Web ページのアスペクトの発見

Discovering Aspects of Web Pages

是津 耕司[†‡] 木俵 豊[†] 田中 克己[‡]

Koji ZETTSU Yutaka KIDAWARA
Katsumi TANAKA

Web ページは他の様々な Web ページからリンクによって参照されており，どのような内容から参照されているかを調べることでそのページの Web 上での見られ方を知ることができる． こうした Web ページの外面的な意味を，我々は Web ページの“アスペクト”と呼ぶ．本論文では，アスペクトの基本概念とその発見方法について説明する．提案手法では，Web の論理構造に基づき，Web ページへの参照を特徴付ける適切な範囲の Web コンテンツをそのページへの参照コンテキストとして抽出し，類似したコンテキストをクラスタリングすることで，典型的な参照内容を表すキーワードをアスペクトとして抽出する．アスペクトの発見は，あるページの Web 上における位置づけや評判を把握するのに有用であると考えられる．

Considering that a web page is referred to by other pages in various contexts through links, these contexts indicate the reputation of the page. Such references are called “aspects” of the web page, as distinguished from the content of the page. In this paper, we propose an approach for discovering aspects for characterizing web pages based on their context. Based on the logical structure of the web, our approach discovers the appropriate range of surrounding contents and assigns them as the context of the web page. The aspects of the web page are discovered by clustering multiple contexts so that each aspect represents “typical reference” to the page. The idea of aspect provides a context-based means of distinguishing web pages from each other. The paper explains the details of the aspect discovery method and its experimental implementation.

1. はじめに

Webの普及により，今日Webは重要な情報源の１つとなっている．Web上には，様々なユーザによって公開された大量のWebページが存在している．Webの特徴は，Webページが相互にハイパーリンクによって関連付けられていることであり，あるWebページは他のWebページから様々な内容で参照されている．例えば，ある企業のWebページは，一般的に，その企業の製品やサービスに関する情報を載せている．しかし，他のWebページからは，地元の優良企業として参照されていたり，ある共同研究のメンバーとして参照されていたりすることがある．これら“地元の優良企業”や“共同研究のメンバー”は，このWebページ（あるいは企業）が外部からどう見られているかを表している．

このように，あるWebページには，その中身に関するセマンティクスだけではなく，他のページからどのような内容で参照されているかというセマンティクスも存在する．我々は，こうした参照に基づくセマンティクスを，Webページの外面的な意味ということで**アスペクト**と呼ぶ．Webページのアスペクトは，Web上におけるページの社会的位置付けや評判を表していると考えられる．例えば，ある企業のWebページのアスペクトは，その企業のWebにおけるブランド戦略に役立つと考えられる．また，ある製品やサービスに関するWebページのアスペクトは，それらのWeb上における評判やトレンドを分析するのに役立つと考えられる．

本論文では，Webページのアスペクトのコンセプトを提案すると共に，Webからアスペクトを発見する方法を提案する．第2章では，あるWebページへの参照を表すコンテキストを，Webの論理構造に基づいて定義する．第3章では，参照コンテキストの抽出方法について説明する．第4章では，抽出されたコンテキストからアスペクトを発見する方法について説明する．第5章では，プロトタイプの実装と評価実験について述べる．最後に関連研究と今後の課題について述べる．

2. Web ページのコンテキスト

あるWebページを参照する**コンテキスト**は，Webページへのリンクアンカーとその周辺に存在するWebコンテンツによって表される[1]．単語のコンテキスト[2]とは異なり，Webページのコンテキストは，Webを構成するHTMLやXMLなどの文書構造とリンク構造の影響を受ける．例えば，上位段落のコンテンツは下位段落のコンテンツの主題を表したり，リンク元のコンテンツはリンク先のコンテンツの分類（リンク集など）を表したりする．したがって，Webページのコンテキストは，Webの論理構造を反映した周辺コンテンツによって表される．

Webの論理構造は，HTMLやXMLなどによって定義される文書要素のツリー構造と，それら文書要素のリンクによってモデル化される．このモデルは，各文書要素をノード，文書要素間の親子関係もしくはリンク関係をエッジで表した有向グラフ$G = (V, E)$によって表すことができる（V, E は，それぞれノード集合およびエッジ集合）．図１に例を示す．Webの論理構造に基づき，あるWebページpの参照コンテキストを$c(p)$は，以下の条件を満たすGの部分木 $c(p) = (V_c, E_c)$ $(V_c \subset V, E_c \subset E)$ として定義される：

1. $c(p)$ はページpへのリンクアンカーaを含む．
2. ある文書要素$v \in V_c$に対し，以下のいずれかを満たす文書要素$v' \in V_c$が存在する：
   - vはv'の親，子，もしくは兄弟に位置する文書要素である．
   - vはv'のリンク元の文書要素である．
3. どの２つの文書要素$(v, v') \in V_c$も，コンテキスト内のある１つの文書内の文書要素に同時にリンクしていない．

コンテキストは，直感的には，対象ページへのリンクアンカーを基点に，親子・兄弟およびリンク元の文書要素へと広がっていく．ただし，コンテキストの一貫性を保つため，文書要素への多重リンクが存在する場合，各々のリンクごとに独立したコンテキストとして分岐させる． 図１の例では，灰色のノードとそれらの間のエッジによって構成される部分

† 正会員 独立行政法人通信総合研究所
{zettsu, kidawara}@crl.go.jp
‡ 正会員 京都大学大学院情報学研究科
{zettsu, tanaka}@dl.kuis.kyoto-u.ac.jp

木が，Webページpのコンテキスト$c(p)$を表す．

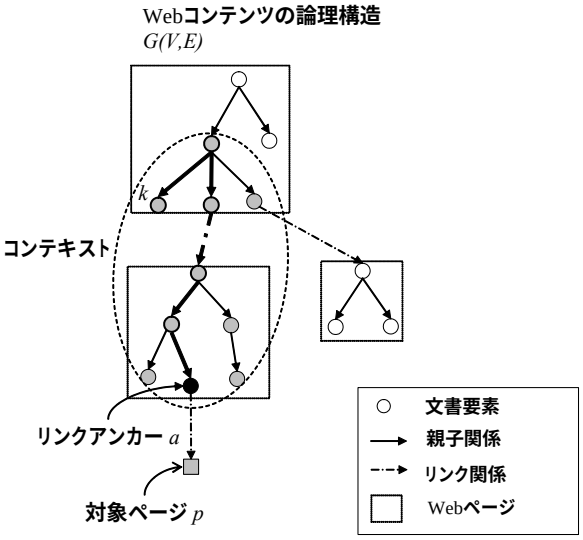

図１ Web ページのコンテキスト

Fig.1 Context of Web page.

3. コンテキストの抽出

コンテキストの抽出では，適切なコンテキスト範囲を特定しなければならない．以下の節で，具体的な方法について述べる．

3.1 コンテキスト・キーワードの抽出

あるコンテキストは，以下のようなキーワードによって特徴付けられる：

1. 対象ページへのリンクアンカーの近くに出現する．
2. そのコンテキストにのみ頻繁に出現し，他のコンテキストにはあまり出現しない．

第１の基準は，コンテキストに含まれる各キーワードとリンクアンカーaとの間の距離に基づいて評価される．あるキーワードkとリンクアンカーaの間の距離は，コンテキストを表す部分木において，kを含む文書要素とリンクアンカーaを結ぶパスの長さとして表される．図１では，kと記されたノードとaのノードとをつなぐパス（太線）で表される．一般にあるキーワードは複数の文書要素に出現するため，実際の距離は，キーワードの各出現とリンクアンカーとの距離の平均値として以下のように求められる：

$$d(k,a)=\frac{\sum_{i=0}^{n_k}(w_s d_s(o_i,a)+w_l d_l(o_i,a))}{n_k}$$

ここで，o_iは，キーワードkの各出現を表す．$d_s(o_i, a)$および$d_l(o_i, a)$は，それぞれ親子関係とリンク関係に基づく距離を表し，w_sおよびw_lは各々の距離に対する重み付けを表す．

第２の基準を評価するために，我々は*tficf*値 (text frequency, inverted context frequency)を定義する．*tficf*値は，従来のテキスト検索における単語の*tfidf*値[3]と同じ考え方に基づいており，以下のように求められる：

$$tficf(k)=n_k\cdot\log\frac{M}{m_k}$$

ここで，n_kはキーワードkのコンテキスト内での出現頻度を表し，Mおよびm_kはそれぞれ全てのコンテキスト数とそれらの中でキーワードkが出現するコンテキスト数を表す．

上記２つの評価基準に基き，あるキーワードkがコンテキストを特徴付ける度合いを表す**文脈貢献度**を定義する．文脈貢献度は，以下のように求められる：

$$ccd(k)=\frac{1}{d(k,a)}\cdot tficf(k)$$

あるコンテキストの内容は，文脈貢献度の高いキーワード集合によって特徴付けられる．これらのキーワードを**コンテキスト・キーワード**と呼ぶ．

3.2 コンテキスト範囲の特定

コンテキスト範囲の特定では，対象ページへのリンクアンカーを基点に周辺範囲を徐々に広げながら，コンテキストの内容が大きく変化する点を見つけ出す．したがって，抽出されたコンテキストは，ある意味的なまとまりを表している．コンテキスト範囲の特定は，以下の手順に従って行われる：

1. リンクアンカーaをコンテキストの初期値$c_0(p)$に設定する．
2. 現在のコンテキスト範囲に含まれる文書要素の親子・兄弟要素およびリンク元要素へとコンテキスト範囲を拡大し，新しいコンテキスト範囲とする．
3. 現在のコンテキスト範囲$c_i(p)$からコンテキスト・キーワード集合を抽出する．
4. 1つ前のコンテキスト範囲に対するコンテキスト・キーワード集合と現在のコンテキスト範囲に対するコンテキスト・キーワード集合の変化を以下のように求める：

   $$shift(c_i(p))=1-similar(\mathbf{K}(c_{i-1}(p)),\mathbf{K}(c_i(p)))$$

   ここで，$\mathbf{K}(c_i(p))$は，コンテキスト$c_i(p)$のコンテキスト・キーワードを要素，その文脈貢献度を要素値とするキーワード・ベクトルを表し，*similar()*関数はキーワード・ベクトル間のコサイン相関値[4]に基づく類似度を表す．
5. 1つ前のコンテキスト範囲と現在のコンテキスト範囲での，コンテキスト・キーワード集合の変化量の違い（変化の急激さ）を，以下のように求める：

   $$\delta_{shift}(c_i(p))=\max((shift(c_{i-1}(p))-shift(c_i(p))),\ 0)$$

6. $\delta_{shift}(c_i(p))$が与えられた閾値θ_{range}を超えた時点で，現在の範囲のコンテキストを結果として返す．そうでなければ，2から繰り返す．

4. アスペクトの発見

ある Web ページが異なる複数のコンテキストから参照されている場合もあれば，異なる Web ページが類似したコンテキストから参照されている場合もある，アスペクトは，類似したコンテキストを要約した記述であり，個々のコンテキストに比べより抽象化された参照内容を表す．以下の節で，コンテキストからアスペクトを発見する具体的な方法について述べる．

4.1 コンテキストのクラスタリング

与えられたWebページ集合に対し，類似したコンテキストをクラスタリングしてアスペクトを発見する．図 2 に，基本的な考え方を示す．各クラスタZ_jが，個々のアスペクトA_jに相当する．

提案手法では，クラスタリング手法として最大距離アルゴリズム(maximin-distance algorithm)[5]を用いる．各コンテキストはコンテキスト・キーワード集合によって表され，距離はキーワード・ベクトルの類似度に基づいて評価される（3.2 節参照）．このクラスタリング手法では，既存のクラスタから一番遠いコンテキストを見つけ，その距離が十分に遠い場合に，そのコンテキストを中心とした新しいクラスタを作成する．最も離れた 2 つのクラスタ間の距離に対する比率をパラメータとして与えればよく，学習の必要がないことが特徴である．クラスタ中心は，クラスタに含まれるコンテキストの平均キーワード・ベクトルによって表され，これが対応するアスペクトA_jの記述となる．

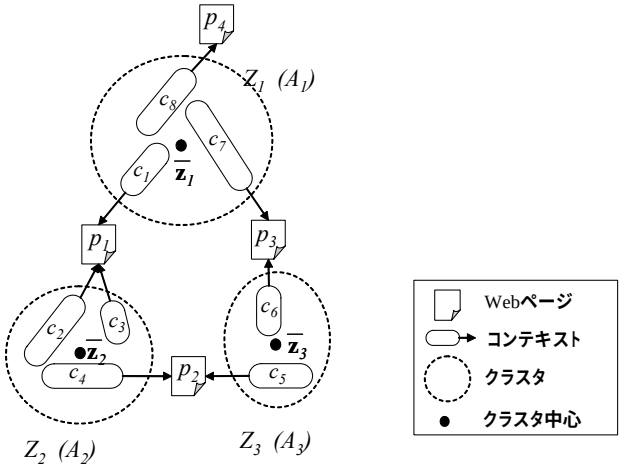

図 2 コンテストのクラスタリングによるアスペクト発見

Fig. 2 Discovering aspects by clustering contexts.

4.2 アスペクトとの関連性の強さ

Web ページとアスペクトとの間の関連性の強さを評価する基準として，我々は**アスペクト確信度**(aspect confidence)と**アスペクト支持度**(aspect support)を定義する．これらは，データマイニングにおける相関ルールの確信度と支持度[6]に基づいており，以下のようにして求められる：

アスペクト確信度：

ページp_iのコンテキスト集合のうち，conf(p_i⇒A_j)%のコンテキストがアスペクトA_jのコンテキスト集合（クラスタZ_jに含まれるコンテキスト集合）に含まれている場合，ページp_iはアスペクトA_jとconf(p_i⇒A_j)%の確信度で関連付けられている．

$$\mathrm{conf}(p_i \Rightarrow A_j) = 100 \times \frac{|C(p_i) \cap Z_j|}{|C(p_i)|}$$

アスペクト支持度：

全てのコンテキストのうち，supp(p_i⇒A_j)%のコンテキストがページp_iのコンテキスト集合とアスペクトA_jのコンテキスト集合の両方に含まれていれば，ページp_iはアスペクトA_jにsupp(p_i⇒A_j)%の支持率で関連付けられている．

$$\mathrm{supp}(p_i \Rightarrow A_j) = 100 \times \frac{|C(p_i) \cap Z_j|}{\left|\bigcup_i C(p_i)\right|}$$

ここで，$C(p_i) = \{c_k\}$は，ページp_iに対するコンテキスト集合を表す．$|C(p_i) \cap Z_j|$は，ページpのコンテキスト集合とアスペクトA_jに対応するクラスタZ_jのコンテキスト集合との積集合に含まれるコンテキスト数を表す．例えば，図 2 において，p_1⇒A_2のアスペクト確信度とアスペクト支持度は以下のように計算される：

$$\mathrm{conf}(p_1 \Rightarrow A_2) = 100 \times \frac{|\{c_2, c_3\}|}{|\{c_1, c_2, c_3\}|} = 67\%$$

$$\mathrm{supp}(p_i \Rightarrow A_j) = 100 \times \frac{|\{c_2, c_3\}|}{|\{c_1, \ldots, c_8\}|} = 25\%$$

5. 評価実験

5.1 プロトタイプの実装

本研究で実装したアスペクト抽出のプロトタイプの詳細について，以下にまとめる．

- Web の論理構造の解析では，HTML で定義されているタグの中から文書の論理構造を表すために使われるタグを選択し，それらを使って Web ページの文書構造を再構成する．表 1 に，プロトタイプで用いたタグを示す．
- 英語のキーワードを対象に，コンテキスト抽出およびアスペクト発見を行った．Web ページ内のテキストを単語に分割し，ストップワード除去，語幹抽出を行って得られた単語をキーワードとして用いている．
- 同一サイトからのリンク参照は，外部からの見られ方を表すアスペクトの価値を損なうと考えられるため，コンテキスト抽出から除外する．同一サイトは，URL のホスト名の一致によって判断する．

タグ名	分類
html, body	ページ全体と本文
table, th, tr, td	表
ol, ul, li, dl, dd, dt	リスト
p	段落
a	リンクアンカー
h1, ..., h6	見出し

表 1 文書構造解析に使用する HTML タグ

Tab. 1 HTML tags for document structure analysis.

5.2 実験結果

相互に競合すると見なされている Web ページどうしがアスペクトによってどう特徴付けられるかを示すため，コンピュータ科学で著名な 3 つの大学（Stanford University, MIT, UCB）のホームページを対象にアスペクトの発見を行った．図 3 に結果を示す．実験結果では，全てのページが{college, state, university}というアスペクトに関連付けられている．一方，{web, w3c, access, active, organization}と{w3c, signature, xml, note}という 2 つのアスペクトは MIT のページにのみ関連付けられ，MIT と他のページを差別化してい

る．実際，MIT が Web 技術で著名な標準化団体である W3C の中心的メンバーである（W3C のオフィスが MIT にある）ことを考えると，これは合理的な結果である．このように，アスペクトを発見することにより，外部からの見られ方に基づいて Web ページを特徴付けることが可能である．

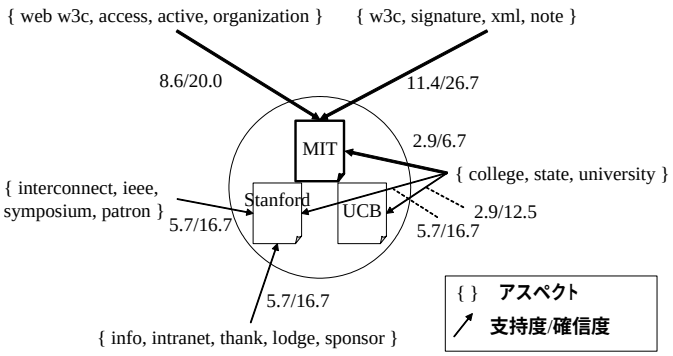

図 3　アスペクトによる Web ページの差別化

Fig. 3　Distinguishing Web pages by aspects.

6. 関連研究

Web ページの位置付けを発見する技術として，Web の評判検索[7]は，テキストマイニング技術に基づき，ある商品の評判を記述しているテキストを Web から抽出する． また，Web コミュニティ発見[8, 9]は，リンクによって相互に強く関連付けられたページ集合を発見する．さらに，HITS[10]は，リンク関係の偏りを手がかりに，相対的に重要なページ(hub/authority)を発見する．これら従来手法と比較し，提案手法では，Web の論理構造（文書構造とリンク構造）に基づいて Web ページと Web コンテンツとの間の相関関係を発見し，Web ページをその参照内容で特徴付けている点が異なる．

7. まとめと今後の課題

本論文では，Web ページのアスペクトの基本概念を示すとともに，アスペクトの発見方法について説明した．提案手法では，Web の論理構造に基づき，Web ページへの参照を特徴付ける適切な範囲の Web コンテンツを参照コンテキストとして抽出する．そして，類似したコンテキストをクラスタリングすることで，典型的な参照内容を表すキーワードをアスペクトとして抽出する．また，プロトタイプによる評価実験では，アスペクトにより競合する Web ページを差別化できることを示した．今後は以下の課題に取り組む：

- より質の高いコンテキストの抽出方法を提案する．例えば，集約やナビゲーションなどより意味的な関連性を反映したコンテキストの抽出を行う．
- アスペクトの表現を，現在のように Web ページのキーワードをそのまま使うのではなく，より抽象的な表現に置き換えること行う．

[謝辞]

本研究は，一部平成 15 年度科研費特定領域研究「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：15017249　代表：田中克己）による．ここに記し謝意を表します．また，本研究は，一部独立行政法人通信総合研究所と京都大学の共同研究「インターネット・コンテンツの意味構造発見に基づく新しいコンテンツ検索・配信方式の研究」による．ここに記し謝意を表します．

[文献]

[1] Google Webquotes: http://labs.google.com/cgi-bin/webquotes.
[2] H. P. Luhn: Keyword in context index for technical literature (kwic index). American Documentation, No. 11, pp.288–295 (1960).
[3] Salton, G., Buckley, C.: Term weighting approaches in automatic retrieval, Information Processing and Management. Volume 24, pp.513–523 (1988).
[4] Salton, G., McGill, M.: Introduction to modern information retrieval, McGraw Hill (1983).
[5] Tou, J.T., Gonzalez, R.C.: Pattern Recognition Principles, Addison-Wesley Reading (1974).
[6] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Morgan Kaufmann (2000).
[7] Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining product reputations on the web, Proceedings of the 8th ACM Conference on Knowledge Discovery and Data Mining, pp.341–349 (2002).
[8] Gibson, D., Kleinberg, J., Raghavan, P.: Inferring web communities from link topology, Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, pp. 225–234 (1998).
[9] Kitsuregawa, M., Toyoda, M., Pramudiono, I.: Web community mining and web log mining:commodity cluster based execution, Proceedings of the 13th Australasian conference on Database technologies. Volume 5, pp.3–10 (2002).
[10] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46 pp.604–632 (1999).

是津 耕司　Koji ZETTSU
1992 年東京工業大学工学部情報工学科卒業．1992 年日本アイ・ビー・エム株式会社．2003 年 独立行政法人通信総合研究所専攻研究員．現在，京都大学大学院情報学研究科博士後期課程に在学中．マルチメディアデータベース，情報検索の研究開発に従事．日本データベース学会，情報処理学会，ACM 各会員．

木俵 豊　Yutaka KIDAWARA
独立行政法人通信総合研究所主任研究員．1999 年神戸大学大学院自然科学研究科情報メディア科学博士後期課程修了．工学博士．次世代インターネットアプリケーションの研究に従事．日本データベース学会, 情報処理学会, IEEE 各会員．

田中 克己　Katsumi TANAKA
1974 年京都大学工学部情報工学科卒業．1976 年同大学院修士課程修了．1979 年神戸大学教養部助手，1986 年同大学工学部助教授．1994 年同大学工学部教授（情報知能工学科）．1995 年同大学大学院自然科学研究科情報メディア科学専攻専任教授，2001 年京都大学大学院情報学研究科社会情報学専攻教授，現在に至る．工学博士．主にデータベースの研究に従事．日本データベース学会，人工知能学会，日本ソフトウェア科学会，IEEE，ACM 等各会員．