを分類し，各分類から対象ページの個々のアスペクトを抽出する．

今，キーアンカー a の見出しコンテンツ $T(a)$ を，見出しコンテンツに含まれるキーワードの特徴ベクトル

$$f(T(a)) = (v_1, v_2, \ldots, v_n)$$

で表す．特徴ベクトルの各要素 v_i は，対応するキーワード k_i の見出しコンテンツ $T(a)$ に対する重要度を表す．キーアンカーの近くに現れ，かつ特定のタグに囲まれたキーワードほど重要であると考え，各キーワードの重要度を以下のように定義する：

$$v_i = \sum_{j=1}^{N_{k_i}} \frac{b}{d(a, o_j(k_i))}\, w(o_j(k_i))$$

ここで， N_{k_i} はこの見出しコンテンツ $T(a)$ におけるキーワード k_i の出現回数を表し， $o_j(k_i)$ はキーワード k_i の j 番目の出現が含まれるノードを表す． $d(a, o_j(k_i))$ はノード $o_j(k_i)$ のキーアンカー a からの距離を示し，キーアンカー a からノード $o_j(k_i)$ までの間に辿る親関係の数によって表される．また，$w(o_j(k_i))$ はノード $o_j(k_i)$ の重みを示し，ノード $o_j(k_i)$ を囲むタグ（見出しタグ）の種類ごとに予め決められた値が与えられる． b は定数である．

対象ページ p に対する見出しコンテンツ集合 $\{T(a_i)\}$ （ a_i は対象ページ p の i 番目のキーアンカー）の分類には，各見出しコンテンツ $T(a_i)$ のキーワード特徴ベクトル $\boldsymbol{f}(T(a_i))$ 間の類似度 [3] に基づく $\{T(a_i)\}$ のクラスタリングを行う．クラスタリングの結果得られた各クラスタ C_j が，対象ページ p に対する個々のアスペクト A_j に対応する．アスペクト A_j の内容は，対応するクラスタ C_j のクラスタ中心を表すキーワード特徴ベクトル $\boldsymbol{f}_{C_j}$ から特徴量（重要度）の大きい順に一定数 m のキーワードを抽出して得られるキーワード集合 $\{k_l\}\,(1 \leqq l \leqq m)$ によって表される．

抽出されたアスペクト集合 $\{A_j\}$ は，各アスペクト A_j に対応するクラスタ C_j に含まれる見出しコンテンツ $T(a_i)$ の数に基づいてランキングされる．即ち，より多くの見出しコンテ

ンツを含むクラスタから得られたアスペクトほど代表的なアスペクトと見なされ，高位にランクされる．最終的に，アスペクト抽出の結果として，各アスペクト A_j を表すキーワード集合 $\{k_l\}$ がアスペクトのランキング順に呈示される．クラスタリングの結果表示方法の概要を図 3 ， 表 1 に示す．図 3 において， f_{C_j} は各クラスタ C_j の中心ベクトルを表している．また， $f_{C_1} = v_a, \ldots, v_d, v_e$ の表記は，クラスタ C_1 の中心ベクトル f_{C_1} の構成要素のうち， v_a, v_d および v_e の特徴量が特に大きかったということを示す． 表 1 において， Aspect A_j は各クラスタ C_j に対応するアスペクト A_j ， Representative word は各アスペクト A_j を表すクラスタ C_j の中心ベクトルの中で特徴量の大きい上位 3 単語を表している．

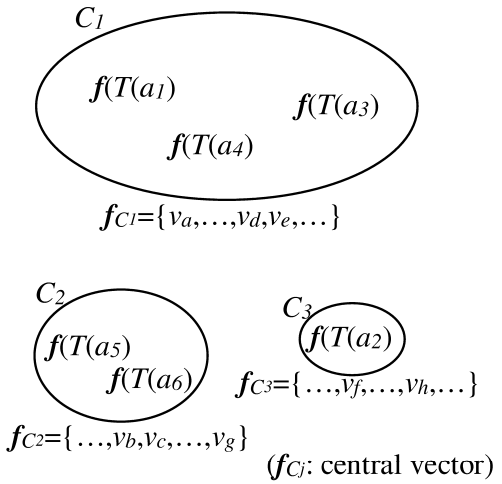

図 3　クラスタリングの結果表示方法の概要
Fig. 3　A compendium of a result of clustering.

表 1　クラスタリング結果の例
Table 1　An example of a result of clustering.

Aspect A_j	**Representative word**
Aspect A_1	k_a, k_d, k_e
Aspect A_2	k_b, k_c, k_g
Aspect A_3	$k_f, k_h, -$

5. 実　　験

本章では，プロトタイプシステムによるアスペクト抽出実験とその考察について述べる．なお，プロトタイプシステムにおいて，対象ページのリンク元ページを検索する部分については，Google Web APIs [6] を使用した．

5.1　実験の目的

以下の 2 点についての検証を行うため，プロトタイプシステムを実装し，実験を行った．

実験 1　見出しタグと HTML のツリー構造を用いた提案手法により，どの程度の見出しコンテンツを取り出すことが可能か．

実験 2　提案手法により，対象ページ内に存在しないキーワードを対象ページのアスペクトとして抽出することが可能か．

5.2　実　　験 1

5.2.1　実験 1 の内容

アスペクト抽出のプロトタイプ・プログラムを実装し，実際の Web ページを対象にアスペクト抽出の実験を行った．実験 1 のプロトタイプ実装では，提案手法に対し以下のような制限や拡張が加えられた：

- リンク集判定は行わず，全てのリンク元ページを見出しコンテンツ抽出の対象とする．
- 実際の HTML 文書においては，キーアンカーの先祖ノードだけでなく，先祖ノードの兄弟ノード（厳密には，現在のノードよりも前に存在する同階層のノード）に見出しタグを持つノードが存在する場合がある．このため，実際の抽出プログラムでは見出しタグを持つ兄弟ノードも見出しコンテンツの抽出対象とする．

- 提案手法では，リンク集ページが階層的に見出しコンテンツを持っている場合，その全てを抽出するが，実験 1 の段階ではキーアンカー a の直近の見出しコンテンツのみを抽出対象とし，それ以上の階層の見出しコンテンツは抽出しない．

- 見出しコンテンツからのアスペクト抽出では，見出しコンテンツの類似性判定は行わず，各見出しコンテンツを個々のアスペクトとして抽出する．また，見出しタグには全て同じ重みを与える．

5.2.2　実験 1 の結果

実験 1 では，有名なハンバーガー・ショップのホームページを対象にアスペクト抽出を行った．見出しタグは，<Hn > $(n = 1, 2, ...)$, <B> および <I> を対象とした．これらのタグは実験前に行った予備調査の結果において，見出しコンテンツの多くが囲まれていたタグである．表 2に，対象ページごとに取得されたリンク集ページ，およびそれらから抽出されたアスペクトの数を示す． 図 2において， “Page”, “# Link Collections” および “# Aspects” の欄は，それぞれ対象ページ，対象ページへのリンクを含むリンク集ページ数および抽出されたアスペクト数を示している．この結果からは，リンク集ページの約 30 %–50 %からアスペクトが抽出できたことが分かる．

表 2　アスペクト抽出統計
Table 2　Statistics of aspect extraction experiment.

Page	**# Link collections**	**# Header Contents**
Burger King	83	46
Mc Donalds	98	31
Subway	88	42

表 3に，プロトタイプシステムによるアスペクト抽出の結果の一部を示す． 表 3において， “Page”, “Aspect”, “Tag” および “Link Collection” の欄は，それぞれ対象ページ，アスペクト，このアスペクトが抽出された見出しタグおよびリンク集ページを示している． 表 3から，ハンバーガー・ショップのホー

表 3　アスペクト抽出結果例
Table 3　A sample of aspect extraction result.

Page	Header Contents	Tag	Link Collection
Burger King	Restaurant Websites	<H3>	ToyBidder's World
	Altri siti di cucina	<H1>	Altri siti di cucina
	Fast Food	<h3>	DefianceWeb (tm) — Dining in Defiance
	All Are Invited	<h1>	Pursuit of Excellence Marching Band Festival - Food Court Information
	BERNIE'S Where all dragonmasters drink!	<b>	Welcome to Bernie's!
McDonald's	Patron Sponsors	<b>	Helen Keller Festival Sponsors
	Slot Machine Economics	<h2>	Las Vegas
	Buttered Muffin for Breakfast	<b>	Parker Software
	Fast Food	<h3>	efianceWeb (tm) — Dining in Defiance
Subway	Healthier fast-food alternatives	<b>	Fast Food Pitfalls — ahealthyme.com
	Subway Sandwiches	<b>	the Ridge Online
	2000 Sales: $3,700,000,0002000 Earnings:	<b>	AmericanCompanies.com

ムページに対し，いわゆる食事関連のアスペクトだけではなく，“Patron Sponsors” や “2000 Sales” など経済に関連するようなアスペクトも抽出されていることが分かる．なお，今回の実験結果は，クラスタリングを行う前のものであり，実際のアスペクトは，これらをクラスタリングすることによって抽出する．

5.2.3　実験 1 の考察

実験 1 段階のプロトタイプシステムでは，リンク集ページ判定および見出し語の特徴ベクトルによるクラスタリングを行っていないため，見出しコンテンツとしては不適切だと考えられるコンテンツが含まれてはいるものの，今回の実験結果から，本研究の提案手法により，対象ページの見出しコンテンツをリンク集ページから抽出することが可能であることが分かった．しかしながら，不適切な見出しコンテンツを排除し，対象ページのアスペクトをより明確にするためには，リンク集ページ判定および抽出した見出しコンテンツの分類を行う必要がある．

また，本プロトタイプシステムにより見出しコンテンツを抽出できなかったリンク元ページについて実際に閲覧し，調査したところ，見出しコンテンツの抽出対象とすべきリンク元ページがいくつか見つかった．これらの見出しコンテンツは，本実験で設定した見出しタグに含まれない HTML タグで囲まれており，これらのタグも見出し語タグに含めることにより，見出しコンテンツを抽出することが可能になると考えられる．

5.3　実　　験 2

5.3.1　実験 2 の内容

プロトタイプシステムの改良を行い，実験 1 と同様に実際の Web ページを対象として，アスペクト抽出の実験を行った．本実験では，実験 1 で行わなかったリンク集ページの判定および，見出しコンテンツのクラスタリングを行っている．また，本実験では，実験 1 の結果を踏まえ，前回の見出しタグに加え，全 12 種類のタグを見出しタグとして設定した．

5.4　実験 2 の結果

表 4 に，神戸大学ホームページのトップページ [7] を対象ページとした，アスペクト抽出実験結果について示す．表 4 において，“# H.C.”, “Keyword”, “Title” の欄はそれぞれ，含まれる見出しコンテンツ数，代表語の上位 3 単語，およびそのクラスタに含まれるリンク集ページのタイトルを表している．代表語とは，アスペクトに対応するクラスタのクラスタ中心を表すキーワード特徴ベクトル中で，特徴量（重要度）の大きい単語のことであり，そのアスペクトを代表する語である．また，リンク集ページのタイトルには，各アスペクトに対応するクラスタのクラスタ中心に最も近いキーワード特徴ベクトルを持つ見出しコンテンツを抽出したリンク集ページのタイトルを記載している．また，タイトルが長いリンク集ページについては一部省略して表記している．

各リンク元ページ 100 ページのうち，リンク集と判定されたのは， 44 ページであった．そのうちの 18 ページから見出しコンテンツ抽出に成功し， 10 のアスペクトを抽出することができた．

5.4.1　実験 2 の考察

神戸大学の一般的なアスペクトは「大学」であると考えられるが， 図 4 より，本実験結果においても，最大のクラスタを持つアスペクトは “University(-ies)”, “College(s)” といった単語で表されており，最大のクラスタを持つアスペクトが，一般的なアスペクトを表すということがいえる．

一方,「生物」や「物理」などのキーワードがアスペクトとして抽出されているが，これらのキーワードは，対象ページには含まれていない．このようなキーワードを，例えば，検索結果と共に呈示することにより，予備知識を持たない検索者が追加キーワードの選択・修正や次に閲覧する Web ページを決定するための判断材料になりうると考えられる．

また,「関連」,「リンク」等のリンク集によく使用される単語も抽出されているが，これはストップワードとして排除すべきキーワードである．

表 4　神戸大学 HP についてのアスペクト抽出実験結果
Table 4　A result of aspect experiment about Kobe University.

	# H.C.	Representative word 1	Representative word 2	Representative word 3	Title
Aspect 1	7	University(-ies)	Kobe	College(s)	Japan Prefectures ...
Aspect 2	2	神戸大学	大阪経済大学	大阪市立大学	近畿地区
Aspect 3	2	関連	リンク	—	関西 TLO
Aspect 4	1	全国	大学	生物	全国の生物学科
Aspect 5	1	高	エネルギー	物理	Particle Physics ...
Aspect 6	1	入居	プロジェクト	一覧	自然科学...
Aspect 7	1	Asia	Far	East	BRAINTRACK ...
Aspect 8	1	Academic	Institutions	Organisations	Academic Institutions ...
Aspect 9	1	Information	—	—	Logic, Statistics ...
Aspect 10	1	National	—	—	Google Directory ...

6. まとめと今後の課題

本稿では，ハイパーリンク構造に基づく Web ページの周辺情報から抽出した，その Web ページの意味的側面を “アスペクト” と呼び，リンク集の見出しからアスペクトを抽出する手法について述べた．また，提案手法に基づき，アスペクト抽出のためのプロトタイプシステムを実装して実験を行い，その結果について考察を行った．

今後の課題として，まず，アスペクト抽出精度の向上のための改善策が挙げられる．リンク集ページの取得に関して，最近の Web ページは，デザイン性を上げるため，ページの一部分を使ってリンク集を構成することが多い．この様なリンク集では，ページ全体ではなく，ページ内の適切な部分コンテンツを “リンク集コンテンツ” として特定し，そのリンク集コンテンツからアスペクト抽出を行うようにすることで，アスペクトとは関係のないキーワードの混入を防ぐことができると考えられる．また，良い Web ページ（例えば，Authority と呼ばれる Web ページ）にリンクしている）リンク集ほど質が高いと考え，リンク集の質を考慮した特徴ベクトルの重み付けを行うなどの改善策にも議論の余地がある．また，ストップワードを導入し，今回の実験の際にアスペクトとして抽出された「関連」や「リンク」などのリンク集によく使用される単語の除去等も行う必要がある．

また， Web 探索において，探索途中で得られる Web ページのアスペクトを使って質問を修正する方法についても研究を進めていく．アスペクトによる質問修正の特徴は，欲しい対象の中身が正確に分らない場合でも，途中経過の Web ページのアスペクトから得られる用途や評判など外見的な意味情報を使って質問を修正しながら対象を絞り込んでゆけることである．そのためには，アスペクトに基づく Web ページ検索を実現する必要がある．

謝　　辞

本研究の一部は，平成 14 年度科研費特定領域研究 (2)「Web の意味構造に基づく新しい Web 検索サービス方式に関する研究」（課題番号：14019048，代表：田中克己），および，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」によるものである．ここに記して謝意を表します．

文　　献

[1] Welcome to McDonald's Japan. *http://www.mcdonalds.co.jp*.
[2] 荒木良, 中島伸介, 角谷和俊, 田中克己. Web ページのアスペクトに基づくクラスタリングとその応用. 情報処理学会研究報告 2002-DBS-128, pp. 289–296, 2002.
[3] G. Salton and M. McGill. Introduction to modern information retrieval. In *McGraw Hill*, 1983.
[4] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using web structure for classifying and describing web pages. In *Proceedings of the WWW2002 International World Wide Web Conference*, pp. 562–569, 2002.
[5] 是津耕司, 角谷和俊, 田中克己. Multimedia corpus: マルチメディアの用例のデータベース化. 情報処理学会研究報告 2002-DBS-128, pp. 367–374, 2002.
[6] Google Web APIs. *http://www.google.com/apis/*.
[7] Kobe University. *http://www.kobe-u.ac.jp*.