大規模サイトにおけるWebページへの公式度の格付けに関する検討
Approach to Scoring Web Pages with Authorization on a Large-Scaled Site

堤智也* 平林真実‡ 大月一弘¶
Tomoya TSUTSUMI† Masami HIRABAYASHI§ Kazuhiro OHTSUKI‖

1 はじめに

WWWの登場をきっかけとしたインターネットの発展により，従来からの柔らかい情報に加えて，企業や官公庁などの発信する情報が急速に増加している．このような環境により，インターネットは個人型の柔らかい情報と企業型の硬い情報，さらにその中間形態にあたる情報が融合した情報メディアとなってきている．しかしながら，インターネットにおける情報発信の中心となっているWWWでは巨大化し融合しつつある情報に対して適切な情報の取得が難しくなってきている．

本研究では，インターネットの持つ柔軟性と自由度を保持しつつ，複雑化する情報への効果的なアクセスを目指し，構造と信頼性に注目する．組織構成員が各自の権限で情報発信を行なえる融合型の大規模組織サイトに対して，公式度を導入し，組織構造をWeb構造に反映させたデータベースとリンク特性を利用することで公式度に基づいた精度の高い情報発信を行うための評価方式について検討する．

2 公式度の定義

信頼度の一要素として公式度は，権威付けと保証機構と考えることができる．本システムでは構造的情報にのみ注目し組織構造による格付けとリンクによる保証機構によって公式度を決定している．組織の上下関係（事業部→課→個人など）による格付けの違いと各組織の責任において指定された自組織ノードへの格付けから，公式度を含む組織構造をWeb構造に反映するためデータベースを用意し基礎的な公式度として各組織に与えることができる．リンクによる保証機構ではリンク特性を利用することで，下部組織では自組織の責任において作成されたページ群の意味関係の構造化を行ない，同時にサイト全体でリンクを介してのページ間の格付けの伝播を行なうことでサイト全体の公式度が決定される．リンクには作成者の意図によりいくつかの意味があり，このリンク特性によって保証機能にも違いが生じてくる．格付けの伝播ではリンク元からの保証によりリンク先の格付けが上がる場合もある．

3 組織構造データベース

Webデータとは別に，組織構造データベース(StDB)を準備する．StDBは，Webページ作成者（あるいは部署）の組織内における位置付けに応じた公式度をWebの構造に射影する．これは各部署の公式度の基礎点数となる．一般には，上位に位置する部署の権限で作成されたWebページには高い公式度点が与えられる．

Webデータは，通常，各部署ごとにサブサイトやディレクトリでまとめられてサーバに格納されているため，URLをみれば比較的容易にページ作成者（部署）の判別ができる．例えば，後述の例では，各部署のトップページのURLとその部署に対応する公式度点数のみを記述しているだけである．

4 リンクのもつ意味の定義

リンクによるWebの構造を明らかにするためリンクの分類を行った．様々な機能を持つリンクの特性を分類し，Classified Hyper Link (CHL) として定義する．例えば以下のようになる．

- ページ作成者・部署本人へのリンク

  official 公式な情報へのリンク
  equivalent 同等な情報へのリンク
  personal 非公式な情報へのリンク

- ページ作成者・部署外へのリンク

  endorse 裏書保証する情報へのリンク
  recommend 推薦（内容を関知している）情報へのリンク
  introduce 紹介（内容を保証しない）情報へのリンク

リンク元の公式度ポイント並びにCHLに基づいてこのページからのリンク先に対する公式度点数が伝播される．

5 公式度計算方式

1. すべてのノードからなるグラフを用意する．グラフにおけるリンクはグラフ内の閉じたリンクでありCHLから導出される値を持っている．
2. 組織構造データベースのエントリを適用することで各組織単位のトップノードに公式度を割り当てる．公式度は仮のものであり，より公式度の高いノードからのリンクによって公式度が上昇することもある．
3. グラフ中で最大の未確定公式度をもつノードから順に公式度を確定する．その確定ノードをリンク元とするリンクを辿り，リンク先ノードに公式度を代入する．その代入の際，リンク先ノードの既存の公式度を P_d，リンク元ノードの公式度を P_s，リンクに指定された減衰量を att とすると，リンク先ノードの公式度 P は次のものとなる．

$$P = max(P_d, P_s - att)$$

4. 未確定ノードが残っていれば 3. に戻る．

*神戸大学大学院総合人間科学研究科
†Graduate School of Human Science, Kobe University
‡岐阜県立国際情報科学芸術アカデミー
§International Academy of Media Arts and Sciences
¶神戸大学国際文化学部
‖Faculty of Cross-Cultural Studies, Kobe University

ページ名	責任	Nmz	URL	HOP	StDB	提案方式
T:F501	学生	58	5	2	3	7
T:テレコム	学生	128	5	3	3	9
T:vninfo	学生	287	5	5	3	3
F:NGO	学生	658	4	2	3	4
文シ:体験記	講座	17	3	2	8	5
広報:研究者総覧(O)	委員会	34	3	2	9	10
O:研究概要	教官	30	4	4	6	7

表 1: 各パターンで得られ指数の例（検索語: 「震災」「ボランティア」）

6 評価システム

評価システムは Linux 2.0.36 上で稼働している．非対話型ネットワークファイル取得ツール GNU Wget を用い，公式度計算は Perl 言語による自作スクリプトを用いた．

1. wget を用いてサイト全体をミラーし，ディレクトリ形式でファイルシステムに格納する．また，組織構造データベースを記述する．
2. 各ページをグラフのノードと看做し，これらに対し一意なノード ID を付加しノードデータベースとする．
3. 各ページからリンク情報を抽出し，リンク元ノード ID，リンク先ノード ID，CHL リンククラスをリンクデータベースに格納する．

以下では提案方式を評価するために検索システムとして全文検索システム Namazu を利用して他の方式による Web ページの評価と比較する．

今回のシステムの評価のために以下の 6 パターンの条件において検索実験を行った．なお，項目名の括弧内は以下の表 1 の項目を示す．

Namazu のみ (Nmz) Namazu のみを利用した語句検索（元データ）．数値は出現頻度などに基づく Namazu のスコア．

URL 深さ (URL) URL のディレクトリ階層の深さ（スラッシュの数）

hop 数 (HOP) CHL を用いずにリンクの hop 数のみから公式度を計算

組織構造データベース (StDB) 組織構造データベースによって指定された公式度を利用．数行程度の組織構造データベースを記述する．

提案方式 CHL と組織構造データベースの両方を利用し公式度を計算する．

なお，URL 深さと hop 数は他と異なり小さいほど評価が高いことを示す．

7 結果

実験は約 3000 ファイル（HTML のみ）のサイトでおこなった．計算処理時間は，「準備」処理と「公式度計算」処理との合計で 1 分以内であった．また CGI を用いた検索処理は数秒で終了した．表 1 に検索語「震災ボランティア」で検索した際の結果を示す．

図 1 に今回の実験対象となったサイトの構造を示す．

T:F501 学生 T の震災ボランティアの活動記録．位置付け: 個人的なメモ．しかし，「震災時の学生の活動記録」ページからリンクされている．

T:テレコム 学生 T が特定研究の一貫として作成した共著論文．懸賞論文に応募した．位置付け: 学部の研究プロジェクトの成果物．

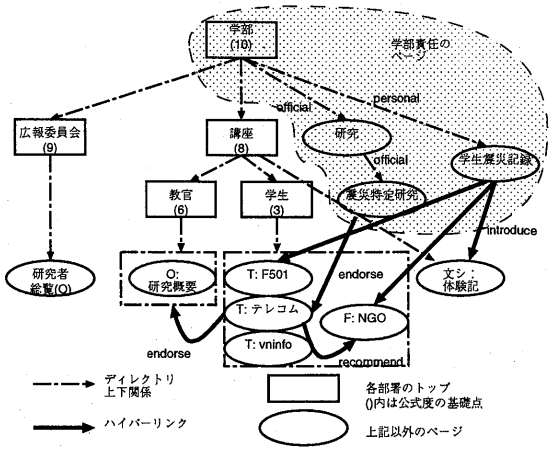

図 1: 対象とした Web サイトの模式図

T:vninfo 学生 T がボランティア活動の中で作成した文書の転載．位置付け: 個人的メモ．外部からはリンクされていない．

F:NGO 学生 F が作成した「阪神大震災地元 NGO 救援連絡会議」ページ．位置付け: 学生の個人的メモ．「T: F501」同様，「震災時の学生の活動記録」ページからリンクされている

文シ:体験記 文化システム論講座の学生によるレポート．位置付け: 講座公式ページ．「震災時の学生の活動記録」からもリンクされている．

広報:研究者総覧 (O) 教官 O の研究者総覧．位置付け: 学部広報委員会公式ページ．

O:研究概要 教官 O が作成した自己紹介的な研究概要．位置付け: 教官ページ．

実験の結果，学部の公式ページである研究者総覧についてみると Namazu 方式では語句の出現頻度が低いために高いスコアにならなかった．一方，提案方式や組織構造データベース方式では高い値を得ており，また URL と hop 数の方式でも低い値をとる（高い評価）ことが分かる．ここから，組織的に高い位置付けのページは Namazu 以外の方式でいずれも上位に位置することが分かる．

一方，F:NGO は Namazu 方式と HOP で高い評価を得ているが他では低い位置付けになる．これは単なる学生の書いたページであるため，組織的な位置付けは低い．URL・組織構造データベース・提案各方式はいずれもこの点を満たしている．

T:テレコムについて見ると，提案方式と組織構造データベース方式との間に大きな差がある．これは学部公式ページである「震災特定研究」から endorse リンクが張られていることが原因になっている．単なる組織構造データベースの適用では学生ページに過ぎないが，endorse リンクにより実際に特定研究の一環のページであることを反映して高い評価値が得られている．

これらのことから，組織構造データベースとリンク探索を併用する提案方式を用いると，組織の中で高位に位置付けられるページのうち各部署がもつ基本公式度の高いページと組織的に低位の部署であっても高位の部署から強い保証機能を持つリンクが張られているページの双方ともが上位に浮かびあがることが分かった．

以上より，提案方式はサイトの情報発信の精度を高めるだけでなく検索時の並べ替え手段としても充分に機能すると考えることができる．