



Webページの信頼性の評価手法に関する研究

平林, 真実

(Degree)

博士 (工学)

(Date of Degree)

2004-03-31

(Date of Publication)

2010-02-23

(Resource Type)

doctoral thesis

(Report Number)

甲3075

(URL)

<https://hdl.handle.net/20.500.14094/D1003075>

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。



博士論文

Web ページの信頼性の評価手法に関する研究

平成 16 年 3 月

神戸大学大学院自然科学研究科

平林 真実

目次

第 1 章	序論	1
1.1	研究の背景と目的	1
1.2	関連研究	5
1.3	本論文の構成	10
第 2 章	社会的信用に基づく Web ページ評価手法の構築	15
2.1	はじめに	15
2.2	社会的信用とページの信頼性	16
2.2.1	ページ信頼性の判断要因	17
2.2.2	Web における社会的信用	18
2.2.3	リンクによるページ評価	19
2.3	ページ評価手法	20
2.3.1	ページ評価の伝播	26
2.3.2	リンクによる推薦の解釈	28
2.4	ページスコア計算方法	29
2.4.1	評価点の導出	29
2.4.2	アルゴリズム	30

2.5	適用可能性と議論	31
2.6	まとめ	33
第 3 章	Web サイトの特徴を利用した評価手法のサイトへの適用	35
3.1	はじめに	35
3.2	サイトにおける信頼性	36
3.2.1	サイトの種別	36
3.2.2	ページの信頼性と責任	38
3.3	適用手法	38
3.3.1	社会的信用の記述方法	39
3.3.2	リンク推薦度の記述方法	40
3.4	実験	42
3.5	議論	50
3.5.1	サイト構造と組織構造	50
3.5.2	様々なサイトへの適用可能性	51
3.6	応用の可能性	53
3.6.1	キーワード検索との連携	53
3.6.2	可視化ツール	54
3.7	まとめ	54
第 4 章	良いページ悪いページの判定手法	59
4.1	はじめに	59
4.2	良いページと悪いページ	61
4.2.1	良いページと悪いページの基準	61

4.2.2	ページの直接的評価	62
4.2.3	リンクによる間接的評価	63
4.2.4	リンク関係によるページ判定	64
4.3	ページ評価モデル	65
4.3.1	リンク関係と評価に対する仮定	65
4.3.2	ページ状態の判定	66
4.3.3	漸化的なページの判定	70
4.4	ページ評価手法	75
4.4.1	良いページ/悪いページの初期集合	75
4.4.2	ページとサイト	75
4.4.3	実験	76
4.5	考察と議論	78
4.5.1	仮定に矛盾するサイト	79
4.5.2	初期ページ指定の扱い	81
4.6	まとめ	82
第5章	結論	83
謝辞		87

目次

2.1	リンク関係を用いたページ評価例	22
2.2	複数リンクがある場合の評価例	23
2.3	経路が2つある場合の評価例	24
2.4	評価点の採用の判断	25
2.5	評価の伝播木	27
3.1	公式度定義ファイルの例 (一部)	41
3.2	各方式の評価点	45
3.3	自己管理付き評価点	48
3.4	他作成者ページへの公式度の伝播	50
3.5	組織構造とサイト構造	51
3.6	公式度の可視化	55
3.7	組織構造の可視化	56
4.1	Web のグラフとその状態	67
4.2	到達可能ノードの行列	72

表目次

3.1	作成者の種別と公式度	40
3.2	リンク属性一覧	42
3.3	ページ評価の例	44
3.4	自己管理付き評価の例	47
3.5	サイトのタイプと適用可能性	53

第 1 章

序論

1.1 研究の背景と目的

WWW が誰でも情報発信を行えるメディアとしての地位を確立するなか，個人の判断で情報発信が行えるという特徴ゆえに，正確でない情報，主観に頼った情報，内容が更新されていない情報などが数多く存在し，Web ページのコンテンツの信頼性が問題となっている．実際，WWW には信頼性の判断ができない小さな企業が多数存在しており，その中の企業のページを信じて取引したことによるトラブルなども発生している．さらに，インターネット上の情報の多くを占めている個人の発信する情報は，どのような人が書いた内容なのかが判断できなかつたり，内容が古いままであったりするなど，内容がその程度信用できるのかわからないものが多い．このため，インターネットでコンテンツを探す際に利用されている検索システムにおいても，信頼性を考慮して各ページにスコアを付ける方法が一般的になってきている．このようなシステムは広く利用され，コンテンツの判断において一定の成果を挙げている．しかし，スコアが高いページが必ずしも信用できるページであるとはいえず，信頼性判断には十分ではない．結局，内容の信頼性の判断は，内容を読んだ個々による判断に任されているのが現状である．

本研究では、Web ページの信頼性を評価する方法について検討する。一般に WWW においてコンテンツの信頼性を評価する際には、次のような問題があると考えられる。

1. 信頼性の基準となるようなものが一意に決定することが難しい

誰が書いたものなら信用できる、どのような考え方を支持するかなど、信用の基準は様々な観点から設定可能であり、評価する人により異なる可能性がある。

2. 内容からのみからコンテンツの信頼性を判断することが難しい

内容からはページが関連するテーマや伝えたい内容については判断することができるが、その内容について詳しく知っている場合やページの作者を良く知っている場合を除いて、内容が信用できるのかといった信頼性を評価することは難しい。したがって、キーワード検索や構文解析などは、「目的とする内容を記述してあるページ」を探すことには有効であるが、信頼性を評価することにはあまり適していない。

これらの問題に対し、本研究では以下のような考えにより信頼性の評価を行う。まず、信頼性に対しどのような基準が可能であるかについて考える。我々が通常利用している信頼性の基準について考えると、誰が作成したページであるかなどの作成者に関する判断に基づいて信頼性を評価することはよくあることである。例えば、官公庁や有名な企業あるいは良く知っている人によって書かれた内容ならば一定の信用を置くことができるし、知らない人の書いた内容は一概には信用できないといった判断を行っている。このような信頼性の判断は WWW からではなく評価する人の社会的な知識に基づいて行われ、その結果がページの判断に反映されることが考えることができる。本手法では、このような我々が通常行っている信用についての評価、つまり、WWW 以外で決定される外部評価をページ信頼性の基準として用いる。

次に、2の問題に関しては、内容に依存しない評価方法が必要であると考えられる。このような方法の一つとして、内容を解析せずにページを評価する方法であるソーシャルフィルタリング的な手法が注目されている。代表的な手法として、WWW 構造であるリンク関係を解析する方法がある。現在、最も利用されている検索エンジンである Google では、「キーワード検索や構文解析」に加えて、リンク解析による評価もページ評価に利用されている。Google でのページ評価の中心となる PageRank 法では、「多くの良いページからリンクされているページは良いページである」という考えに則り、マルコフ過程を用いてリンクをたどる際のページの滞在確率を求め、各ページの評価として点数化する方法を採用している。この方法はページの人気度を評価することで信頼性の向上に貢献しているといえる。ただし、どの程度信用できるかという信頼性そのものを評価したものではないため、点数が高いページの方が必ずしも信頼が高いページでないという問題がある。

本研究で提案するページ評価手法では、WWW におけるグラフ構造を一種のネットワークフローの問題として捉え、初期値としてページに対する信頼度を指定し、ページ間のリンク構造の解析を行うことで、自動的に各ページに対し信用に基づいた信頼性の評価を得るというアプローチを採用する。

これらのアプローチにおける基本的な概念は、

- いくつかのページに対しては信用を評価することができる。
- リンク構造を利用することにより、他ページの信頼性を推測できる。

という前提に基づいたものである。

本提案手法の特徴として、以下の2点がある。

1. 信頼性の基準となるものに WWW 以外の外部により決定される「社会的信用」の評価を用いて、一部のページに対し評価を与える。すなわち、既知のページの社会

的な信頼性を外部評価として利用する。

2. リンク構造解析を利用して、それらのページの評価からすべてのページの評価を求める。すなわち、リンクによる信頼性が伝播するという考えに基づきページ評価を行う。

本研究で評価しようとしている内容の信用を表わすような信頼性は、前述したように本来は主観的なものであり、同じ情報に対する信頼性も個々の人間の主観によって異なる場合がある。したがって、信頼性の評価を行う場合に、誰にも共通する客観的な評価のみを初期値として採用するという考え方と、個々の主観を許しその人の主観に沿った評価結果を導出するという考え方のどちらも可能である。

本手法における外部評価として社会的知識に基づく信頼である「社会的信用」という概念を導入することで、誰が信用できるかという判断において、客観的評価の中にある程度の主観性を含んだ信頼性の評価を行うことができる。主観を含む評価を許すことにより、統一的な評価とともに多様な側面を持つ広義の信頼を評価することも可能にしている。

リンク構造解析においては、リンクの意味をどのように捉えて信頼性の判断に利用するかがページ評価において重要となる。多くの場合、我々がリンクを作成するのはリンク先ページの内容を評価した上で見る価値がある、あるいは見てほしいと判断した場合である。このようなリンクは一種の推薦を意味するものと考えることができる。リンクによる推薦を信頼という観点から見ると、リンクを作成した人はリンク先ページに対し推薦した責任を持つ、すなわち、推薦することで信頼を付与していると捉えることができる。これは、リンクを介してページへ信用が伝播していると考えることができる。本手法では、このような考えに基づきリンク構造の解析を行うことで各ページの評価を行う。

また、リンクの意味を解釈する際には、より正確にリンクの意味を捉えるためにリンクの解析において、次に2つの方法を提案する。

1. リンクに属性を付与することで、リンクの意味を明示させる方法
2. 属性が付与されない場合には、信頼性評価の初期値を複数持つことにより、それぞれの初期値に適したリンクの意味を考慮した方法。

以上に示したように社会的信用を基準とする信頼性とリンクによる信頼性の伝播を利用することでページ評価を行うことで、キーワードや内容に依存することなく様々な信頼に応じたページスコアリングやページフィルタリングなどへ応用を行う。

1.2 関連研究

WWW における信頼性評価

WWW における情報の信頼性に関する研究は、現状ではそれほど多くはなく、むしろ実用的に使えるシステムとして提供されているものが多い。一般の社会における信頼は、Luhmann [1] も述べているように一義に定義可能なものではないため、目的に応じて多少の主観的あるいは恣意的な判断が含まれることが多い。ここでは、社会的信用の概念に関連する直接信頼を評価するような定性的な評価手法について述べる。

実用的なシステムにおいては、目的に従った信頼に関する情報を利用することで、ページのレーティングやフィルタリングなどに利用している。このようなシステムでは、商取引における安全性の確保やいわゆる有害なコンテンツを排除することを目的としたものがほとんどである。したがって、既知の信頼情報を利用しているページを評価している点では、本研究における方法と同様であるが、それらを元にすべてのページの信頼性を評価し、一般的な内容の信頼性を判断するためには利用できないという問題がある。

電子商取引においては、信用のモデルに関する研究 [2][3] もなされているが、実用的にはベリサイン [4] などの信用を保証する会社により行われる認証が利用されている。これ

らの方法は、信用できる商店であることを示すことで取引における信頼性を確保し、安全な商取引を実現するために有効な手法である。

有害なページを排除するためのフィルタリングとしては、有名なポータルサイトである MSN [5] や Yahoo! [6] で導入される Internet Content Rating Association(ICRA) [7] によるレーティングが広く利用されている。マイクロソフトの Internet Explorer でも利用されている RSACi は、Recreational Software Advisory Council(RSAC)(現在は ICRA に吸収合併された)により作成されたレーティングシステムであるが、ICRA はこれを見直して作成されたレーティングシステムである。ICRA によるレーティングは、コンテンツ提供者側が World Wide Web Consortium(W3C) [8] によって規定された Platform for Internet Content Selection(PICS) に準拠したラベルを自身のページに対して記述するセルフレイティングを行うものである。セルフレイティングを行うことで、子供に対して不適切なページを見せないことと、表現や発言の自由を確保しようとしている。

ICRA のシステムでも利用されている PICS は、レーティングすること(格付けすること)とフィルタリングすることを分けているため、レーティングに基づいてフィルタリングを行うか否かは、フィルタリングする側に任されている。このようなシステムとすることで、PICS に基づいたデータベースがいくつか運用され利用できるようなになっている。国内でも財団法人インターネット協会 [9] によるシステムが運用されている。PICS におけるデータベースは本研究における WWW 以外の外部情報の一種であり、本研究における手法にも一部は利用できるものと考えられる。このような方法は、本研究の目的にそのまま利用できるものではないが、外部情報を利用することで Web ページの意味づけを行っていく点においては同様のものである。

WWW 以外の外部情報を用いる方法として、本研究に関連するものとしては、近年 Semantic Web[10] が注目されている [11]。W3C により規定された Semantic Web では、

Resource Description Framework(RDF) [12] を利用することで、ページに対してその意味を記述することができ、キーワードだけではなくその意味に基づいたページの処理を可能にしている。ただし、意味とその体系を定義するためにオントロジーが必要であり、的確なオントロジーの定義 [13] が Semantic Web の適用における課題となっている。Semantic Web におけるページに対する意味付けの手法は、本研究における外部情報として信用情報を指定する方法として利用することが可能であり、本研究の広範な実施における適用手段となる。また、実装手法として RDF のみを用いるという方法も考えられる。

ページに対する評価を直接指定するような定性的なページ評価手法は、的確な指定が可能であり、信用を判断するためには有効かつ確実な方法の一つである。しかしながら、数十億と存在するすべてのページに対してこのような判断を行うことは困難であり、多くのページを対象とするためには、解析的な方法を用いて計算により評価することができるページ評価手法が必要とされる。

解析的なページ評価手法

解析的な手法を用いてページを評価する方法としては、本研究においても採用している WWW をグラフとみなす考え方を含め、関連する多くのページ評価手法が存在している。これらのページ評価は、目的に対し適切な情報を提示するためにページの重要度や人気度を評価するためのスコアリング、ページの関連性や違いを示すためのクラスタリングやコミュニティ発見などを行うために利用されている。また、手法としては、キーワードやトピックに基づくもの、リンク構造解析に基づくもの、リンク構造とトピックを両方を用いるものなどが代表的である [14]。これらの方法は、本研究で用いているページ評価手法と高い関連性を持つ。特に、リンク関係によるページ評価の考え方は、ページの信用や信頼性を考える上で重要な考察をもたらすものである。

WWW 登場の初期においては、キーワードを中心としたページ評価やトピックによる分類が中心に行われてきた [15]。初期の WWW において AltaVista[16] などの検索エンジンにおいて行われたページスコアリングは、ページ内や META タグに含まれるテキスト抽出によるものであった。しかし、ページ数の増大とキーワードの羅列等による意図的なスコア操作のために適切な結果を得ることが難しくなっていった。キーワードを中心としたページ評価は、前に述べたように文章の内容のみにより信用を判断できないため、信頼性の判断には適していないと考えられる。また、初期の Yahoo![6] のようなトピックによる人為的な WWW ディレクトリの構築は現在も行われているが、すべてのページに及ぶような網羅的なものは困難となっている。

近年では、ページ間のリンクによる構成されるリンク構造を解析することによるページ評価手法や、ページ間の関係性を示すための Web コミュニティ発見などの手法が行われている [17]。

トピックに基づくリンク構造解析によるページ評価 [18] [19] [20] [21] [22] [23] [24] [25] としては、HITS アルゴリズム [26] [27] [28] が有名である。HITS において Kleinberg は、トピックに関する Hub ページと Authority ページを発見する手法を提案している。特定のトピックに関する多くの情報を持つ Authority ページと、それらの情報への多くのリンクを持つ Hub ページの発見手法は、リンク構造解析を用いて WWW 上の重要なページを見つけることができる。ただし、HITS ではトピックが一般的な意味を持つものにずれていくというトピックドリフトの問題が指摘されている [29]。

HITS における手法は、初期値となるページ群からリンク構造解析を用いてページを評価するという点において、本研究と近い手法である。ページの重要性の評価は、WWW 内における関係から判断される信頼性的一种として考えることもでき、Authority ページは多くの重要な HUB サイトから推薦されているという意味では、信頼度が高いと考えられ

る．また，HITS では解析における初期値はトピックに基づいて指定されるが，これらの初期値を信用に基づいたものを利用することで，信用に関する何らかの評価を得ることができる可能性もある．

HITS のような WWW における特徴的なページ群を発見する方法は，WWW コミュニティ発見としての研究 [30] [31] [25] [32], [33] [34] [35] [36] や似ているページの発見 [37] などが行われている．これらの手法は，WWW における特徴的なページを見つけることができるが，トピックに関する重要性や関係性を判断するために有効である．

トピックに依存しないリンク構造解析としては，WWW の構造に注目した方法 [38] [39] [40] [41] [42] [43] や社会的関係に注目した方法 [44] [45] [46]，確率的な手法による方法 [47] [48] [49] がある．特に Google [50] におけるスコアリング [51] に利用されている PageRank [52] [53] が有名である．PageRank アルゴリズムでは，多くの良いページからリンクされているページは良いページであるという考えに基づき，各ページでの滞在確率をマルコフ過程により求めることで，ページの人気度を評価している．PageRank は人気のあるページ，有名なページに高いスコアを与えるページスコアリング方法として効果的である．ただし，WWW 内のみの関係性によるものであり，HITS と異なり初期値となるようなページ群は必要としていない．

PageRank に対してトピックを考慮することでスコアリング精度を向上させる手法として，人為的に作成されたディレクトリ情報を利用した重み付けを行うもの [54] や，リンクのアンカーテキストをトピックにより重み付けすることでリンクにより他ページへの遷移確率を変える手法 [55] が提案されている．さらに，HITS や PageRank におけるページ評価の概念を拡張した手法 [56] [57] [58] [59] [60] も多数存在する．これらの手法は，人気度によるスコアに対し信用に基づいた情報を利用した重み付けを行っており，リンク構造解析に外部情報を利用した評価手法の一種とであり，信頼度の向上に有効であると考えら

れる。

解析的な手法では、HITS のように初期値が必要な場合や PageRank のように初期値無しで確率的に求めるなどの違いはあるが、計算により多くのページの評価を自動的に得ることができ、得られた評価によりスコアリングなどの定量的評価やクラスタリングなどの分類を行うことができる。これらは、膨大な Web データから必要なページや構造の抽出を行う一種のデータマイニングと考えられる。しかしながら、これらの手法では信用や信頼性を基準としたマイニングを目的としたものでないために、ある種の信頼性の向上に貢献しているものであるが、信用や信頼性の基準としてはそのまま使えないこともある。たとえば、HITS における authority ページはトピックに関する重要なページと示し、PageRank においては人気のある有名なページが高いスコアを得ることになるが、authority ページや高いスコアを得たページがスコアの低いページより信用が高いことを示すわけではない。

また、HITS や PageRank においては、リンク関係が多数存在するようなページが高い評価を得やすい傾向があるが、リンク関係が少ないが信用が高いページは、スコアの下位に埋もれてしまい易く、信頼性の判断を行うことが難しくなってしまうという問題も存在する。本解析手法においては、信用や信頼性を基準とした考えによりページ評価を行うことで、リンク数や人気などとは独立したページ信頼性の評価を行う。

1.3 本論文の構成

本論文の構成は以下の通りである。

第 2 章では、WWW 以外の外部情報としてページ作成者の社会的信用 [61] という概念を導入し、社会的信用に基づいた Web ページの評価を一般化モデルとしてグラフ理論を用いて数学的に定義し各ページの評価を導出するアルゴリズムを提案する [62] [61]。

本モデルでは、

1. いくつかのページに対しては信頼性が評価されている
2. リンク関係をリンク元ページからリンク先ページへの推薦と解釈する
3. 信頼度の高いページから推薦されたページの信頼度は高くなる．すなわち、リンクを介して信頼度が伝播する

という仮定の妥当性をしめす．

さらに、ページ作成者の URL と信用度を指定することと、グラフ理論を用いてページとリンク間の構造解析を行うことで、各ページの信頼度を計算する手法および WWW へ適用するためのアルゴリズムを提示する．本アルゴリズムはダイクストラ法を改良することにより、妥当な計算量によりすべてのページの評価を行うことができる．また、提案方法の汎用性に関する議論を行い、一般化モデルで使用するパラメータを調整することで、信頼性を表すいくつかの代表的な基準に適用できることを示す．

第 3 章においては、大規模な Web サイトにおけるサイト構造に注目し、Web サイト内のページの信頼性を評価するための手法について述べる [63] [64] [65] [62] ．

ここでは、大学や研究所などの組織が持つサイトにおいて多くの作成者が各々のページを作成する環境を想定し、信頼性の基準として公式度という概念を導入する．組織においてはページを作成する人や部署に応じて、情報を発信する際の責任や裁量が暗黙あるいは明確に規定されており、誰の責任でページを作成したかによってページの公式度が異なっている．また同一作成者によるページでも、公式ページや趣味のページなどで明らかに公式度が異なるページも存在している．

本章では、第 2 章で提案した一般化モデルにおいて、

1. 組織構造を利用して各ページの初期値を決定する

2. リンクに属性を付与することで、リンクの意味を明示させる

ことにより、サイト内の各ページの公式度を導出する方法を提案する。

さらに、実在のサイトを対象とした評価実験を行い、提案手法の妥当性と有効性を明確にしている。また、提案手法を利用することにより、現在の企業のような公式情報しか存在しないサイトにおいても自由度の高い情報発信が可能になることを示した。

第 4 章では、社会的信用に負の信用という概念を導入し、正と負の信用から WWW 全体における良いページと悪いページを判定する手法の提案する [66]。

ここでは、良いページと悪いページの持つ性格によりリンクの性質に違いがあることを利用し、リンクによるページ関係をモデル化し、良いページ/悪いページ/悪い可能性のあるページを判定する手法を導出した。

第 2 章で示した方法に対し、

1. 評価者が良いページ、悪いページと判断されるページ群を指定する。
2. すべてのページに対して推移的閉包を用いてリンクを解析することで、良いページ群/悪いページ群との到達可能性を調べる
3. 得られた到達可能性の関係によって、各ページの良い/悪い/gray(悪い可能性)の判定を行う

ことにより、各ページの信用を判定している。

さらに、実在のサイトに対して調査を行い本判定基準の妥当性であることと、ポータルサイト、検索サイト、ランキングサイトなどでは様々なページに網羅的にリンクしているために、本手法においては取り除いて解析するべきであることを明らかにした。また、WWW に適用する場合の有効性と問題点を示すことで、WWW におけるリンク関係についての考察を行う。

第 5 章では，本研究におけるページ信頼性の評価手法について総括を行う．

第 2 章

社会的信用に基づく Web ページ評価手法の構築

2.1 はじめに

利用者が Web ページを見てページの内容について判断する際には、誰がページを作成したのか、どのような立場で書いているのか、どのような目的で書かれているのかなど Web ページ（集合）自体からは導出できない Web とは独立した外部評価を行っている。つまり、利用者が Web を見るときに暗黙的に行っている外部評価は、Web を見る目的や環境に応じて恣意的な判断の下に行われていると考えられる。

Web ページの信頼性においても、最終的な判断はページの閲覧者が Web とは独立した知識に基づいて判断することは多い。誰が作成したページなのかを評価するためには、作成者（誰）がどのような人物あるいは組織なのかを判断できる社会的な知識が必要となる。作成者に関する信用が評価することができれば、その作成者のページを書いた際の立場からページの信頼性を判断することや、リンク先のページに対する信頼性の判断において適切な評価を行うことができる。

本章では、ページ作成者に関する社会的な信頼性を社会的信用と考え、社会的信用に基づき、リンク構造解析によりすべてのページに評価点を付け、ページの信頼性を判断する方法を提案する。

本方法では、Kleinberg[26]も指摘したように従来は推薦の意味でのみ捉えられることの多かったリンクに対して、推薦度を付与できるように拡張した。リンクに推薦度を付けることで、リンク構造解析を利用した評価に役立つだけでなく、ページ作成者のリンク設定に関する意図を明示的に表明でき、ある種の意味づけが可能となる。

本手法により、ページに対する責任や信用の存在を示すために利用することができ、検索手法などと組み合わせることで、インターネット上での信頼性を確保したコミュニケーションを実現できると考えている。すなわち、本手法を応用することにより、過去に問題となったようなリンクすることの責任に関する問題 [67] への適用や、情報のフィルタリングなどへも応用可能である。

本章では、まず社会的信用の考え方を示し、次に推薦度付きのリンク構造からページ間の相対評価を用いて行うページの評価の考え方について示す。さらに、ページ評価点として評価最大値を採用することでページ評価を行う手法を提案する。最後に本ページ評価点の利用法についても議論を行う。

2.2 社会的信用とページの信頼性

本手法におけるページ評価は、社会的信用と推薦度付きのリンクによりページの信頼性を評価する。本節では、WWWにおける社会的信用の考え方と、ページ間のリンク関係から得られるページの信用の考え方について述べる。

2.2.1 ページ信頼性の判断要因

一般的にページを判断するための要因としては、

- 誰が作成したか（企業，官公庁，学校，個人など）
- どのような目的で作成したか（広報，広告，趣味など）
- どのような立場で作成したか（組織として，公人として，個人としてなど）
- どこに作成したか（企業のサイト，官公庁サイト，プロバイダなど）
- いつ作成されたか
- どのような内容か

などの要因がある。

我々が Web ページの内容を判断するためには、最終的には内容を読むことを行っているが、同時に上記のような要因についても暗黙的に判断を行っている。

ページにおける信頼性を判断するためには、

- どこに作成したかを URL を見ることにより企業や官公庁，個人などの区別
- 誰が作成したのか，作成者の社会的責任の度合い
官公庁や有名な企業などの社会的責任の高い作成者
- 個人などの責任の低い作成者，知らない企業や NPO など責任が不明なものなど
ページが書かれた状況（目的や立場）により，内容についての責任の所在

などの要因の影響が大きい。

逆に，いつ作成されたかなどの要因は，情報の有用性には影響するが信用に対する影響は少ない。

すなわち、情報の信頼性に判断する要因としては、誰がどのような立場で作成したかという作成者の信用が重要であると考えられる。

このようなページに信頼性の元となるページ作成者に関する信用をページ作成者の社会的信用 (**Social Credibility**) と呼ぶ。ある作成者がどの程度信用できるかの評価である社会的信用は、現実の存在している社会的な状況によって決定されるものであり、Web から得られるものではない。すなわち、Web と独立した外部評価として得られるものである。

2.2.2 Web における社会的信用

次の例に示すように同種の情報を持つ複数のページが存在する場合のページの信頼性評価の例を考える。

例 2.1 それぞれ以下の作成者によって作成された、あるコンピュータシステムについてのマニュアルが存在するとする。

- コンピュータシステムを管理しているコンピュータセンターが作成
- 講義に使うために教員が作成
- 学生が個人的に使うために作成

もしページの閲覧者がページ作成者が誰かを知っているならば、最初にコンピュータセンターが作成したマニュアルを選択するはずである。なぜならば、コンピュータセンターがそのシステムについて最も責任を持つべきであり、利用者の要求に答えられるように内容を正確に保っていると考えられるからである。したがって、コンピュータセンターが最も信用できる作成者であると判断できる。逆に、学生の作成したマニュアルは最も信頼性が低いと判断できる。

すなわち、各マニュアルの信頼性は、

コンピュータセンター > 教員 > 学生

の順番であると考えられる。この順番は、すなわち作成者の社会的責任に基づいた信用に由来するものである。

このようにページの信頼性は、現実の社会において誰がシステムを管理しているのか、責任を持っているのか、といった知識に基づいて判断された作成者の社会的信用によって評価することができる。すなわち、高い社会的信用を持つページ作成者は、同時にそのページ内容について高い責任を負っているとこととなる。

社会的信用を決定する基準は、一意に決定できるものではなく目的や状況によって異なる可能性がある。誰が信用できるか誰が正しいかなどの判断は、誰がどのような立場で社会的信用を決めるかに応じてある程度の裁量が許されると考えられる。

2.2.3 リンクによるページ評価

リンク関係に基づいたページ評価手法として代表的なものとして、リンクによって内容の関係性を示すと考えるもの [68] [23] とリンクによりページの推薦を行うと考えるもの [52] [26] が存在する。これらに共通するのは、Kleinberg [26] も指摘したようにハイパーリンクは潜在的な人間による判断を含むものであり、**authority** を付与するものであるという考え方である。すなわち、ページ作成者がリンクを行うことにより、リンク先ページに対して信任を行うこととなり、リンクによる責任を持った推薦として解釈することができる。

ページの信頼性の観点からみると、E-Commerce における研究でも言及 [69] されているようにリンクをすることにより信用が伝播する、すなわちページ作成者の社会的信用をリンクを介してリンク先ページに対して付与することになる。

しかしながら、同じく Kleinberg [26] が指摘するようにリンクは信任や推薦のみの意図

により作成されるものではなく、様々な意図を持っている。単純な参照や「戻る」、「トップ」などのナビゲーションを行うためのリンク、章ごとに分かれた文書などの構造を表すためのリンクなど、推薦を意図しないリンクも多数存在する。また、自分の作成したページと他人の作成したページに対するリンクでは、リンクの持つ意味が異なっている。従来これらのリンクの意味は抽出することが困難であり、またリンクの意図を表明する方法がなかった。そのためリンク構造解析によるページ評価においては、アンカーテキストなどによる重み付けは行うことはあっても、推薦度の判定についてはリンクの有無によって行うのが一般的であった。

本手法では、リンク作成時に推薦度を付与できる推薦度付きリンクを導入することで、ページ作成者がリンク先ページに対してどの程度好意的に評価しているか、あるいは構造だけを示し推薦を含まないリンクなのかといった、リンクの持つ意図を明示的に表明できるようになる。同時に推薦度を利用することで、リンク先ページに対してどの程度の信用を付与するかを指定することができ、責任を明確にしたリンクを作成できる。したがって、社会的信用に基づくページ信頼性評価も推薦度付きリンクを用いることで、より正確な評価が可能となる。

2.3 ページ評価手法

本方式では、ページによるページの評価という考えによりリンクによる評価を行う。つまり、リンクを介したページ間の相対評価という考え方を用いている。また、ページの信頼性の元となる社会的信用を数値を用いて示すことで、各ページの信頼性を評価点として得る。

ここで、リンクにより構成される Web 構造を、ページをノード集合 N 、リンクを枝集合 E とする有向グラフ G とし、以下のような表記を行う。

$$G = (N, E)$$

$N = n_i$ | ページ集合中のページ

$E = e_{ij}$ | ページ集合中の n_i から n_j へのリンク

q_i : n_i の社会的信用の値

w_{ij} : リンク e_{ij} に付与された推薦度

(ただし, $0 \leq w_{ij} \leq 1$)

w_{ij} の値については,

- 自ページの評価点以上を他ページの評価点として与えることは不自然であること
- 推薦しないことを負の推薦度で表現することも考えられるが, 本章では負の評価点を考慮しないこと

を意図している.

図 2.1 にページ n_a を起点とし, ページ n_i にいたる経路 x 上の推薦度を用いたページ評価の例を示す. 図 2.1(a) は初期状態である. ページ n_a の社会的信用の値 q_a とページ n_a からページ n_b へのリンクに付与された推薦度 w_{ab} , ページ n_b からページ n_i のリンクに付与された推薦度 w_{bi} が与えられている. また, ページの社会的信用の値 q_b, q_i が 0 点となっているのは, 予め与えられていないため便宜上 0 点としている.

図 2.1(b) は, リンクによる評価が行われた結果である. ページ n_a を起点として, q_a を 100 点と評価すると, ページ n_b の評価点は, 推薦度が 0.8 であるため 100×0.8 つまり 80 点となる. さらに, ページ n_i の評価点はページ n_i への推薦度が 0.5 であるため 80×0.5 つまり 40 点となる.

ここで, $s_j(n_j, x)$: ページ n_i からページ n_j へのある経路 x による評価点とし, 各ページに社会的信用の値が与えられているとすると,

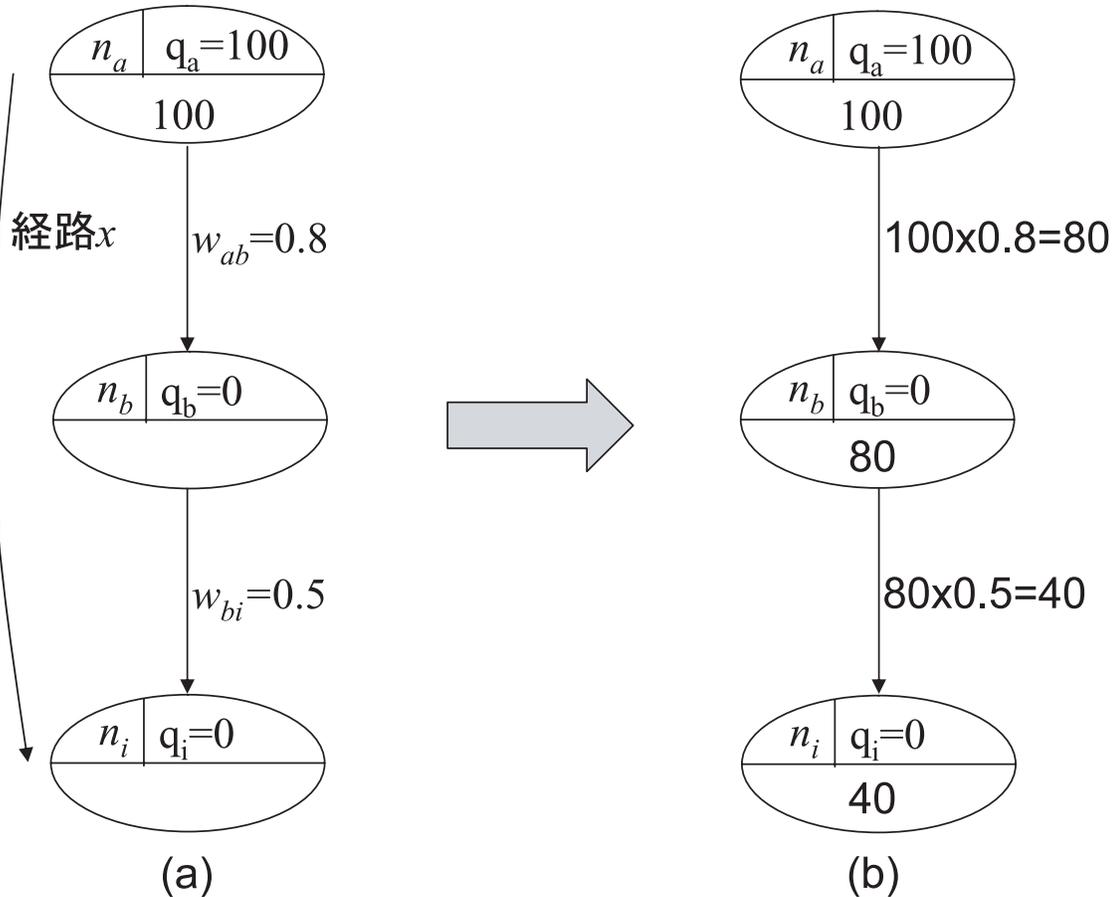


図 2.1 リンク関係を用いたページ評価例

$$s_i(n_a, x) = q_a \cdot w_{ab} \cdot w_{bi} = 100 \times 0.8 \times 0.5 = 40 \quad (2.1)$$

となる。

図 2.2 では、ページ n_i に対して複数の被リンクが存在する場合を考える。この図ではページ n_i に対して、 $n_a \rightarrow n_b \rightarrow n_i$ を通る経路 x と $n_c \rightarrow n_d \rightarrow n_i$ を通る経路 y が存在する。

ページ n_a を起点とした場合は、図 2.1 と同様にページ n_i の評価点は 40 となる、しかしページ n_c を起点とした場合には、ページ n_i の評価点は 24 となる。つまり起点となる

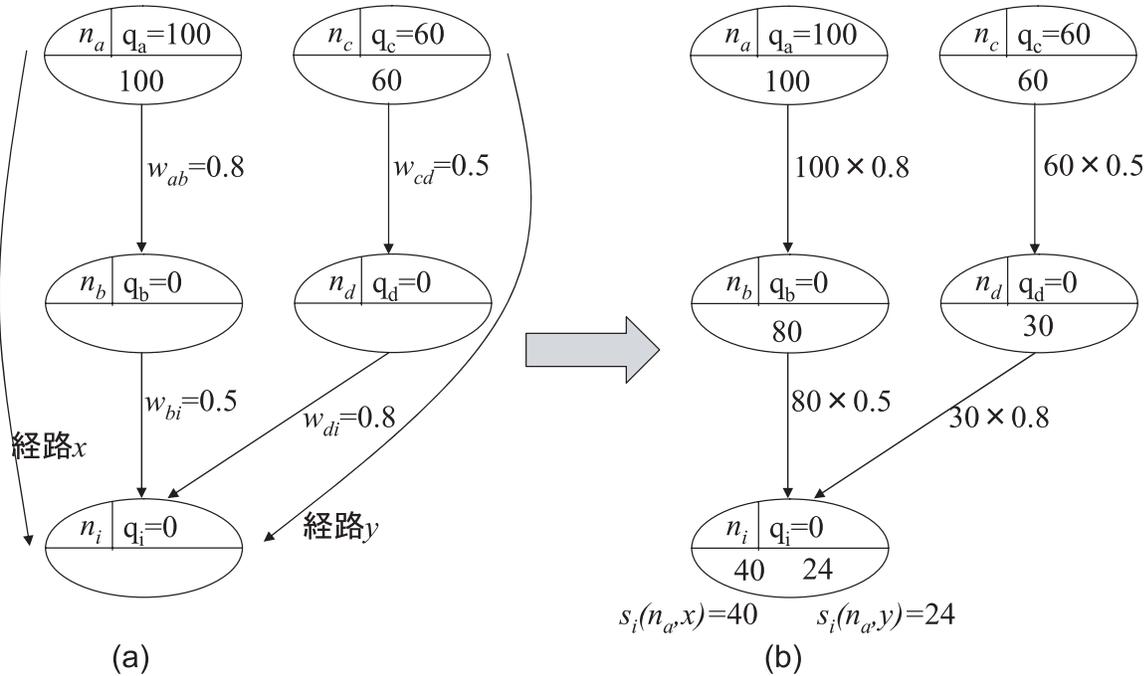


図 2.2 複数リンクがある場合の評価例

ページの違いによって、ページ n_i の評価点は異なることになる。

すなわち、ページ n_i の評価点は一意には決まらず、

$$\begin{aligned}
 s_i(n_a, x) &= q_a \cdot w_{ab} \cdot w_{bi} = 100 \times 0.8 \times 0.5 = 40 \\
 s_i(n_c, y) &= q_c \cdot w_{cd} \cdot w_{di} = 60 \times 0.5 \times 0.8 = 24
 \end{aligned}
 \tag{2.2}$$

となる。

さらに図 2.3 では、ページ n_a からページ n_i に対して 2 つの経路が存在する。この場合には、経路上のリンク推薦度の違いによって、ページ n_i の評価点が異なることになる。

すなわち、経路 $n_a \rightarrow n_e \rightarrow n_i$ を x' すると

$$s_i(n_a, x') = q_a \cdot w_{ae} \cdot w_{ei} = 100 \times 0.6 \times 0.8 = 48
 \tag{2.3}$$

となる。

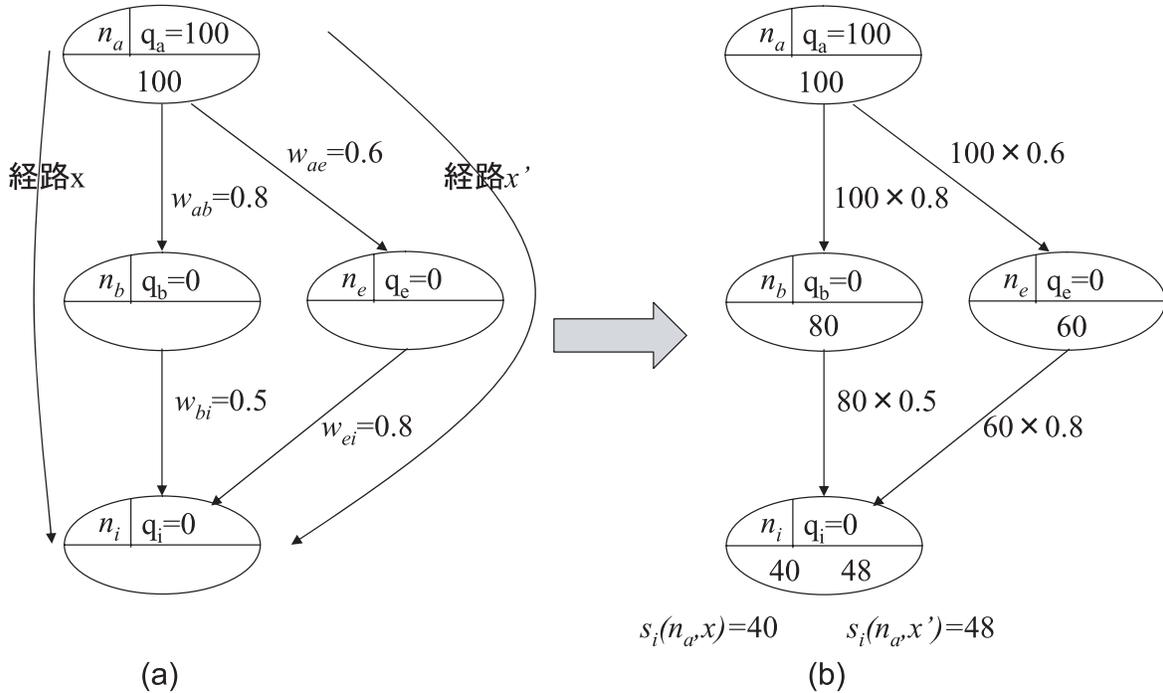


図 2.3 経路が 2 つある場合の評価例

以上のように，あるページの評価点は，任意のページを起点としてどのような経路を辿って評価されたかによって決まる．すなわち，あるページに注目したとき，その評価点は起点となるページと経由するリンクにより異なることになり，あるページを起点とした，そのページの評価として非常に多くの評価点が存在する．

我々が Web ページを見てリンクを辿るときを考えると，リンク先ページがどのようなページにリンクされているかは既知ではなく，自分の辿っているリンクによってのみ判断していると考えれば，一つのページに起点や経路によって複数の評価が存在するのは妥当といえる．多数の評価点からページ評価点を決定する方法としては，評価の個数や偏差を計算するなど，幾つかの方法が考えられる．

我々は，評価点の決定について，最大評価点の採用を方針とした．すなわち，ページ n_i に対する評価点を ps_i とし，集合 Z_i をあるページ n_i への評価点の集合とすると，

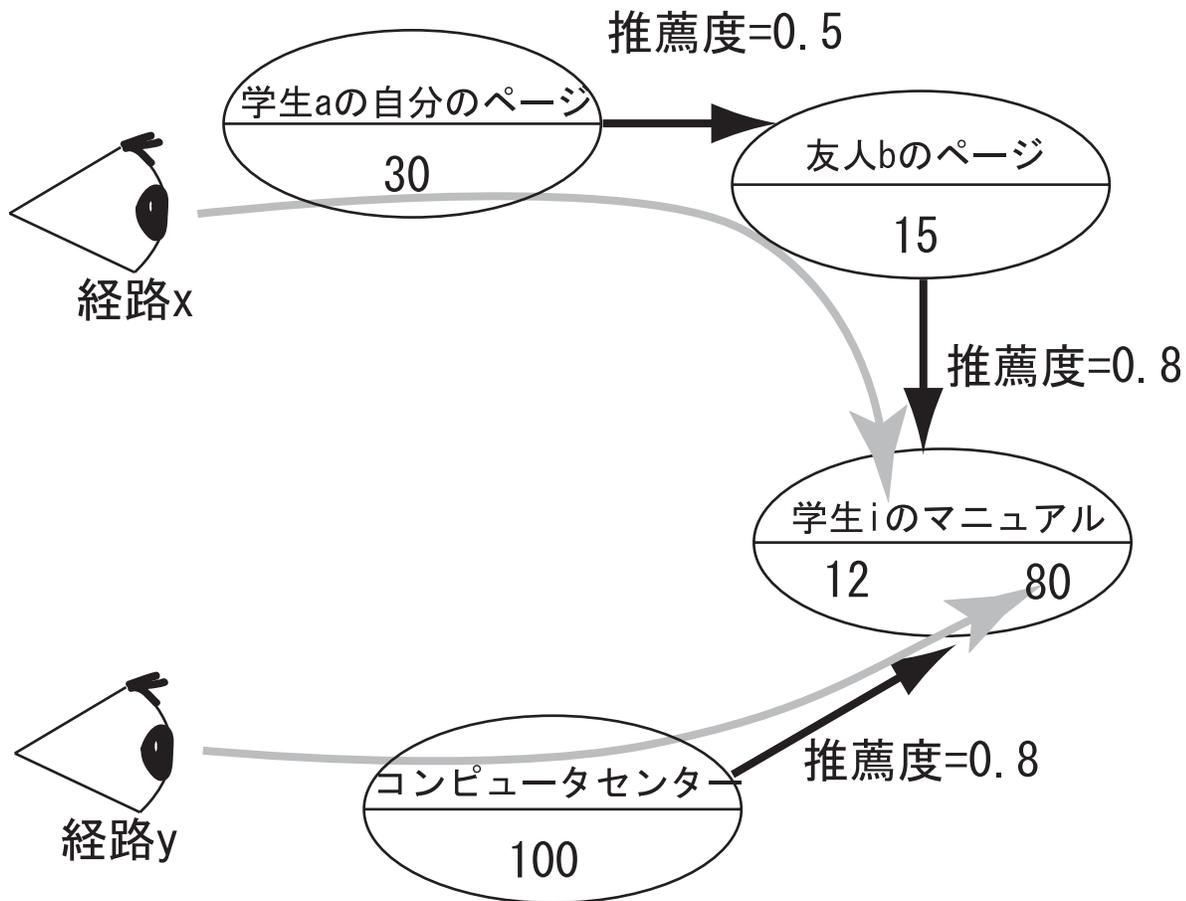


図 2.4 評価点の採用の判断

定義 2.1 ページ n_i の評価点は

$$ps_i = \max(q_i, \{z | z \in Z_i\}) \text{ ただし, } \{z | z \in Z_i\} \text{ は } Z_i \text{ に属するすべての要素を表す (2.4)}$$

と表される。

評価点として最大値を採用することは、少なくともそれだけの評価がなされているという評価のポテンシャルを示していると考えられる。起点となるページからの何らかの積極的な評価が、リンクにより伝播しつつ、あるページの評価点が決定されることから、少なくともある程度の推薦という評価を行ったページ群が存在している。つまり、ある視点からの評価において、良いページである可能性を持っていると考えられる。

例えば、図 2.4 に示すように、大学において学生 a があるシステムの使い方を探してい

るとき、経路 x の場合、自分のページを起点として友人 b のページのお勧めリンクを経由し学生 i のページに到達し、評価点は 12 となる。ところが、学生 i 書いたマニュアルが良くできていて、システムを管理しているコンピュータセンターからお勧めリンクが張ってあるとする。このとき、経路 y が存在し、学生 i のページの評価点は 80 となる。

この例では、学生 i のマニュアルページに対する評価点は、コンピュータセンターのお勧めを信頼して 80 点とするのが妥当と思われる。少なくともコンピュータセンターは、学生 i のページを良いページだと判断している。経路 x を経由して学生 i のページにたどり着いた学生 a は、そのページを 12 点相当にしか判断できない。このとき、そのページに 80 点が付けられていることやコンピュータセンターからお勧めリンクが張られていることを知れば、ページの判断において利用できる。

このように、評価点として最大評価点の採用することは、なんらかの良い評価がされているポテンシャルを示すものとして、ページ評価における要素として有効であると考えられる。

2.3.1 ページ評価の伝播

本方式の特徴の一つとして、評価点として最大評価点を採用することで、各ページがどのページを起点とするリンクによって評価が伝播したかが一意に決定されることがある。各ページがどのページの社会的信用の値を起点として評価が伝播したか決まることで、Web の持つグラフ構造から、評価伝播の木構造を抽出することができる。Web 構造に存在しているサイクル構造も、推薦度が $0 \leq w_{ij} \leq 1$ であるため、サイクルしても評価点は減少するだけで評価の伝播がサイクルすることはない。

図 2.5 にリンクによる評価点の決定と、その評価の伝播の例を示す。

図 2.5 において社会的信用の値がないページは、信用の値が評価がされていないか 0 点

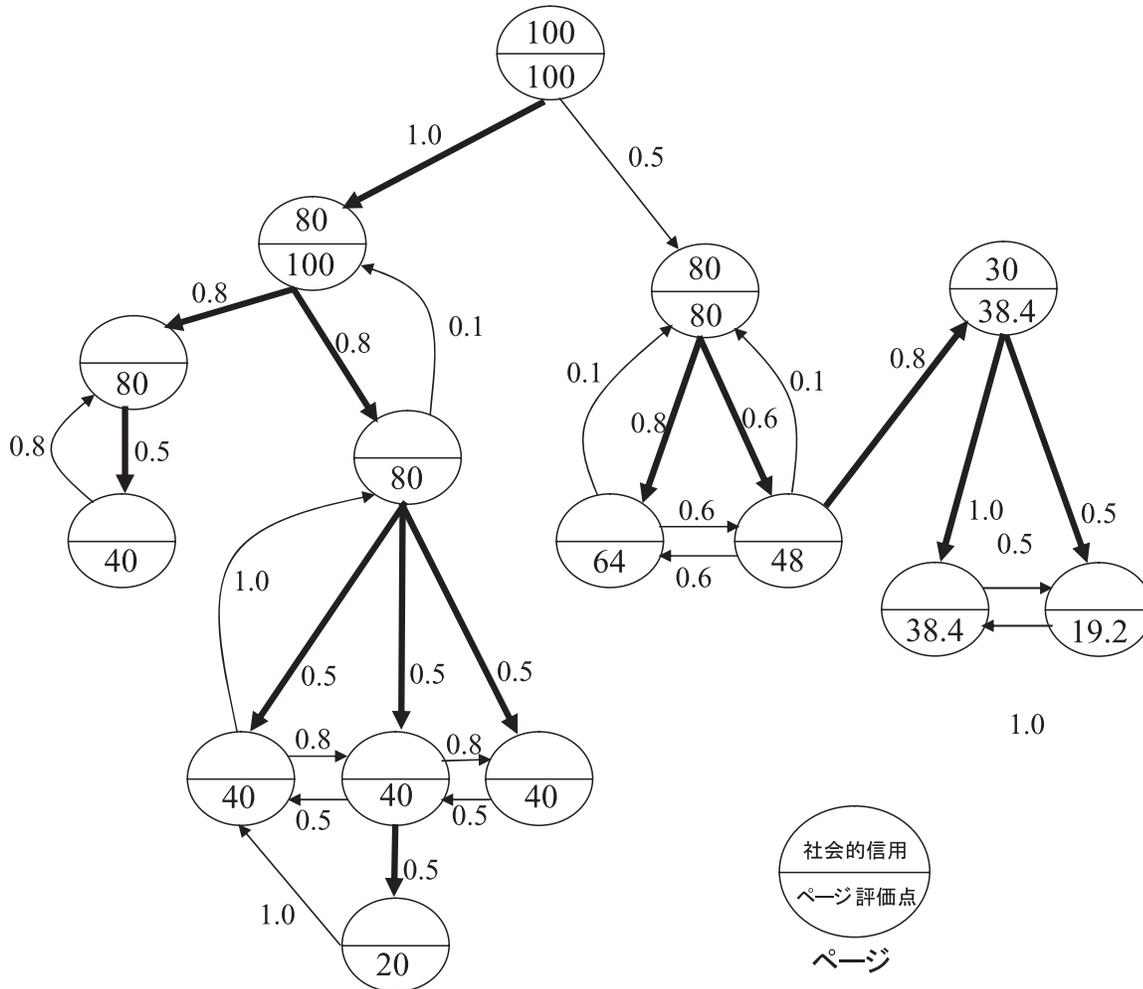


図 2.5 評価の伝播木

であるかどうかである。

本来の Web のグラフ構造から，評価点に関するリンクだけを抜き出し（太線），それ以外のリンクを消す（細線）ことにより，どのページを起点として評価が伝播したかが木構造となって現れている．この木構造により，あるページがどのページを起点として推薦されているか，誰が推薦しているのかが明確に示される．このような Web におけるページ間の関係，あるいはページ作者間の関係は，評価点の基準となる社会的信用の視点によって多少異なる解釈が許すものである．また，自分の作成したページに対して評価の

伝播を確認することで、Web 読者の視点からどのページを推薦したかになっているか、誰に推薦してもらっているかを確認できる。Web における自ページの評価関係の確認は、ページの自己管理にも有効と考えられる。

2.3.2 リンクによる推薦の解釈

ここで、本手法におけるリンクによる推薦が意味するものについて考える。

ページ n_a が低い推薦度によってページ n_b をリンクしているとする。このとき、「ページ n_a はページ n_b が低い信用しか持っていないと判断している」と考えるより、「ページ n_a はページ n_b が高い信用を持っているか判断できない」と考える方が妥当であると思われる。つまり、低い推薦度でリンクを作成する場合、内容を良く知っているページを敢えて（批判の意味を込めて）リンクするためより、リンク先ページを十分に知らないときや信用について正確な判断ができないことが理由で作成されることが多いと考えられる。

例 2.2 コンピュータセンターが学生が書いたマニュアルの存在を知ったときに、その中身を十分に読まずにリンクを作成したとする。このとき、コンピュータセンターは学生のマニュアルについて責任を負うつもりはなく、単にリンクは学生の書いたマニュアルの存在を紹介しているだけである。したがって、リンクのもつ推薦度は低い。低い推薦度のリンクとすることで、リンク先の学生のマニュアルに対して最低限の責任しか負っていないことを意味している。

この例に示すようにリンクによる推薦は、リンク先ページに対して負うことができる最低限の責任を示すであると考えることができる。すなわち、任意の $s_i(n, x)$ に対するページの評価点 ps_i について、以下の式が成り立つ。

$$ps_i \geq s_i(n, x). \quad (2.5)$$

すなわち， V_i をノード n_i に到達可能なすべてのノードの集合， X_i を V_i に属するノードを基点としノード n_i に到達するすべての経路の集合とすると，

$$ps_i \geq \max(s_i(n, x) | n \in V_i, x \in X_i). \quad (2.6)$$

したがって，本手法においてはノード i の最終的な評価点として最大値を採用する方針とする．

2.4 ページスコア計算方法

前節で示したページ評価モデルより，Web に対して適用を行うための計算方法を示す．

2.4.1 評価点の導出

ノード n_i への評価点の集合 Z_i を求めるには， X_j に属するすべての経路に対する評価点（すなわち，起点となるノード $n_i \in V_j$ の社会的信用の値 q_i に対して経路上にあるすべてのリンク e_{ab} 上の推薦度 w_{ab} を積算した値）の中から最大値となる評価点を見つけなければならない．経路集合 X_j には，サイクルが含まれていることもあり，その要素数は相当数に上る可能性があり，2.4 式を直接計算することは困難である．そこで本節では，2.4 式を変形しページ n_j の評価点 ps_j を求めるためのアルゴリズムを導出する．

今，ノード n_a がノード n_j に対するリンクを持つ，つまり e_{aj} が存在するものとする． X_j の要素のうち，経路に枝 e_{aj} を含む経路の集合を Y_j^a とする．

Y_j^a のうちノード n_j に対して最も高い評価を行う（起点ノード，経路）の対は， n_a に対して最も高い評価を行う対と同じものである．なぜなら，各対の n_j に対する評価は，各対の n_a に対する評価に w_{aj} を積算したものであるからである．ここで，この対には， n_a 自身を起点とする対も含まれる．すなわち，

$$\max(s_j(A, x) | x \in Y_j^a) = ps_a \cdot w_{aj} (\text{ただし, } A \text{ は } V_a \text{ に属するすべてのノード}) \quad (2.7)$$

と表すことができる．ここで, N'_j をノード n_j に対して直接リンクしているノードの集合とすると, 2.4 式は, 2.7 式を用いることにより, 隣接ノードの評価点 ps_a とそのノードからの推薦度 w_{aj} のみの関係式で表すことができ,

$$ps_j = \max(q_j, \{ps_a \cdot w_{aj} | n_a \in N'_j\}) \quad (2.8)$$

となる．この式を解けば, 各ノードの評価点 ps_i を求めることができる．

ここで, すべての w_{ij} において, $0 \leq w_{ij} \leq 1$ であったことに注目すれば, 次の性質が成り立つ．

性質1 すべてのノードの中で最大の社会的信用の値 q を持つノードを n_i とすれば,

$$ps_i = q_i \text{ となる.}$$

性質2 すべてのノードに対する評価点 ps_i のうち, $1 \sim k$ 番目までに大きい値を持つ

ノードの ps_i が求まったものとする．このとき, $k+1$ 番目に大きな値を持つノードは, 既に求まっている $1 \sim k$ 番目のノードのいずれかから直接リンクされているか, あるいは $1 \sim k$ 番目以外のノードの中で最大の社会的信用の値 q_j を持つノードである．

2.4.2 アルゴリズム

以上から, 各ページの評価点計算のためのアルゴリズムを求めた．本アルゴリズムでは, 各ページに対し社会的信用を利用してページの初期値を設定し, まず最大値を持つノードを決定し, さらに次に大きい値を持つノードを決定するという繰り返しにより, 高い評価点を持つノードから順番に評価点が決定していく．

また、このアルゴリズムによりすべての評価点は、ページ数を n 、リンク数を m としたとき、 $O(n \log(n + m))$ で計算可能となる。以下にアルゴリズムを示す。

すべての ps_j は以下のアルゴリズムによって得ることができる。ここでは、 pc_j を評価の現在値を入れるもの、 $label_j$ はノード n_j の現在の状態、 $cnode$ は現在注目するノードをそれぞれ示すものであるとする。

[手順 1:] [初期化]

すべてのノードに対して、 $pc_j = q_j, label_j = 0$ とする。ここで、 $i = 0, cnode = 0$ とし、手順 2 に進む。

[手順 2:] [最大の評価点を持つページへの注釈]

$label$ の値が 1 でないすべてのノード n_k に対して、最大の pc_k を持つノード n_k を探す。[同値の場合、任意のノードを選択する]

$cnode = k, ps_k = pc_k, label_k = 1, i = i + 1$ と置く。

$i > n$ ならば手順 4 に進み、それ以外なら手順 3 に進む。

[手順 3:] [注目ノードからの評価点を計算する]

$cnode$ からのリンクを持ち、 $label$ の値が 1 でないすべてのノードに対して、

$pc_j = \text{Max}(pc_j, pc_{cnode} \times w_{cnode_j})$ とする。手順 2 に進む。

[手順 4:] 終了 [すべての ps_k の値が計算された]

2.5 適用可能性と議論

本提案方式の利用の可能性と展望について議論を行う。

提案方法では、社会的信用をページ評価者に対してある程度の恣意的な視点から決定を許容する外部評価として利用している。しかし、より広く公平にページの信頼性を評価す

るためには、格付け機関のような組織による一般的な社会的信用の決定が必要となる。単一の格付け機関ではなくても、分野ごとに信用できる機関による社会的信用の決定が行われれば、正確なページ信頼性を実現できる。社会的信用を決定するための情報として、Yahoo! [6] のディレクトリや DMOZ のような公開されているオープンディレクトリ [70] などを利用することで、信頼性を高める方法も考えられる。

一般的に判断される社会的信用だけではなく、個人的な興味や趣味に基づいて主観的評価を含む社会的信用に基づきページ評価を行うこともできる。主観的な評価を行うことで面白いと思われるページや、自分に適したページを発見するなどの応用も考えられる。ただし、社会的信用についてはある程度の主観的な視点にて決定できるが、リンクによる評価は推薦度を利用したため推薦という視点のみで評価されている。したがって、社会的信用と推薦の組み合わせによっては、評価点が正確に社会的信用の値を反映するものとは言えない場合もあり得る。しかし、社会的信用の視点として、何を以てするかで全く異なる意味を持つ評価点を得ることで、何らかのページ評価基準として利用できるだろう。

本手法はページの内容に全く依存しないページ評価手法であるため、検索エンジンとの組み合わせも有効である。キーワード検索の結果として得られた URL に対して、社会的信用の値を指定することで本手法におけるページ評価値を適用すると、興味ある話題についての関連ページ群を木構造として、引き出すこともできる可能性がある。

Web の構造を可視化することにより、その構造や関係性を表現する手法も数多く行われている [71] [72] [73]。本手法の応用として Web を閲覧する際に利用する場合に、ページ評価の伝播木を可視化し、利用者へ提示することが考えられる。伝播木を参照しながらページを閲覧することにより、読んでいるページが誰によって推薦されているかを確認することができ、より安心して Web 閲覧を行うことができる。評価の伝播木を可視化することでページ作成者と推薦の関係を概観することができ、閲覧者の情報判断の手段として

利用可能である。

2.6 まとめ

本章では、社会的信用と推薦度を記述できるように拡張したリンクを用いた、新しいページ評価手法の提案を行った。本方法では、Web データから導出できない外部評価による社会的信用と、ページ作成者の意向を反映するリンク推薦度を導入することで、ページの信頼性を判断するための新しい評価点を導入した。

第 3 章

Web サイトの特徴を利用した評価 手法のサイトへの適用

3.1 はじめに

WWW における情報の信頼性を考える上で、情報の発信元である Web サイトにおける信頼性の確保は重要な要因である。本章では、サイト内ページの信頼性を評価し、提示する手法について述べる。

情報を発信する側として見た Web サイトは、WWW における初期の頃と比較し、大きく変化している。初期のサイトでは、サイトとして公式の情報と個人で作成されたページが混在する自由な情報発信が行われていた。しかし、現在のサイト、特に企業のサイトはプレスリリース、製品情報、求人情報、企業情報などの公式情報しか載せていない。一方、大学や研究所や NPO などのサイトでは、従来のような公的な情報と個人ページが混在している。

後者のような公的なページと個人ページが混在した環境では、異なる社会的信用を持ったページ作成者が混在している。このようなサイトに対し第 2 章において提案したページ

評価手法を適用することにより、ページの評価点をサイトによる一貫した公式性の度合いや信頼度として利用することが可能となる。また、サイト内で各ページ作成者が自己ページにリンク推薦度を付与することによりページの自己管理を行うことで、ページ信頼性の自己管理を実現できる。すなわち、サイトに対して評価手法を適用することにより、サイトと各ページ作成者の意図を反映したページ信頼性を明示した情報発信を行うことができる。

本章における研究では、第2章のページ評価手法をサイトの特徴を生かし効果的に適用を行うための適用手法を示す。本適用手法では、

- 組織構造を利用した社会的信用の定義ファイルの利用
- リンク推薦度のリンク属性による指定

を行うことにより、サイトにおけるページ評価を行う。さらに、実際のサイトに適用実験を行うことで本手法の妥当性を検証する。

本評価手法は、大学などサイトのページの信頼性を提示するだけでなく、企業などのサイトに対して適用することで、かつてのような自由な情報発信をページに対する責任を管理した上で行うために有効である。

3.2 サイトにおける信頼性

3.2.1 サイトの種別

現在多くのサイトが存在しているが、情報の管理という視点からみると以下の2つのタイプに分類できる。なお、本章におけるページ作成者とは、実際にページを作成した人とは限らず、外部委託などにより実際に作成はしていない場合でも、内容を決定し、かつ責任を持つ人(あるいは部署など)を表すものとする。

単一型 サイト内のすべてのコンテンツがサイトを主催する組織の公的な情報のみで構成され、単一のページ作成者によって作成されたサイト。

企業のサイトの多くはこのようなタイプである。これらのサイトでは、部署や個人による個別コンテンツは存在が許されていない。

分散型 サイトを主催する組織による公的な情報と、部署や個人などの複数のページ作成者による分散管理されているサイト。

大学、研究所、NPO/NGO などにこのタイプが多い。これらのサイトでは、各ページ作成者が自分の作成したページを自分の責任において管理している。例えば、大学の場合では、サイトを主催する組織の公式情報は広報委員会、学部や研究科の入試情報はそれぞれの入試委員会、教員や学生の個人ページは各個人など、組織や部署や個人の責任において作成され管理されているページが混在している。

単一型のサイトでは、すべてのページの社会的信用は均一であり、すべてのページの責任はサイトを主催する組織に存在する。

分散型のサイトでは、作成者ごとに社会的信用は異なり、同様にページの信頼性も作成者に依りて異なる。さらに、同じページ作成者によるページであっても、作成者の立場によりページの信頼性も異なる。例えば、同じ教員が書いたページであって研究業績などの公式な情報と趣味に関するページでは、同じ責任を持って作成したとはいえずページの信頼性も異なる。このような様々な社会的信用を持つ作成者が存在する環境は、第2章におけるページ評価手法における環境と同様のものである。本手法を適用することで、分散型サイトにおけるすべてのページの信頼性を評価することができる。

3.2.2 ページの信頼性と責任

Web 全体の場合と異なり，サイト内での社会的信用の定義においては，サイトとしての信頼性を確保した上で管理することができるために，統一した基準により信用を決定することが可能である．サイト内の作成者である部署や個人は，組織内における地位や立場に応じて責任を持っている．すなわち，組織におけるページ作成者の社会的信用は，組織によって明確に決定できるものである．例えば，大学であれば広報委員会は大学として全責任を負った情報を発信しているが，学生のページに対してはある程度の管理責任のみ負っているが，内容に対して責任を追うものではない．

このような組織として責任の度合いによりページを評価することで，ページに対する組織としての保証の度合いや，公式度として解釈できるページ信頼性の担保とすることができる．各ページ作成者は，組織（サイト）により与えられた責任を自分の作成したページに対して自己責任において付与することで，サイト全体としての責任を明確にできる．このようなページの自己管理はリンク推薦度が利用できる．自己作成ページに対するリンクでは，リンク推薦度をリンク先ページに対してどの程度の責任を持っているか，あるいはどの程度公式なものかを示すために利用することができる．ただし，自己作成ページに対するリンクには，ナビゲーションや文章の章分割などの構造を示すための推薦を含まないものが多数存在している．したがって，自己ページに対するリンクと他者ページへのリンクにおける性質の違いについては考慮する必要がある．

3.3 適用手法

本適用手法では，以下の方法によりサイト内の社会的信用とリンク推薦度の指定を行う．

3.3.1 社会的信用の記述方法

サイトにおける適用において、社会的信用の値としてサイトを主催する組織としてのページ作成者に対しどの程度責任を持つかを表す値あるいは組織としての公式度を表す値を採用した。以下、本章における社会的信用の値を公式度と呼ぶ。この公式度を数値として示すため、責任を持たないあるいは関知しない場合を公式度 0、全責任を持つ最大値を公式度 100 とした。

ページ作成者を特定するためには、サイトにおける URL を利用する。しかしながら、多くのページ作成者が存在する大学などの大規模サイトにおいて、サイト内のすべての作成者の URL を指定することは膨大な量となり実質困難である。そこで、本手法ではページ作成者の種別毎に公式度を指定することにより指定の簡素化を図った。サイト内のページ管理は、部署や個人などの作成者毎にディレクトリ単位で行われることが多い [74]。したがって、ディレクトリを示す URL に対して公式度を指定することで、部署全体に対して値の指定を行うことができる。例えば、広報委員会の管理するディレクトリ、プロジェクトのディレクトリ、教員用のディレクトリなどの単位での公式度が指定できる。ディレクトリを指定することで、公式度はディレクトリ全体（ディレクトリに含まれるページ群）、あるいはディレクトリを代表するインデックスページ（通常は index.html など）に対して与えられる。また、特定の URL を指定することで単一のページに対して公式度を与えることもできる。

本章における実験では、表 3.1 のような種別と公式度を設定した。この作成者の種別、対応する URL を記述した定義ファイルを用意することで、サイト全体における公式度の管理を行う。作成者の種別ごとに公式度を管理することで、作成者の移動（教員の新任や退官、学生の入学や卒業など）などが発生してもほとんど変更を行う必要がないという

表 3.1 作成者の種別と公式度

種別	内容	公式度
faculty	学部当局	100
division	学科	60
relevant	関連研究	20
committee	学部内委員会	80
teacher	教員	50
student	学生	30
admin	サイト管理部署	70
project	学部内プロジェクト	80

利点も得られる。なお、実際の運営においては、この定義ファイルはサイトを主催する組織の責任で公式度の指定がなされ、記述はサイト管理者などが行うことになることを想定している。

実際の定義ファイルでは、ページ作成者の種別をクラスとして表し、クラスの公式度のみを指定し、個別のディレクトリに対しては、対応するクラスを指定することで指定を行っている（図 3.1）。

3.3.2 リンク推薦度の記述方法

リンクによる推薦には、3.2.2 節でも述べたように、自己作成ページに対するリンクと他者作成ページに対するものではリンクの持つ性質が異なる。特にサイト内においては、それぞれにリンクに含まれる意図が存在し、[75] に指摘されているように WWW 全体とは異なっている。そこで、リンクによる推薦度をあらわすために、直接数値で表すのでは

```

class    division      60
class    relevant      20
class    committee     80
class    teacher       50
class    student       30
class    admin         70
class    project       80

instance faculty top   _root_  ccs.cla.kobe-u.ac.jp/  ccs.cla.kobe-u.a
c.jp/index-jp.html
instance division     ame_d  top     ccs.cla.kobe-u.ac.jp/America/
ccs.cla.kobe-u.ac.jp/America/index-jp.html
instance teacher     ame_t  ame_d  ccs.cla.kobe-u.ac.jp/America/Kyoukan/
ccs.cla.kobe-u.ac.jp/America/Kyoukan/Kyoukan.index-jp.htm
instance division     asi_d  top     ccs.cla.kobe-u.ac.jp/Asia/
ccs.cla.kobe-u.ac.jp/Asia/index-jp.html

instance teacher     asi_t1  asi_d  ccs.cla.kobe-u.ac.jp/Asia/Kyoukan/
ccs.cla.kobe-u.ac.jp/Asia/Kyoukan/index-jp.html
instance teacher     asi_t2  asi_d  ccs.cla.kobe-u.ac.jp/Asia/Nakano/
ccs.cla.kobe-u.ac.jp/Asia/Nakano/index.html

```

図 3.1 公式度定義ファイルの例 (一部)

なくリンクの意図に応じた表 3.2 に示すようなリンク属性を用意した。この属性を各ページのリンク (<a href>タグ内) に埋めこむことでリンクの推薦度を指定できる。

ここでは、推薦の意図を 6 種類、本来同一の値になるようなページへのリンク (一つの論文を章毎に別ページにした場合など) には、equivalent を、「戻る」「back」などの推薦とは直接関係のないリンクには、ignore を付与した。また、一般的なリンクに関しては、同一作成者内のリンクか他の作成者へのリンクを区別し、それぞれ推薦度の高さに応じて official と personal、endorse と introduce、の各 2 種類の属性を準備した。実験における属性付け作業においては、「"back" や "戻る" などと書かれたリンクは ignore とする」、「同一作成者内のリンクのデフォルトは official とする」、「他作成者へのリンクのデフォルトは endorse とする」などといった自動化処理を図ることにより作業を軽減した。また、属性として推薦度を指定することで、サイトに応じて属性と推薦度の値を調整したり、反対に推薦度を調節することでサイト全体におけるページ信頼性を調整するなどの柔軟な運営を行うことができる。

表 3.2 リンク属性一覧

属性名	説明	推薦度
equivalent	章毎に分割された文章など意味的に同一なページへのリンク	1.0
official	ページ自己評価において公式なページへのリンク	0.95
personal	ページ自己評価において非公式なページへのリンク	0.4
endorse	他作成者ページに対する推薦を意味するリンク	0.8
introduce	他作成者ページに対する単なる紹介を意味するリンク	0.3
ignore	「戻る」などのナビゲーションだけのないリンク	0.0

3.4 実験

実験では、本提案方式、各ページ作成者の公式度のみによる評価、PageRank[53] の3つの方法により各ページの評価点を求めた。PageRankによる評価点は、本提案方法の合計点と同じになるように正規化を行っている。なお、PageRankにおける random surfer model の値は 0.15 として計算した。

表 3.3 に、いくつかの代表的なページを示す。この表では各ページについて、提案方法による評価点、PageRankによる評価点、ページの公式度、提案方法の評価点の決定される際の評価点の伝播を示している。表中の(1)~(4)は、「林」で検索を行った場合に検出されるページである。本サイト内には、教員である林が作成したページ(表中の(1)と(2))、学生の林が作成したページ(同(3))、ならびに、学部委員会が作成した教員紹介の中に林教授のページ(同(4))が存在する。(1)は林教授のトップページであり、講座の

ページからのみリンクされているが、リンク属性は ignore (自分の属する組織からの単純なリンクの評価は評価点に折込済みのため無視することにしたため) となっており、公式度である 50 点が提案方式の評価点となっている。また、(3)、(4) も同様にそれぞれの公式度がそのまま評価点となっている。(2) は、林教授の業績の自己紹介ページである。このページは、(4) のページから「最新の詳しい業績」としてリンクが張られている。つまり、このページは、学部委員会 (評価点 80) が公式性の高いページとして推薦はしているものの、コンテンツ自体を管理しているわけではない。このため、提案方法による評価点は、(4) > (2) > (1) > (3) という順になっているが、公式度という観点からは妥当であると思える。

これに対し、PageRank の結果は、作成者に対する評価とは関係なく、一部のページで突出した点を持つページが存在している。これらのページは、多くの被リンクを持つものであり、PageRank の持つ特性を反映したものといえる。なお、ページ (1) の PageRank の値は正確には 0 点ではないが、被リンクが 1 つしかなく、小数点以下 3 桁で丸めたため 0 点となっている。

ページ (5) は学生 A が作成したページであり、あるプロジェクトのレポートが書かれている。このページは学生 A の個人ページ内に置かれているが、プロジェクトの公式レポートの一つであり、プロジェクトページから endorse リンクが張られている。プロジェクトのページからの endorse リンクにより学生 A のプロジェクトに関するページが評価されることで、プロジェクトの公式度である 80 点が影響し 60.8 点となる。このようにリンクにより高い公式度を持つページからの評価伝播が発生している。ただし、学生 A のページのうち高い評価点を得ているのは、プロジェクトに関するページだけであり、それ以外のページでは学生への公式度である 30 点となっている。これは、リンクによる評価伝播では「戻る」などのリンクは ignore 属性として判断されるため、評価伝播は発生しな

表 3.3 ページ評価の例

ページ内容	提案方法	Page Rank	公式度 (q_i)	評価伝播
(1) 林教授のトップページ	50	0	50	50
(2) 林教授のプロフィールページ	64	4	50	80 → 64
(3) 林という学生のページ	30	29.2	30	30
(4) 教員情報 林教授	80	600	80	80
(5) 学生 A の作成したレポートのページ	60.8	7.2	30	80 → 64 → 60.8
(6) 学生 B の作成したページ	30	85.5	30	30

いことによる。一般に同一作成者の作成したページでも「戻る」などのリンクを取り除くと関連性の深いページへのリンクしか張っていないことが多く、関連性の低いページにまで評価点が上がってしまう可能性は低いと考えられる。

ページ (6) は学生 B の作成したページであり、PageRank における特徴的な評価を示している。このページは多数のページからリンクされているため、PageRank において高い評価点が与えられている。これらのリンクの多くは学生 B 自身のページからのリンクが多くを占めているが、PageRank ではこのようなページの評価点も上がってしまう。これ

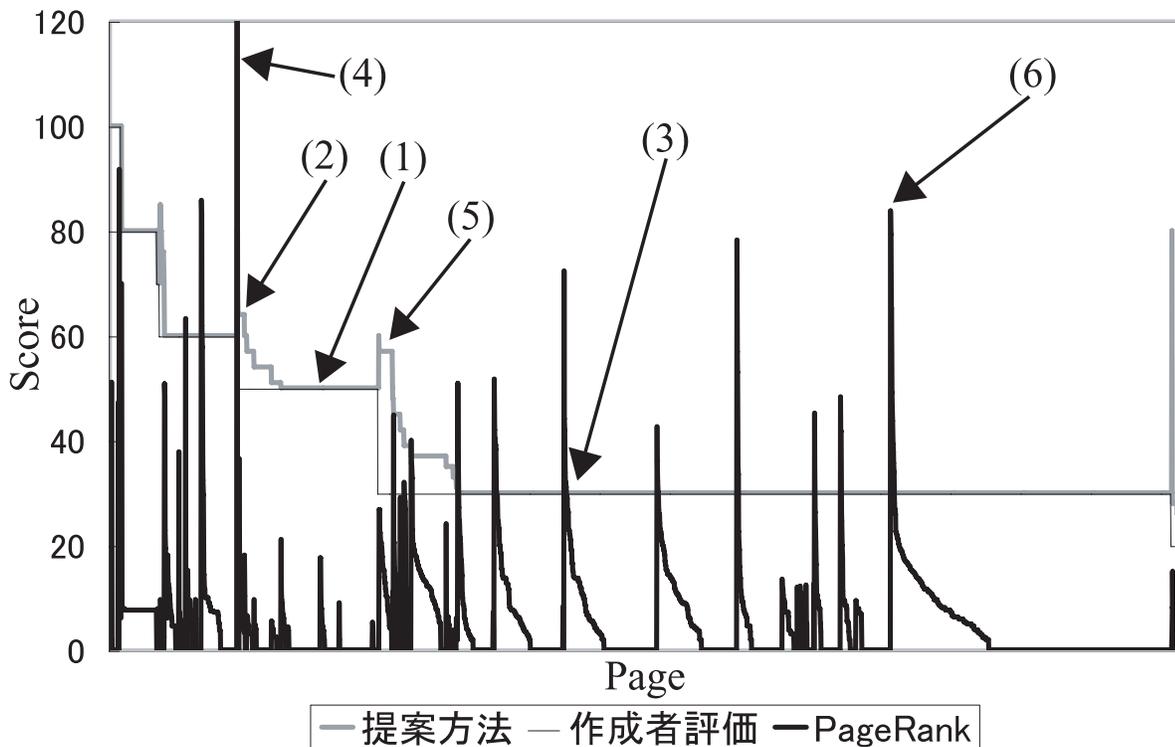


図 3.2 各方式の評価点

に対し提案方法では公式度である 30 点のみが与えられている。

図 3.2 に各評価方法による評価点の全体分布と、表 3.3 の各ページの位置を示す。この図では、ページ作成者への公式度、提案方法による評価点、ページ番号の優先度によりページを並べ替え、 x 軸としている。

提案方法による評価点は、ほぼ作成者評価と一致しているが、一部のページで作成者評価を上回っている。評価点が上がっているページは、高い公式度を持つ作成者により作成されたページからリンクされたページであり、我々の想定するページ公式度として妥当性を持つと考えられる。

本提案方法では、基本的には作成者評価に準じた評価点を持ちつつ、ページ (2) やページ (5) のように評価の高い作成者からリンクをされることで、個人ページ内のページでありながら高い評価を得ることができる。上の実験では、各ページの評価点は、最低でも

その作成者の公式度が付与された。しかし、同じ教員が作成したページにおいても、プロフィールや研究業績、趣味のページとでは、公式さが異なるはずである。そこで、作成者に対する公式度を各ページ作成者の管理ページ群のトップページ（インデックスページ）に対してのみ与え、その他のページに対する公式度を0点として実験を行った。すなわち、

$$q_i = \begin{cases} \text{作成者へ公式度 (トップページ)} \\ 0 & \text{(他ページ)} \end{cases} \quad (3.1)$$

とした。

公式度を各ページ作成者のトップページに対してのみ与えることで、ページ作成者が自己作成ページを推薦度付きリンクで自己評価を行う。つまり、公式度は作成者に対して与えられる作成者評価として考えることができ、各ページ作成者がその評価の範囲内で自分の作成したページを評価することができる。これは、作成者による自己作成ページの自己管理と考えられる。

表 3.4 にこの実験でのいくつかの代表的なページを示す。表中の自己管理付き評価がこの実験での評価点である。一律外部評価の結果は表 3.3 における提案方法と同じ方法であり、作成者のページに対し一律に公式度を与えている。また、作成者評価とは作成者に対する公式度（その作成者のトップページの公式度）を示すものであり、表 3.3 の公式度と同じ値となっている。

ページ (1)~(6) は表 3.3 のページ (1)~(6) と同じページである。ページ (2), (3), (4), (5) においては、 q_i は0点であるが、トップページからのリンクのより official 属性の評価がされ評価点が付けられている。

ページ (7) は学部の関連組織の紹介ページであるため、作成者は学部当局であり作成者評価は100点である。しかし、紹介としての introduce 属性のリンクが貼られているため

表 3.4 自己管理付き評価の例

ページ内容	自己管理 付き評価	一律評価	公 式 度 (q_i)	作成者評 価
(1) 林教授のトップページ	50	50	50	50
(2) 林教授のプロフィールページ	60.8	64	0	50
(3) 林という学生のページ	27	30	0	30
(4) 教員情報 林教授	76	80	0	80
(5) 学生 A の作成したレポートの ページ	57.8	60.8	0	30
(6) 学生 B の作成したページ	30	30	30	30
(7) 学部関連組織の紹介	30	100	0	100
(8) 学生の趣味のページ	12	30	0	30
(9) 終了したシンポジウムのペー ジ	0	100	0	100

に評価は 30 点となっている。同様にページ (8) では、趣味のページに対して学生自身が personal 属性のリンクを貼ることで評価点を低く抑えることができる。ページ (9) は、学部主催のシンポジウムの案内が書かれたページである。シンポジウムの開催に先立って学部当局の管理領域内に作成され、トップページからリンクがされていた。その後シンポジウム終了とともにトップページからのリンクが削除されている。学部トップページからのリンクが削除されたため、いくつかの被リンクは残っているものの、提案方法による評価点は 0 点となっている。

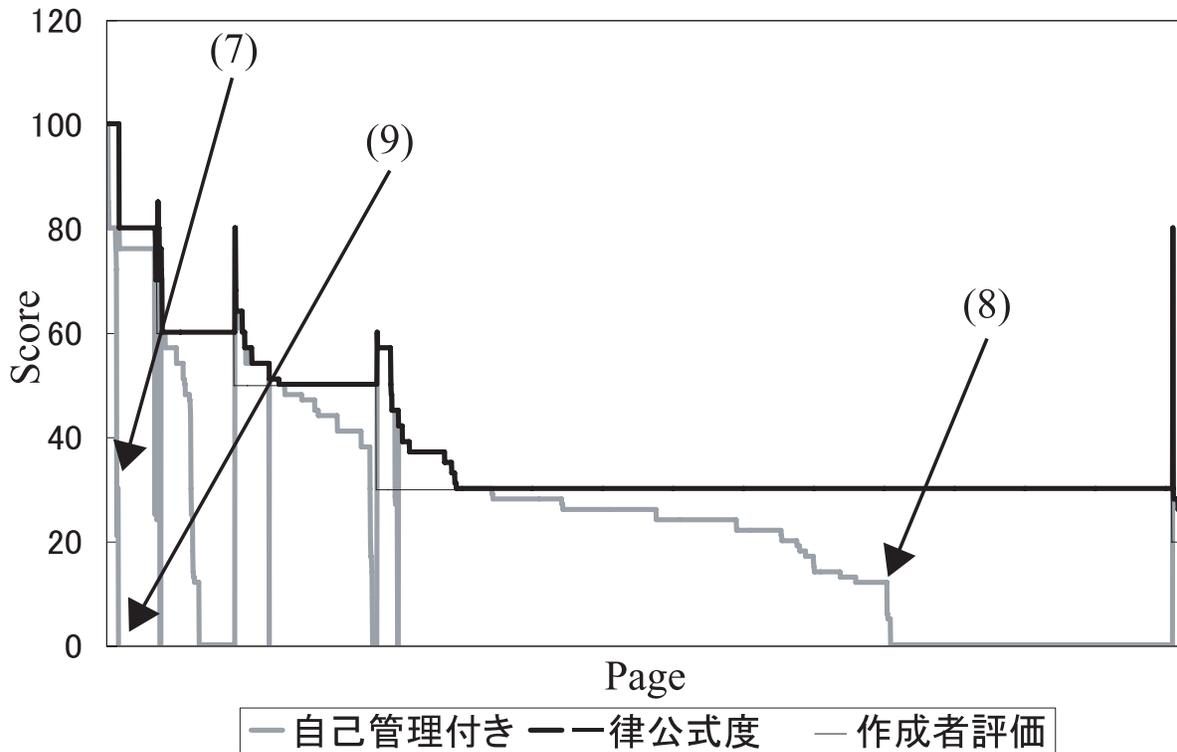


図 3.3 自己管理付き評価点

図 3.3 にこの方法による各評価方法の実験結果を示す．この結果でも提案方法は作成者評価に概ね準じた結果となっているが，高い作成者評価を持つが低い評価を持つページや，0点となるページが現れている．しかし，図 3.2 とは異なり，高い公式度を持つ作成者のページであっても点数の低いものや 0 点となっているページが存在する．これらのページはリンク切れや作成者によるページ自己管理の結果として現れたものである．

この結果から，公式度を 3.1 式に従い，作成者トップページにのみ与え，その他のページを 0 点とした場合でも，公式性の高いページは高い評価点を得られ，リンク切れなどで有用性が低くなったページの評価点が下がっている．全体としては公式度に従いつつ，ページの公式性をより反映した評価点となっていると考えられる．

図 3.4 では，太い線が評価の伝播木を表す．この図には，プロジェクトトップページ，学生 A トップページ，学生 B トップページの異なる 3 つのルートを持つ木，伝播木 1 ~

3が存在している。また灰色の領域は各作成者の管理領域を示す。学生 B の作成したすべてのページは伝播木 2 に属し、学生 B 自身の管理領域内に収まっている。一方、学生 A のレポートのページは、学生 A の管理領域にありながら、プロジェクトをルートとする伝播木 1 に含まれている。つまり、学生 A のレポートは、公式さという視点からはプロジェクトのページであるとみなすことができる。一方、学生 A の管理領域では、トップページをルートとする伝播木 2 に属するページと、伝播木 1 に属するページが存在している。つまり、同じ学生 A の作成しているページであるが、プロジェクトに属するページと学生 A に属するページかが伝播木によって区別することができる。このように評価の伝播木によって、高い公式度を持つどのような人が、そのページを推薦しているかを知ることができる。ページ間の推薦関係を知ることによって、ページ内容に対する評価をする際の判断材料として利用することができる。

提案方法では、高い公式度を持つページからのリンクの存在に大きな影響を受ける。この性質を利用することでリンク有無のチェックにも利用することができる。作成者評価点と提案方法による点を比較することで、本来あるべきリンクのリンク切れや不要になったページの発見に役立つ。つまり、ページ作成者によるページの自己管理だけでなく、サイトとしてのページ管理にも有効であると考えられる。

以上のように提案方法による評価は、サイトとしての公式さの判断や作成者によるページ自己管理において有効な材料となると考えられる。

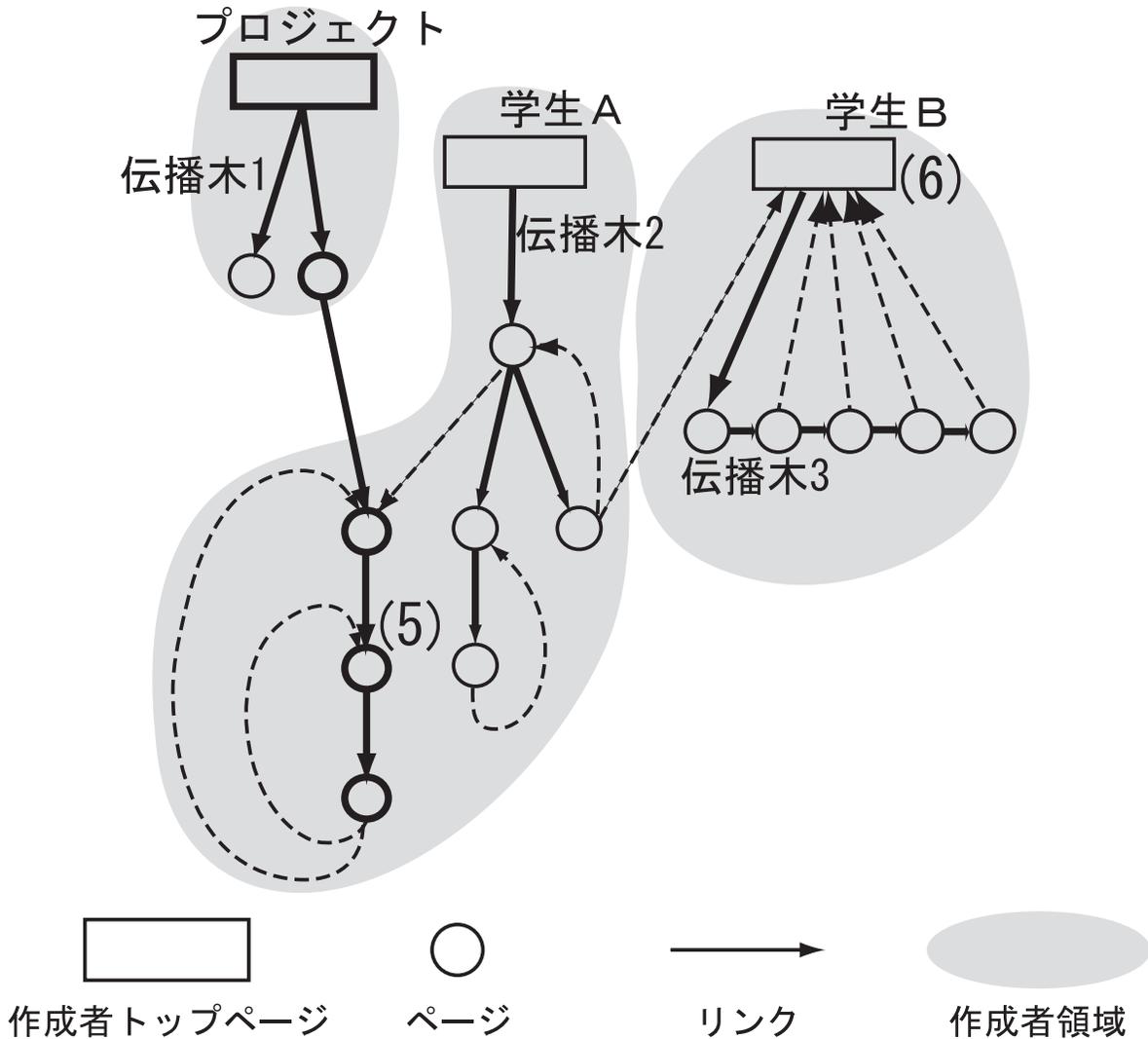


図 3.4 他作成者ページへの公式度の伝播

3.5 議論

3.5.1 サイト構造と組織構造

ページ作成者の公式度と URL の定義ファイルでは、実際のサイトのディレクトリ構成とは関係なく、組織構造とサイト構成、およびそれらの公式度の関係を記述できる。例えば、図 3.5 のように平坦なディレクトリ構造を持っているサイトでも、組織構造とディレ

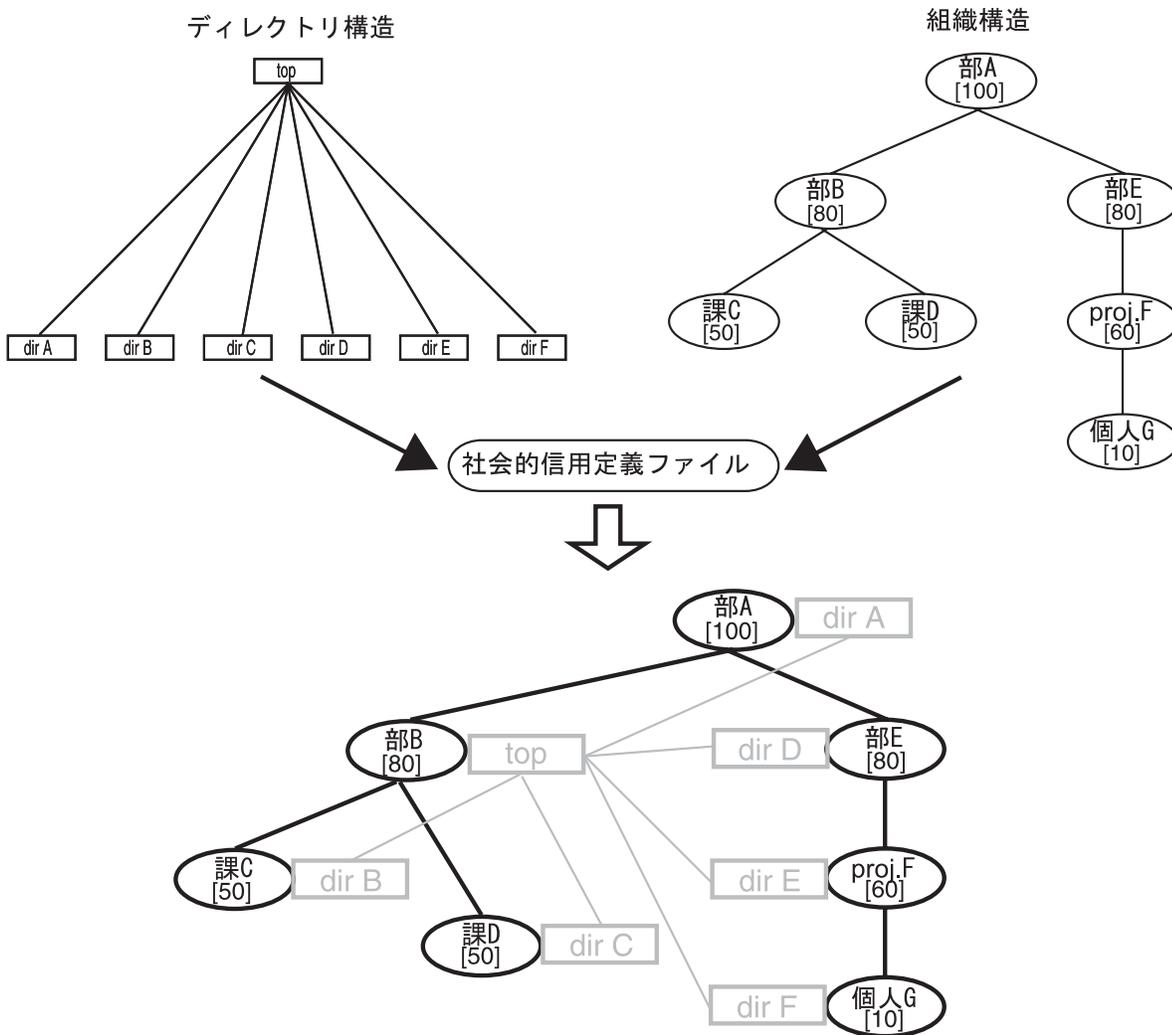


図 3.5 組織構造とサイト構造

ディレクトリ構造の対応を正確に記述することが可能である。これは、組織の各部署と URL の関係を示すことにもなり、信頼性において重要な要素となるものと考えられる。

3.5.2 様々なサイトへの適用可能性

本方式の適用可能性は、サイトの持つ特性により変化すると考えられる。また、社会的信用をどのような基準により決定するかによっても、評価点の持つ意味は異なる。ここでは、本方式の適用可能性とサイトの種類と社会的信用の意味について検討を行う。

例で示したように大学や研究所や NPO などのサイトでは、社会的信用の異なる様々な作成者が混在しているため、評価点はページの公式性を示すものとして利用することができる。また、本方式の特徴である評価の伝播木を示すことで、ページの責任関係を示すことができる。

企業などのサイトは、現在ほぼすべてのページは運営する企業によって統一的に管理され、ほとんど公式度としての違いは存在していない。本方式をサイト管理に適用することで、評価点により公的なページと非公式なページを区別することができる。つまり、組織としての責任を持ちつつ、サイト内にページ作成者の責任で自由度の高いページを持つことが可能である。ページ作成者の社会的信用を定義したファイルによりサイト全体の公式性を管理し、各ページ作成者はリンク属性により自己作成ページの公式性を管理できる。さらに、伝播木によりページ間の責任関係を確認しつつ、サイト構築を行なうことができる。

ポータルサイトでは、公式度的な社会的信用の利用は難しいが、社会的信用をお勧め度としての利用や、自己作成コンテンツとニュースなどの他機関が作成したコンテンツとを社会的信用によって区別することなど利用が考えられる。

ISP などのサイトでは、評価点を公的ページとユーザページの区別や、お勧め度としての利用、公式性を示した上でのオフィシャルページとユーザページのシームレスなナビゲーションなどへの利用ができる。

個人サイトでは、お勧め度としての利用やサイト内のページ管理、リンク管理などへ利用できる。また、社会的信用の意味を自由に設定し、趣味に応じたスコアによりナビゲーションを行うなど、一種の Web におけるパーソナライズ的手法 [76] としても可能性を持つ。

これらの、評価点の利用可能性と社会的信用の意味をまとめると表 3.5 のようになる。

表 3.5 サイトのタイプと適用可能性

サイト	現状		適用可能性	
	自由度	管理	自由度	管理
大学	○	×	○	○
企業	×	○	○	○
プロバイダ	△	×	○	△

3.6 応用の可能性

サイトにおける情報提示および管理ツールとしての応用の可能性について検討を行うため、以下の2つについて実施を行った。

3.6.1 キーワード検索との連携

実際にサイト内を閲覧するためには、リンクを辿るだけでなくキーワード検索によって目的のページを探すことは多く、有効な方法である。キーワードに依存しないことは本手法の特徴の一つであるが、キーワード検索と組み合わせることは、実際のページの閲覧においては非常に有効な手段である。

そこで、日本語全文検索システムである *namazu* [77] を利用することで本手法とキーワード検索との連携を実現した。本章の実験で利用した神戸大学国際文化学部のミラーを WWW サイト内に作成し、本手法の評価と *namazu* を用いたインデックス作成を行い、*cgi* プログラムにより Web ブラウザからの検索を実現した。

さらに、検索結果を公式度とどのような作成者によるページであるかを、サイト全体に

において可視化して提示するプログラムを試作した(図 3.6)。このプログラムにより、あるキーワードを含むページの中で、どの作成者(部署や個人)が作成したページが高い公式度を持つかを容易に確認することができる。したがって、同じキーワードを含む公式度の高いページと、うわさ情報のように学生などが作成した公式度が低いページが存在を図として概観することができ、通常埋もれがちなページの発見や閲覧者によるページ選択に利用できる。また、本手法の妥当性の確認にも有効である。

3.6.2 可視化ツール

社会的信用とページ作成者の定義ファイルを利用することで、組織構造を可視化することで、サイト構造とは独立に組織構造に基づいたナビゲーションおよびサイト管理に利用できる。

図 3.7 に組織構造を可視化したプログラムの実行結果を示す。図における、組織構造を示す領域がボタンとなっており、作成者名をクリックすることで該当するページを閲覧することができる。本プログラムは組織構造に基づくナビゲーションだけでなく、組織構造の全体とその公式度を確認することができるため、サイト管理における公式度の確認にも利用できる。さらに、前述のキーワード検索の可視化とも組み合わせることで、必要なリンクがないために公式度が高いはずであるが公式度が低くなっているページや、逆に古い情報へのリンクの消し忘れにより不必要であるのに公式度が高くなっているページなど、リンク関係に由来する状態を発見するために利用できるだろう。

3.7 まとめ

本章では、サイトの持つ特徴を利用することによるページ信頼性の評価手法の効果的な適用手法について示すとともに、実験を行うことで第2章において提案したページ評価手

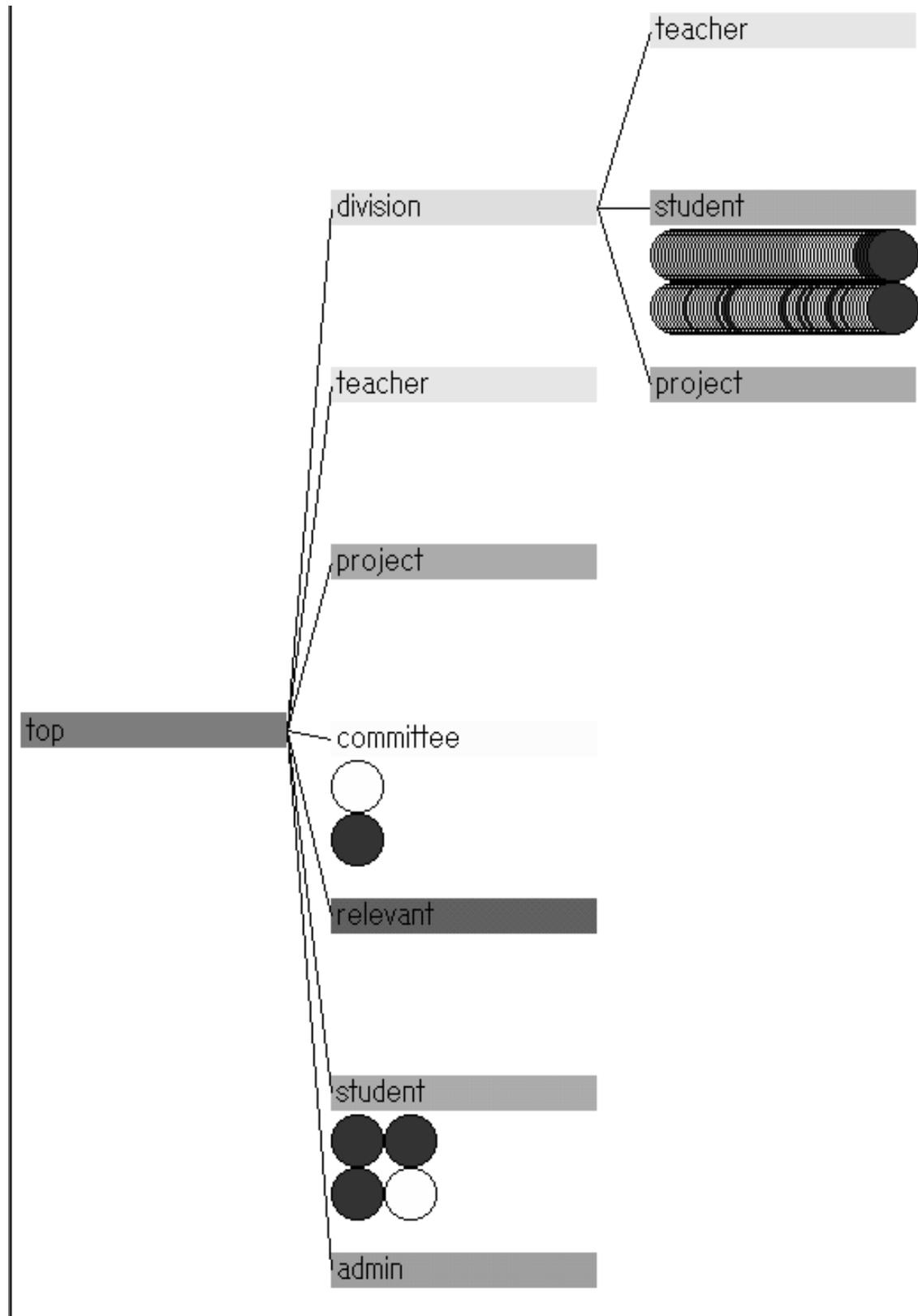


図 3.6 公式度の可視化

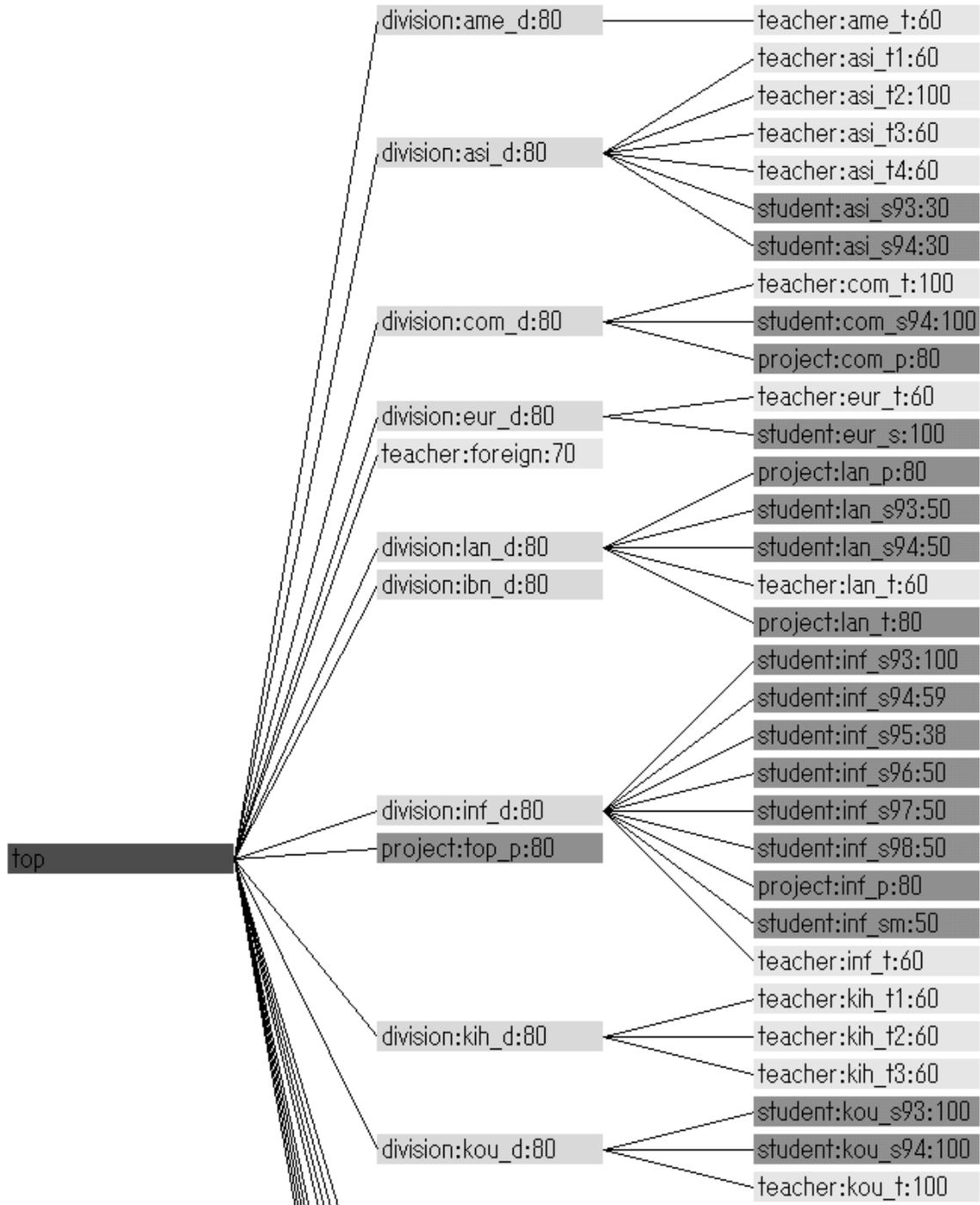


図 3.7 組織構造の可視化

法の妥当性を示した．さらにキーワード検索との組み合わせや結果の可視化プログラムにより，本手法による公式度としてのページ信頼性評価手法の可能性を示した．

本手法を利用してサイトにおける情報の公式性を明示した情報発信を行うことで，WWWにおける一定の規律を確保した上で自由なコミュニケーションへの可能性を示している．また，サイトにおける情報の自己管理手法としても有効であると考えている．

第 4 章

良いページ悪いページの判定手法

4.1 はじめに

本章では，社会的信用の考え方を拡張することで，良いページと悪いページに注目した手法を提案する．第 2 章において示した社会的信用は，信用の高いページ作者に対して指定を行うものであった．本章では，負の社会的信用と考えられる一般に不信があるようなページ作成者をも指定する．正と負の信用を指定することで，良いページと悪いページを判定する手法を示す．

良いページや悪いページを判断することは，適切な利用者に適切な情報を提供し，不適切な利用者には情報を提供しないための有効な手段となっている．キーワードやリンク解析を行う検索エンジンなどでは，ある程度の良いページを得ることができるが，利用者が意図した「良い」ページが得られるとは限らず，また悪いページを判断することはできないことが多い．企業や子供向けの情報フィルタリングでは，キーワードや URL によりコンテンツを制限しているが，データベースの維持が必要であったり，良いことを判断することが難しいといった問題がある．本手法では，コンテンツに依存しないリンク構造解析により，自動的に WWW 上のページに対して良い/悪いを判定することができる．

本手法では、以下の手順で Web 上のページ評価を行う。

- [1] 評価者が良いページ、悪いページと判断されるページ群を指定する
- [2] 指定されたページ群を中心とした Web 上のページのリンク関係を解析する
- [3] 得られたリンク関係により各ページの良い/悪いの判定を行う。

ページ群の指定においては、既知のページの中から良いページと悪いページを選択する。通常、ページに対する評価は、どんなページであるか、誰が書いたかなどを含めて最終的に内容を読んで判断している。既に内容を知ってるページについては評価も確定しているため、良いページと悪いページを判断することができる。

リンク解析においては、良いページと悪いページの持つ性格からリンクの性質に違いがあることを利用し、良いページと悪いページのリンク関係に基づくモデル化を行っている。このモデルからグラフ理論を用いて、ページを判定するためのアルゴリズムを導き出している。このアルゴリズムを使うことで、いくつかのページに対して良い/悪いと指定を行うだけで、リンク構造から良いページと悪いページの判定を自動的に行うことができるようにした。

本提案手法の主な特徴としては、以下の点がある。

1. ページに含まれる単語やコンテンツに依存しない
2. 内容を見ることなく、良い/悪いの判定を行える
3. ページ解析において、主観の違いを反映した初期値を指定できる

例えば、子供向けの良いページ/悪いページや大人向けの良いページ/悪いページなど、異なる主観により初期値を指定できる

4. ページの性格の違いによるリンクの性質を利用してリンク解析を行うことで、良いページだけでなく、従来の方法では判断しにくい悪いページの判定を行える。

本判定方法は、ページの持つ特性や傾向の判断に利用することができる。さらに情報のフィルタリングやパーソナライズなどに応用可能であると考えている。

4.2 良いページと悪いページ

本手法における考える良いページと悪いページの定義および判断基準について述べる。

4.2.1 良いページと悪いページの基準

あるページに対して、良いと判断するか、悪いと判断するかは、評価者や評価する状況によって異なり、一意に決定できるものではない。しかしながら、一般的に良い悪いと判断しているような、ある程度の倫理的な基準は存在していると思われる。例えば、有名な企業や官公庁のページは、安心して見ることができるような倫理性や信用が高く信頼できる内容を持つ良いページと見ることができる。アダルト系のページや違法性の高い内容を扱ったページは、青少年にとって有害といわれるような悪いページと見ることができる。本手法では、このような一般的に使われている倫理的な判断を、良いページ/悪いページの判定基準とする。

ただし、一概に良い/悪いと判断できない場合もある。例えば、競馬や競輪などの公営ギャンブルは、未成年にとっては悪いページと言えるが、公的な機関によって運営されている良いページとも言える。このようなページを良いとみなすか悪いとみなすかは、評価者の主観に依存するものである。ここでは、このようなある程度の評価者の主観的な判断を判定基準に持ち込むことを許容することとする。

4.2.2 ページの直接的評価

実際の Web において、我々はどのようにページの良い/悪いを判定しているだろうか。通常は、直接ページ内容を読むことによって、どのようなページであるかというページの性格を判断することで、良い/悪いを判定している。ここでは、ページ内容自身から直接的にページを判定することをページの直接的評価と呼ぶ。直接的評価においてページの性格を判断するための要因としては、誰が（企業、官公庁、大学、個人など）、どのような立場で（組織、公人、個人として）、どのような目的で（宣伝、広報、趣味など）、どこに（co.jp、ac.jp、or.jp、ne.jp などのドメイン名やディレクトリ）、いつ作成したかなどが考えられる。このような要因からページの性格を判断する際には、評価者の持つ社会的知識が大きく影響する。例えば、誰がページを作成したのかわかったとき、作成した組織や個人を良く知っていれば、どのような立場から書かれたものかなどの内容の持つ背景を含めた正確な判断を行うことができる。

では、良いページと悪いページはどのような性格を持つページだろうか。良いページは、企業や官公庁など信用できる作成者が正確さをもって書いたページであり、いい加減なことや嘘などは書かれていないと思えるページといえることができるだろう。つまり、良いページは、社会的責任が担保された信用を持っている、あるいは信頼できる性格を持つページと考えられる。悪いページは、倫理的に問題があったり、悪意がある作成者によって書かれたページであり、内容の信憑性や作成者への責任を問えないようなページといえることができる。つまり、悪いページは、信用がない、見ることに自己責任が要求されるといえる性格を持つページと考えられる。

ここで、本手法において定める直接的評価における良いページと悪いページの性格と、その要因をまとめると以下ようになる。

良いページ 社会的責任をもって作成された信用できるページ，すなわち社会的に信用があるページ

例えば，有名な企業や官公庁や信頼できる人物など，社会的な信用のある作成者によって公式に書かれたページ．

悪いページ 信憑性の低い，倫理的な問題がある，自己責任が要求されるページ，すなわち社会的に不信をもたれるページ

例えば，アダルト系，風俗系，違法性がある有害な内容を含んだページ．

4.2.3 リンクによる間接的評価

良いページ/悪いページの判定においては，リンクによるページ間の関係も重要な判断要因である．我々がリンクを辿りページを見るとき，どのページからのリンクを介してそのページに到達したかといったページ間のリンク関係を，そのページに対する判断要因として利用している．どのようなページなのか，すなわちページの性格から，そのページから出ているリンクの性質を推測し，リンク先ページを判断し良い/悪いの判定を行う．ここでは，このようなページ内容を直接参照しないリンクによるページの判定を，良いページ/悪いページの間接的評価と呼ぶ．

通常，リンクはページによるページの推薦として解釈されている．良いページのリンクは，社会的責任を持って作成されているという良いページの持つ性格から，良いページの責任を持って信頼できるページを推薦していると考えられる．これは，良いページの持つ信用がリンクによりリンク先ページに伝播していると考えられる [61]．つまり，良いページからのリンクは信用を伝えるという性質を持っている．

同じリンクによる推薦であっても，悪いページからの推薦では含まれる意味が異なる．悪いページでは，社会的な責任をほとんど考慮することなく作成されるという悪いページ

の性格から、リンクに対しても責任を負うことはない。たとえ悪いページからのリンクから不快なページに辿り着いても自己責任でしかない。逆に、もしリンクを辿ることで悪いページに到達した場合には、悪いページを推薦しているページとして、リンク元ページに対し信用とは逆の不信を感じることになる。つまり、悪いページの場合には、信用の伝播するのとは逆に、リンクと逆方向にリンク元ページに対し不信が伝播する性質があると考えられる。

4.2.4 リンク関係によるページ判定

以上により、ページに対して良い/悪いを指定する直接的評価とリンクの性質を利用した間接的評価を行うことで良いページ/悪いページの判定を行うことができる。また、間接的評価の過程においては、ページ内容を直接確認することなくリンク先ページの判断を行っている。

ここで、良いページ/悪いページとのリンク関係に基づいてまとめると、以下のようになる。

良いページからリンクされているページ　良いページから信頼を持った推薦を受けたことになり、良いページである可能性が高い。良いページの責任においてリンク先ページに信用を与えていることとなる。つまり、良いページの持つリンクの性質により、良いページの信用が伝播している。

良いページへリンクしているページ　良いページにリンクしているだけでは良い/悪いの判定要因とはならない。

良いページは良いページにリンクしている可能性が高いが、悪いページはそのリンクの性質から、どんなページに対してもリンクする可能性がある。

悪いページからリンクされているページ　リンクはどんなページからも無許可でされる

可能性があり，良い/悪いの判定要因とはならない．

悪いページはそのリンクの性質から良いページにも悪いページもリンクしている．

悪いページへリンクしているページ 悪いページを推薦することでページ責任を問われる，悪いページである可能性が高い．

悪いページへリンクすることは，ある種の非倫理性を意識した上で行っていると考えられ，社会的信用を持ったページとはいえない．

4.3 ページ評価モデル

これまで述べたページの評価の考え方に基づいて，良いページ，悪いページにおけるページ関係について仮定を行い，ページ評価モデルを構築する．

4.3.1 リンク関係と評価に対する仮定

良いページの基準とそのリンクによる関係から，良いページにおけるリンク関係に対し仮定を行った．

仮定 4.1 良いページからリンクされているページは良いページ

悪いページはどんなページにもリンクするため，悪いページにリンクされてもページ判断には利用できない．良いページでは，悪いページへの推薦は自ページの信頼性や評価に影響するため，関係のあるページや，良いと判断しているページにしかリンクをすることがない．

つまり，良いページからリンクされたページは良いページと判断することができる．

同様に悪いページの基準とそのリンクによる関係から，悪いページのリンク関係に対しても仮定を行った．

仮定 4.2 悪いページをリンクしているページは悪いページ

悪いページを推薦しリンクするということは、同類や仲間、あるいは倫理的な悪意も持っていると考えられる。したがって、悪いページをリンクしているページも悪いページと判断できる。

ここで、仮定に基づき良いページと悪いページの定義を行う。

定義 4.1 良いページ: 仮定 4.1 を満たすページはすべて良いページである

定義 4.2 悪いページ: 仮定 4.2 を満たすページはすべて悪いページである

本手法では、この定義に基づき良いページと悪いページの判定を行う。

4.3.2 ページ状態の判定

ここでは、4.3.1 節で述べた仮定および定義に基づき、各ページの評価を行う。各ページ間のリンク関係をグラフによって表現すると、Web ページをノード、リンクをエッジとした有向グラフで表すことができる。ここで、有向グラフを G 、 G 上のあるノードを n_i 、ノード n_i からノード n_j へのリンクをエッジ $e_{i,j}$ と表記する。

図 4.1(a) にグラフ表現の例を示す。図において、リンクによる直接結び付けられたノードのことを、隣接ノードまたは隣接ページと呼ぶ。例えば、ページ n_1 とページ n_5 とは、エッジ $e_{1,5}$ によって隣接している。

以下では、ページの「良い」、「悪い」などの判定された結果のことを、“ページの状態”と呼ぶ。ページの状態には、good, bad, unknown の 3 つの状態がある。good, bad はそれぞれ「良い」、「悪い」と判定されたページ、unknown はどちらとも判定されていないページである。

定義 4.1, 定義 4.2 より、各ページの状態は隣接ページのリンクの状態に依存する。この

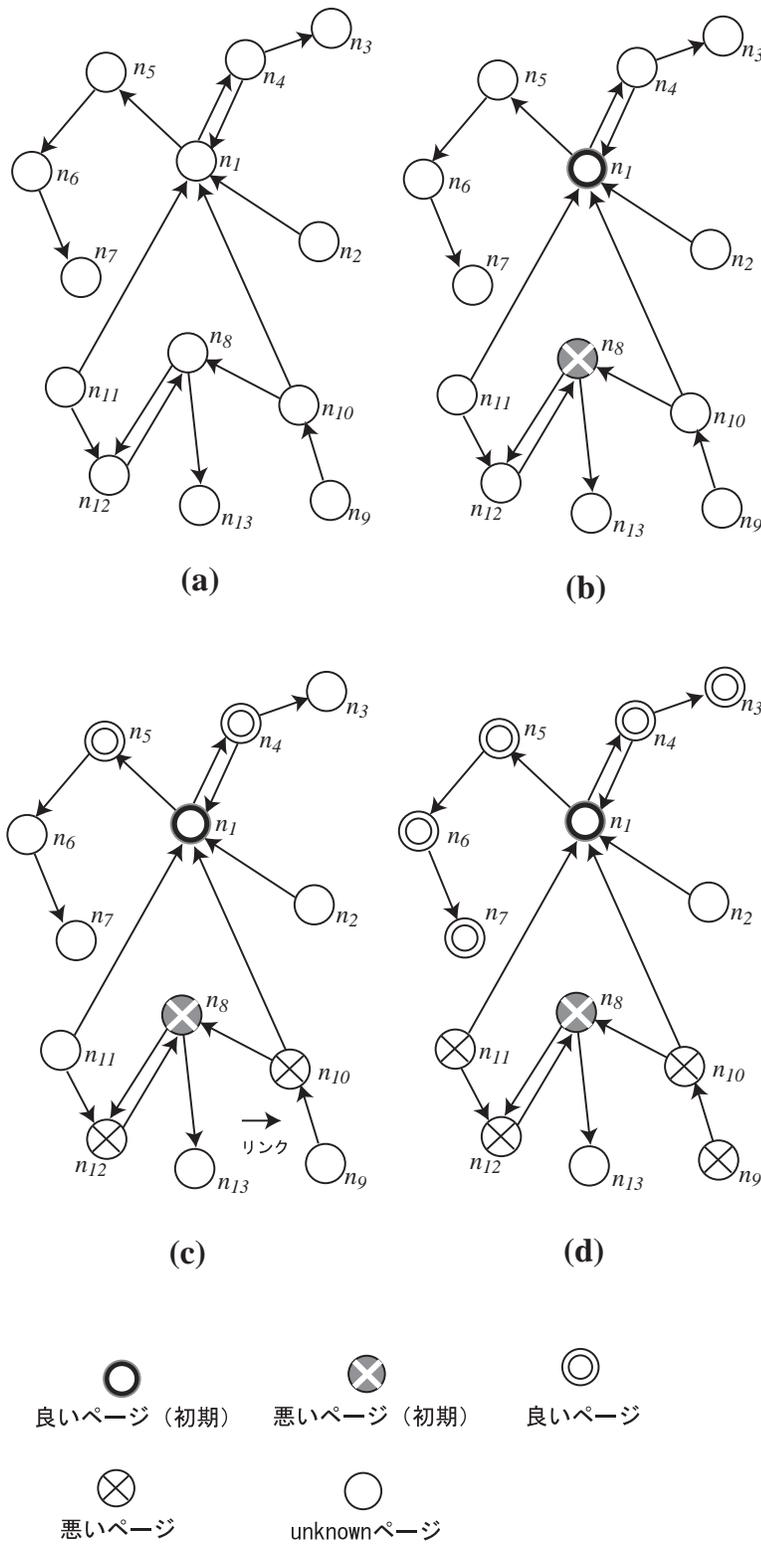


図 4.1 Web のグラフとその状態

隣接ページのリンク状態を”隣接リンク状態”と呼び、あるページ(ノード n_i とする)に対する隣接リンク状態 L_{n_i} を次の状態を用いて表現する。

G_{in} : good ページからのリンクの状態

G_{out} : good ページへのリンクの状態

B_{in} : bad ページからのリンクの状態

B_{out} : bad ページへのリンクの状態

各状態の値は、それぞれ条件を満たすリンクを持つ場合は 1、それ以外の場合は 0 とす

る。 L_{n_i} をこの 4 つの状態の組、すなわち、

$L_{n_i} = (G_{in}, G_{out}, B_{in}, B_{out})$ として表す。

リンク隣接状態の組合わせとしては、以下の 16 通りが存在する。

$p_1 : (0, 0, 0, 0)$ $p_2 : (0, 0, 0, 1)$

$p_3 : (0, 0, 1, 0)$ $p_4 : (0, 0, 1, 1)$

$p_5 : (0, 1, 0, 0)$ $p_6 : (0, 1, 0, 1)$

$p_7 : (0, 1, 1, 0)$ $p_8 : (0, 1, 1, 1)$

$p_9 : (1, 0, 0, 0)$ $p_{10} : (1, 0, 0, 1)$

$p_{11} : (1, 0, 1, 0)$ $p_{12} : (1, 0, 1, 1)$

$p_{13} : (1, 1, 0, 0)$ $p_{14} : (1, 1, 0, 1)$

$p_{15} : (1, 1, 1, 0)$ $p_{16} : (1, 1, 1, 1)$

このうち、良いページ、すなわち定義 4.1 を満たすのは、 $p_9, p_{10}, p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16}$ の 8 通りであり、悪いページ、すなわち定義 4.2 を満たすのは、 $p_2, p_4, p_6, p_8, p_{10}, p_{12}, p_{14}, p_{16}$ の 8 通りがある。特に、 $p_{10}, p_{12}, p_{14}, p_{16}$ は、どちらの定義をも満たす。

しかし、良いページの定義が正しいとすると、「良いページは悪いページをリンクしない(仮定 4.1 の裏)」となり、この 4 つのパターンは存在しないことになる。実際の Web

ページにおいては，このような状態に該当するページも存在することが考えられる．また，評価者の初期値の選択によっては，このような状況が起ることもありうる．しかしながら，それらの取り扱いについては後述することとし，このようなパターンは存在しないものとして話を進める．つまり，以下の仮定が成り立つものとする．

仮定 4.3 良いページの定義と悪いページの定義を同時に満たすページは存在しない

定義 4.1 も定義 4.2 も満たさないのは， p_1, p_3, p_5, p_7 であり，unknown 状態と判定される．しかし，unknown ページの中でも，悪いページからのリンクがあるものについては，リンク関係によってページの持つ傾向を推測することができる．本手法では，このようなページを悪い可能性があるページとして，gray 状態にあるページと呼ぶ． p_3, p_7 が gray 状態に該当する．このようなページは，悪いページから知られ評価されているが，一般的に有名なページであるような良いページには含まれない．つまり，悪いページには知られているが良いページではないページであり，悪いページに近い傾向をもつページと考えられる．

ただし，gray ページは，良いページと悪いページがすべて決定した後に残った unknown ページから得られるものである．したがって，ページ判定の過程では，gray 状態から good になることや bad になることがある．例えば，マクロメディアなどの有名なページでは，当初は悪いページからのリンクが見つからないために gray と判断されているが，良いページからのリンクが見つかった時点で良いページと判定される．このようにすべての good および bad と判定されるページが決定した上でのみ，gray ページの判定が行うことができる．

以上より，仮定 4.3 を満たすものとするとき，gray を加えた上で L_{n_i} のパターンは，次に 4 つに分類される．

1. 良いページ，すなわち状態が good に定義されるページ

パターン $p_9, p_{11}, p_{13}, p_{15}$ がこれを満たす．

2. 悪いページ，すなわち状態が bad に定義されるページ

パターン p_2, p_4, p_6, p_8 がこれを満たす．

3. unknown ページのうち，悪いページからのリンクがあるページ，すなわち状態が gray に定義されるページ

パターン p_3, p_7 がこれを満たす．

4. 上記のどれにも定義されていないページ，すなわち状態が unknown に定義されるページ

パターン p_1, p_5 がこれを満たす．

すべてのページに対する良いという判定において，ページを分類すると 1. は良いページであり 2.3.4. は良くはないページと分けることができる．同様に悪いという判定においては，1.4. は悪くはないページであり，2. は悪いページ，3. は悪いかもしれないページと分けることができる．

4.3.3 漸化的なページの判定

ページ判定の初期状態では，評価者はいくつかのページに対し「良い」，「悪い」を判定しているが．その他のページについては判定を行っていない．つまり，初期状態では，評価者の指定したページのみが good, あるいは bad 状態にあり，残りのページは unknown 状態にある．図 4.1(b) に初期状態の例を示す．この例において，評価者は n_1 を good, n_8 を bad と評価したことを示す．

ここで，ノード n_i の初期状態を $S_0(n_i)$, k 回目のページ判定結果における状態を $S_k(n_i)$ と表す．

図 4.1(b) の例に対して, 4.3.2 節で述べた判定手法を適用すると, ノード n_4 のページ隣接状態が $L_{n_4} = (1, 0, 0, 0)$, ノード n_5 では $L_{n_5} = (1, 1, 0, 0)$ となり, $S_1(n_4)$, $S_1(n_5)$ は good と判定される.

逆にノード n_{10} のページ隣接状態は $(0, 1, 0, 1)$, ノード n_{12} では $L_{n_{12}} = (0, 0, 1, 1)$ となり, $S_1(n_{10})$, $S_1(n_{12})$ は bad と判定される (図 4.1(c)).

さらに漸化的に, ページ隣接状態が変化することにより, ノード n_3 のページ隣接状態が $L_{n_3} = (1, 0, 0, 0)$, ノード n_5 が $L_{n_5} = (1, 0, 0, 0)$ となり, $S_2(n_3)$, $S_2(n_5)$ が good と判定される. また, ノード n_9 のページ隣接状態が $L_{n_9} = (0, 0, 0, 1)$, ノード n_{11} が $L_{n_{11}} = (0, 1, 0, 1)$ となり, $S_2(n_9)$, $S_2(n_{11})$ が bad と判定される.

最終的に, 図 4.1(d) に示すように, ノード n_3, n_4, n_5, n_6, n_7 が good と判定され, ノード $n_9, n_{10}, n_{11}, n_{12}$ が bad と判定される.

ここで, ページ状態の変化に関して, 次の定理を定める.

定理 4.1 一度 good あるいは bad と判定されたページの状態は変わらない

定理 4.1 が成り立つとすると, 漸化的にページ判定が行われる際, 一度判定されたリンク隣接状態の各リンクの状態 ($G_{in}, G_{out}, B_{in}, B_{out}$) が, 1 から 0 になることはあり得ない. したがって, unknown または gray 状態から good あるいは bad になることはあっても, 仮定 4.3 により, リンク隣接状態 L_{n_i} は変化することはあってもページ状態 $S_k(n_i)$ が変化することはない.

この定理 4.1 を利用することで, 良いページおよび悪いページの判定条件は次のように書き直すことができる.

定義 4.3 良いページの定義: 初期状態が good であるノードから到達可能なノードはすべて good である

		reachable from												
		n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	n_{11}	n_{12}	n_{13}
reachable to	n_1		0	1	1	1	1	1	0	0	0	0	0	0
	n_2	1		1	1	1	1	1	0	0	0	0	0	0
	n_3	0	0		0	0	0	0	0	0	0	0	0	0
	n_4	1	0	1		1	1	1	0	0	0	0	0	0
	n_5	0	0	0	0		1	1	0	0	0	0	0	0
	n_6	0	0	0	0	0		1	0	0	0	0	0	0
	n_7	0	0	0	0	0	0		0	0	0	0	0	0
	n_8	0	0	0	0	0	0	0		0	0	0	1	1
	n_9	1	0	1	1	1	1	1	1		1	0	1	1
	n_{10}	1	0	1	1	1	1	1	1	0		0	1	1
	n_{11}	1	0	1	1	1	1	1	1	0	0		1	1
	n_{12}	0	0	0	0	0	0	0	1	0	0	0		1
	n_{13}	0	0	0	0	0	0	0	0	0	0	0	0	

図 4.2 到達可能ノードの行列

定義 4.4 悪いページの定義: 初期状態が bad であるノードへ到達可能なすべてのノードは bad である

ここで、到達可能とは、ノード n_i からノード n_j へのリンクをエッジ $e_{i,j}$ とするとき、パス $P = \{n_i e_{i,1} n_2 e_{2,3} \cdots e_{j-1,j} n_j\}$ が存在することを示す。

図 4.2 は、図 4.1 に対する到達可能ノードを示す行列である。図 4.2 において、行に書かれたノード n_i から列に書かれたノード n_j へ到達可能な場合、その値は 1、到達不可能

な場合は 0 となる。

このような行列を到達可能行列と呼び、 A とする。ノード数を u 、すべてのノードの集合を N とし、初期状態が good であるノードに対応する要素の値を 1、それ以外のノードを 0 で示すノード n_i の good 状態を $S_G(n_i)$ で表すとき、初期状態が good であるノードを表す要素数 u のベクトルを $x = [S_G(n_1), S_G(n_2), \dots, S_G(n_u)]$ とする。同様にして、初期状態が bad であるノード n_i の bad 状態を $S_B(n_i)$ で表すとき、初期状態が bad である要素数 u のベクトルを、 $y = [S_B(n_1), S_B(n_2), \dots, S_B(n_u)]$ とする。

このとき、最終的に good と判定されるノードのベクトルを x' とすると、 $x' = xA$ により x' を求めることができる。 x' の各要素のうち、値が 1 以上となっている要素に対応するノードが good と判定されるノードとなる。このノードの集合を $\alpha = \{n_i \in N | S_G(n_i) > 0\}$ とする。

また、最終的に bad と判定されるノードのベクトルを y' とし、 A^T を A の転置行列とすると、 $y' = yA^T$ による y' を求めることができ、 y' の各要素のうち、値が 1 以上となっている要素に対応するノードが bad と判定されるノードとなる。このノードの集合を $\beta = \{n_i \in N | S_B(n_i) > 0\}$ とする。

gray であるページについては、初期状態が bad であるページから到達可能なページのうち、良いページに判定されているページと悪いページと判定されているページを除いた物である。したがって、まず、初期状態が bad であるページから到達可能なノードのベクトルを z としたとき、 $z = yA$ をにより z を求め、 z のノード n_i の値を $S_R(n_i)$ としたとき、到達可能なノードの集合 $\gamma = \{n_i \in N | S_R(n_i) > 1\}$ とする。このとき、 $\delta = \gamma - (\alpha + \beta)$ により得られた結果 δ が gray 状態となるページの集合となる。

グラフ上の各ノードについて、リンクによる隣接関係を示すような行列、すなわち、行となるノード n_i からリンクしているノード n_j がある場合には、行 i 列 j の値を 1 とし、

それ以外の場合は0とするようなグラフを作成する．このような $u \times u$ の行列を隣接行列と呼ぶ．

隣接行列が与えられたならば， A を求める問題は，推移的閉包 (transitive closure) を求める問題と同一であり，多くのアルゴリズムが提案されている．この中には $O(u + v)$ (v はエッジの数) のオーダーで求められるアルゴリズム [78] も示されている．

以上より，本手法の手順は以下のようになる．

[1]: WWW におけるページをノード，リンクを有向エッジとした有向グラフに対する隣接行列 P を作成する

[2]: 隣接行列から transitive closure を求めるアルゴリズムを用いて，到達可能行列 A を作成する

[3]: 初期状態が good であるノードのベクトル x を作成し， $x' = xA$ により，最終的に good となるノードのベクトル x' を得る．さらに，good となるノードの集合 α を得る．

[4]: 初期状態が bad であるノードのベクトル y を作成し， $y' = yA^T$ により，最終的に bad となるノードのベクトル y' を得る．さらに，bad となるノードの集合 β を得る．

[5]: $z = yA$ により初期状態が bad であるノードから到達可能なノードのベクトルを得た上で，bad から到達可能なノードの集合 γ により， $\delta = \gamma - (\alpha + \beta)$ により good でも bad でもないページの集合 δ を得る．

[6] 得られた集合 α, β, γ がそれぞれ good, bad, gray と判定されたノードであり，それ以外のページの unknown と判定される．

4.4 ページ評価手法

4.4.1 良いページ/悪いページの初期集合

本手法では、良いページ/悪いページの初期集合は、評価者が既に評価済みのページの中から、良い/悪いの定義に基づき任意に決定する。4.2 節で述べたように、良い/悪いの定義は評価者の主観に依存するものであり、初期集合の選び方により本手法により得られる結果も異なる。このような初期集合の任意性は、目的に応じた初期集合を指定することで、様々な状況における適用可能性をもたらし、さらに複数の結果を複合的に利用するなどの応用も考えられる。

4.4.2 ページとサイト

ここまで、本章ではリンクをページ間の関係として扱ってきた。このようなリンク関係を実際に WWW に適用することを考えると、ページ作成者単位（組織、部署、個人など）でページ群をサイトとして縮退することができる。例えば、企業のページの場合、企業の持つドメイン内のすべてのページを企業として管理していることが多い。この場合には、その企業のトップページからリンクを辿ることのできる同ドメイン内のページ群からリンクを抽出することで、サイトからの他サイトへのリンク集という形に縮退することができる。個人ページでは、プロバイダ内の～で表される個人の領域内のページ群を個人サイトとして縮退して考えることができる。

ページ作成者の単位でページ群を縮退することで、サイト内の推薦以外のリンク（構造を作るためのリンクなど）を除外できる。同時に、ページ作成者の単位によってページ群をサイトとして縮退し、広大な Web 空間のリンク関係をサイト間関係として扱うことが

できる。

本手法では、ページをドメイン名に応じてサイトとして縮退することで、サイト間のリンク関係としてみることにする。

4.4.3 実験

本手法を確認するため簡単な実験を行った。本来は、本手法は WWW 全体に対して適用することで結果を得るものであるが、大規模な実験が必要であり、現時点では実現できていない。そこで、予備的な実験としていくつかのサイトに関してリンク関係を調べ、実際に内容を確認していくことで、本手法の有効性および問題点について調査を行った。

実験では、良いページの初期集合として国内の自動車メーカーのページ、悪いページの初期集合として性風俗情報を扱ったページを選択した。

各ページのサイトに対し、同ドメイン内のページを探索し、ドメイン外部へのリンクを抽出するプログラムを作成し、リンク先サイトのリストを得た。また、AlltheWeb(<http://www.alltheweb.com>) の被リンクページの検索機能を利用し、各ページのサイトに対してリンクをしているサイトのリストを得た。

得られたすべての URL 集の中から、無作為にいくつかの URL を抜き出し実験用のモデルとした。サイトの内容については、実際にページ内容を閲覧してページに対する判定を行った。良いページの初期集合として 8 つの国内自動車メーカーのサイト、悪いページの初期集合として 4 つの性風俗店情報サイトを選択した。

良いページからのリンク

良いページからのリンクでは数万以上のサイトが得られた。この中から、URL から単純に判別できるサイトや実際に内容を見て確認したサイトにより、サイト間のリンク関係

を概観した。

自動車メーカーにより多少の違いはあるが、別ドメインへのリンクとしては、各メーカーサイトから特定の車種に特化した別ドメインの自社ページ、海外を含む子会社や地域の販売会社を中心となっている。さらに、自社関連以外のページへのリンクとしては、自動車関連の協会や関連省庁、モーターショーやスポンサーをしているイベント用ページ、地域自治体や地域ポータル（トヨタなら愛知、マツダなら広島など）があった。

これらのページからリンク（次のホップ）では、地域販売会社による各地域のポータルサイトや観光案内サイトなど、各地域でないとわからないような店舗や中小企業へのものが多くなる。しかし、これらのリンクから悪いと思われるページやサイトへの直接的なリンクは発見できなかった。

なお、良いページへの被リンクについては、あまりに量が多いため（例えば、www.toyota.co.jp へのリンクは4万以上）検索結果のタイトルのみで確認を行った。結果の上位では関連会社や自動車関連の会社や店舗があるが、結果の下位になるにしたがって、個人の趣味ページや関連がわからないページが多くなり、予想通り多様なページによるリンクが見られた。

悪いページへのリンク

悪いページへリンクしているページは、各サイトに対して数十から数百程度で同一ドメイン内からのリンクも多いため、それほど膨大な量にはならなかった。

悪いページへリンクを行っているのは、圧倒的に同業のポータルサイトと同業者による相互リンクが多い。直接的な相互リンクではなくてもランキングサイトやポータルサイトを介してつながっていることが非常に多い。

他には、一般的な地域ポータルなどのお店紹介などからのリンクや、プロバイダ内の個人ページ内の日記や雑談、掲示板の広告からのリンクなどがあった。このうち、プロバイ

ダ内のページや、アダルト系だけでない一般のサイトも扱うアクセスランキングサイトでは、ドメイン名だけで判断すると一般の店舗や個人ページなどの悪いとはいえないページも含まれてしまうという問題がわかった。この点については後述する。

gray ページ

悪いページからのリンクは、アドビやマクロメディアなどのツールや有名ポータル、検索サイトへのリンクを除くと、内容的に近いものや広告主サイトへのページなどが中心となる。つまり、このようなページからポータルサイトや良いページであるツール系サイトを除くと、悪いページと同類のページが残ることになる。さらに、残ったページから悪いページを除いたとしても、確認した範囲では悪いページに近い内容のページが得られた。

結果

今回の実験では、半分手作業によるデータの取得とページの確認を行ったため、十分な量のデータについては確認することができなかった。

全体の結果としては、良いページからのリンク、悪いページへのリンクそれぞれで確認する限り、良好な結果が得られた。しかし、ポータルサイト、プロバイダ、ランキングサイトが仮定 4.3 に反するような状態になっていることが確認された。この点については 4.5 節で述べる。

この点を除けば、少なくとも 3 ホップ程度のリンク関係においては、良好な結果が得られたと考えている。現在、さらに大規模な実験をするべく準備を進めている。

4.5 考察と議論

本手法は、本章で行った仮定が成り立つ限り矛盾は生じないはずである。しかし、実験によりいくつかの問題や検討すべき項目が判明している。それらについて考察と議論を

行う。

4.5.1 仮定に矛盾するサイト

ポータルサイト

良いページからリンクされ、同時に悪いサイトへもリンクしているような仮定 4.3 に反するサイトとして代表的なものはポータルサイトである。実験でも、Yahoo!や地域ポータルなどのサイトが仮定 4.3 に矛盾することを確認できた。

多くの情報を持つポータルページでは、一般に良いページと判断できる場合であっても悪い内容を含んでいる場合がある。例えば、18 才以上であることの認証を行った上ではあるが、アダルト系のサイトへのリンクを持つものも多い。アクセスランキングを行っているサイトでも同様な例が見られた。また、ポータルサイト自身のコンテンツとして悪いページへのリンクはなくても、掲示板に広告としてアダルト系コンテンツの URL へのリンクが張ってある可能性もある。大きなポータルだけでなく地域のポータルサイトや地域の店舗紹介サイトでも、地域の企業や自治体からのリンクがされることが多いため同様の問題が起きている。

さらに、ほとんどのポータルサイトは検索機能を有しているため、単語を入力することで悪いページへのリンクを生成する（すなわち悪いページとなる）。つまり、ポータルサイトでは、多様なページへのリンクを持つと同時に、多様なページからもリンクされているために、良いページ/悪いページ両方の定義を満たしてしまう。

このようなポータルサイトの特徴は、本章で定める考える良い/悪いとは別に考えるべきものであり、本手法ではポータルサイトは特別な扱いを適用する必要がある。もし、ポータルサイトを良いページや悪いページとしてしまうと、ほとんどのページが到達可能となってしまう、本手法によるページ判定は不可能である。しかしながら、漸化的にページ

判定を行っていく上で、リンク関係からのみポータルページを探すポータル発見アルゴリズムなどが考えることはできる。これについてはさらなるリンク構造に関する研究が必要である。

現状の実験においては、ポータルサイトと判断されるようなページを明示的に指摘することで、それらのサイトを unknown として処理を行い対処している。

プロバイダ

プロバイダにおいても、ドメイン単位でページを判定した場合に仮定 4.3 に矛盾するページが見つかった。

プロバイダの個人ページから悪いページへのリンクが存在することはよくあることで、実験でも確認されている。このようなリンクをドメイン名だけでみると、プロバイダ内の大半を占める善良な一般ページや、そこへリンクしているページがすべて悪いページと判定されてしまう。同時にそのプロバイダを利用している企業がプロバイダのトップページにリンクしたり、プロバイダが広告を出していることがある。この場合、仮定 4.3 に矛盾することとなる。

同じような問題は、大学や研究機関などの公的な情報と同時に個人ページを有するサイトでも発生する。個人ページ内の趣味のページや掲示板にたまたま張られた広告からアダルトサイトにいくことで、ドメイン全体が悪いページに判定されてしまうこともありえる。

このような問題は、4.4.2 節で述べたようにページ作成者の単位とドメインが一致していないことに起因していると考えられる。今回の実験では、個人ページの判別を行わなかったが、作成者単位でのリンク解析の必要性が確認できた。

その他の矛盾するサイト

ポータルやプロバイダ以外のサイトにおいても矛盾が起こると考えられる。信用できる組織が注意の喚起や告発のためにリンクすることや、法人としてはきちんとしたものであっても扱う内容が悪いページに該当するような場合（成人向けを扱う出版社や新興宗教系の団体など）が例として考えられる。

このようなサイトに関しては、良い/悪いの評価が他ページへ波及しないように、本手法のページ判定においては unknown ページとして扱うことが最も適当であると考えられる。しかしながら、unknown ページであってもページに含まれるリンクの数（良いページからのリンク数、悪いページへのリンク数）や良いページ/悪いページの初期集合からのホップ数によって、良い/悪いのどちらにより近いかの判断を行うなどの方法も考えられるだろう。また、このようなサイトは、特殊な性格を持つサイトとして見ることができ、これらのサイトを発見し提示することによって、ページの判断において有効利用できるのではないかと考えられる。

4.5.2 初期ページ指定の扱い

初期ページをどのようにして、どの程度指定するべきも検討する必要がある。

悪いページにおいては、数ホップ内における相互リンクが非常に多く、似たようなサイトを一つ指定するだけで、相互リンクしているページがすべて悪いページに含まれることになる。したがって、相互リンクしているどのページを悪いページの初期として指定しても、すべてが含まれることになるため、少ない初期ページの指定であっても、悪いページの発見に効果があると思われる。しかし、良いページに関しては、実験での例に見るように自動車メーカーから他業種へのリンクは非常に少ない。したがって、各分野についての

良いページを指定する必要がある。

本手法を WWW に適用させる際には、初期ページの集合を充実させることにより、unknown ページを減らしていくことは重要な要因となる。初期ページを充実させるためには、WWW ディレクトリなどの信頼できるサイト集やページフィルタリングに利用されている URL 集などを利用したり、他の検索エンジンの結果を利用した指定方法などが考えられる。

また、初期ページ集合の扱いについても検討が必要である。現在は、評価者が決定した初期の良いページ/悪いページの集合は、変化しないものとしてページの判断を行っている。しかし、矛盾が生じた場合に、良いページ優先、あるいは悪いページ優先、unknown とするなど、初期ページ集合が変化する方法も考えられる。これらの点については、実際のページに適用していく中で適切な方法を決定していく必要があるだろう。初期ページ集合を適切に設定することで、ページに対する到達可能性だけでなく、初期ページからの到達ホップ数を考慮したスコア付けなどの応用も可能であると考えている。

4.6 まとめ

本研究では、Web 上のページに対して、良いページ/悪いページを判定する手法を提案した。今後、より広範な Web への適用や良い/悪いを判断するため、適切な初期集合の選択方法についての検討を行う予定である。さらに、良いページ/悪いページの違いや悪いページへの到達可能性を可視化するための方法や、ページ特性を判断するための本手法の適切な応用分野についても研究を進めていく予定である。

第 5 章

結論

なんらかの手段によりたどり着いたページの内容が正しいのかを判断することは、たとえば内容を読んだ上でも難しいことがある。あるページに対して、実社会における信用を担保とした広大な WWW 空間における位置づけがどのようなものかわかれば、そのページに含まれる情報を判断するために有力な材料となる。たとえば、内容に詳しくない場合でも、そのページの位置づけから内容の確かさなどの信頼性を評価することができる。

本研究は、WWW における信頼性の問題に注目し、WWW における信用に関する議論を行った上で、初期値として与えたいいくつかのページの信用を利用し、WWW の構造からページ信頼性を評価する方法を示した。

第 2 章では、Web における信頼性の基準として、ページ作成者の社会的信用の考え方を導入し、「リンクすることによりリンク先ページに信頼が伝播する」という概念と、リンク構造からページを評価するモデルとアルゴリズムを示した。本モデルにより、社会的信用という考え方が WWW における信頼性を評価に適用可能であり、リンクによるページ評価が信頼性評価として妥当であることを示した。このモデルを利用することで、いくつかのページ（あるいはページ作成者）に対する信用の指定から、リンク構造解析により妥当

な計算量で多数のページの信頼性評価をできる。したがって、本モデルは実際の WWW に対し十分に適用可能であることと、応用の可能性について議論を行った。

第3章では、サイト内における信頼性に注目し、サイト構造と組織構造、サイトの特性について議論を行った。さらに、評価モデルをサイトに対し効果的に適用する手法を示し、サイトにおける公式度として信頼性を提示する実験を行った。

本手法では、作成者の種別による信用の定義とリンク属性によるリンク推薦度の指定を行うことで、サイトとしての公式度を提示することができ、

- サイトにおいては、サイトの信頼性の確保した情報発信
- 各作成者においては、ページの信頼性の自己管理による責任ある情報発信

が可能であることを示した。また、閲覧者に対してはリンクによる責任関係の提示などの応用についての議論を行い、信頼性を確保した情報発信の可能性を示した。

第4章では、社会的信用の概念を拡張し、正の信用と負の信用という概念を導入した。これらの信用を初期値としたリンク構造解析により良いページ、悪いページ、悪い可能性があるページを評価する手法を示した。本手法により、ポータルやプロバイダのページを除くと、「良いページにリンクされるページは良いページ」、「悪いページにリンクするページは悪いページ」というリンク解析により良いページ/悪いページとほぼクラスタリングできることがわかった。良い/悪いというある程度の主観的判断を許容する評価基準を採用することにより、従来のページ評価では得ることができない、様々な側面から見たページの持つ特徴を二元的に判定することができる。

本研究では、ページの信頼性を評価する方法について示したが、目標とするところは、現在インターネットにおいて行われている一方的な規制と無秩序な情報発信ではなく、一定の規律を持った上での自由な情報発信を行うことである。

現実社会における信頼関係は単純に得られるものではない [1] が、社会的に定義可能な信頼関係を WWW の世界に適用することで、WWW における信頼を評価することは可能である。現実の関係を元にして得られた WWW 上の信頼は、現在問題となっている [79] や [80] で述べられているようなインターネット上の規制と自由の問題や [81] や [82] で議論されているインターネットと法律、さらには情報倫理 [83] などの諸問題を解決する上で一つの手法となるものと考えている。

今後は、さらに Web における信頼性の問題を解決するためのモデルと適用手法について研究を行う。信用の semantic に基づくモデルの検討や Semantic Web において行われているような、RDF を利用した信頼性の指定などの、さらに汎用的な手法についても検討していく予定である。

謝辞

本研究の機会を賜り，神戸大学自然科学研究科情報メディア科学専攻においてご指導を頂きました神戸大学大学院自然科学研究科 北村新三教授 に深く感謝致します。

本研究をまとめるにあたり研究の遂行および論文の作成において，神戸大学自然科学研究科情報メディア科学専攻博士課程後期に在学中の4年間にわたり直接御指導を賜り，多大な御助言と御教授を賜った神戸大学国際文化学部 大月一弘教授 に心より感謝致します。

本研究の遂行にあたり，御指導と有益な御助言を賜り，論文作成に多大なご協力を頂いた神戸大学国際文化学部 清光英成助教授，森下淳也教授に感謝致します。

機会ある毎に御助言と有益な御示唆を賜った神戸大学国際文化学部 鍋木誠教授，退官なされた横尾能範教授に感謝致します。

本研究における第2章，第3章の評価システムの実装を担当して頂いた，神戸大学大学院人間総合科学研究科を修了し株式会社東洋情報システムに勤務されている堤智也氏，議論に参加して頂き論文作成にも御協力頂いた，神戸大学大学院人間総合科学研究科を修了し株式会社ブレーンセンターに勤務されている絹川達也氏に感謝致します。

また，研究において御議論を頂いた神戸大学国際文化学部情報論講座の皆様に感謝致します。

本研究に御理解と御支援をして頂いた岐阜県立国際情報科学芸術アカデミーおよび情報科学芸術大学院大学の皆様，特に坂根巖夫前学長，横山正学長，DSP コースの皆様に感謝

致します。

最後に日頃から研究生生活を支えて頂いた家族と友人に感謝致します。

参考文献

- [1] Luhmann, N.: 信頼, 勁草書房 (1990), 大庭健, 正村俊之 訳.
- [2] Corbitt, B. J., Thanasankit, T. and Yi, H.: Trust on the World Wide Web: A Study of Consumer Perceptions, in *Proc. of the 6th Asia Conference on Information Systems* (2002).
- [3] Gefen, D.: Reflections on the dimensions of trust and trustworthiness among online consumers, in *Proc. of Database for Advances in Information Systems*(33:3), pp. 38–53 (2002).
- [4] ベリサイン: <http://www.verisign.co.jp/>.
- [5] MSN: <http://www.msn.com/>.
- [6] Yahoo!: <http://www.yahoo.co.jp/>.
- [7] ICRA: <http://www.icra.org/>.
- [8] World-Wide Web Consortium: <http://www.w3c.org/>.
- [9] 財団法人インターネット協会: フィルタリング情報ページ:
<http://www.iajapan.org/rating/>.
- [10] Semantic Web: <http://www.w3c.org/2001/sw/>.
- [11] 特集 セマンティック Web, 情報処理, Vol. 43, No. 7, pp. 707–750 (2002).
- [12] Resource Description Framework(RDF): <http://www.w3c.org/RDF/>.

- [13] Web-Ontology: <http://www.w3c.org/2001/sw/WebOnt/i>.
- [14] Dhyani, D., Ng, W. K., and Bhowmick, S. S.: A Survey of Web Metrics, *ACM Computing Surveys*, Vol. 43, No. 4, pp. 469–503 (2002).
- [15] 佐藤光弘, 梶浦正浩, 多田智之, 池野篤司, 森口修: WWWにおける情報検索技術の動向, *電子情報通信学会誌*, Vol. 82, No. 12, pp. 1237–1242 (1999).
- [16] AltaVista: <http://www.altavista.com/>.
- [17] 村田剛志: Web コミュニティ, *情報処理*, Vol. 44, No. 7 (2003).
- [18] Bharat, K. and Henzinger, M.: Improved algorithms for topic distillation in a hyper-linked environment, in *Proc. of the 21st ACM SIGIR Conference on Research and Developments in Information Retrieval*, pp. 104–111 (1998).
- [19] Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., and Rajagopalan, S.: Automatic resource compilation by analyzing hyperlink structure and associated text, in *Proc. of the 7th International World Wide Web Conference*, pp. 65–74 (1998).
- [20] Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Mining the link structure of the World Wide Web, *IEEE Computer*, Vol. 8, No. 32, pp. 60–67 (1999).
- [21] Chakrabarti, S., Berg, van den M. and Dom, B. E.: Focused crawling: a new approach to topic-specific web resource discovery, in *Proc. of the 8th International World Wide Web Conference* (1999).
- [22] Chakrabarti, S., Joshi, M. and Tawde, V.: Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks, in *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 208–216 (2001).
- [23] Terveen, L. and Hill, W.: Finding and visualizing inter-site clan graphs, in *Proc. of ACM*

- CHI98*, ACM Press (1998).
- [24] Dean, J. and Henzinger, M. R.: Finding Related Pages in the World Wide Web, in *Proc. of the 9th International World Wide Web Conference* (1999).
- [25] 豊田正史：WWWにおける関連コミュニティ群の発見, 情報処理学会研究報告 DBS-122, pp. 307–314 (2000).
- [26] Kleinberg, J. M.: Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999).
- [27] Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: The Web as a Graph: Measurements, Models, and Methods, in *Proc. of International Conference on Combinatorics and Computing*, pp. 1–17 (1999).
- [28] Kleinberg, J. M.: Hubs, Authorities, and Communities, *ACM Computing Surveys*, Vol. 31, No. 4es (1999).
- [29] Bharat, K. and Henzinger, M. R.: Improved Algorithms for Topic Distillation in a Hyperlinked Environment, in *Proc. of the 21st ACM SIGIR*, pp. 104–111 (1998).
- [30] Gibson, D., Kleinberg, J. and Raghavan, P.: Inferring Web Communities from Link Topology, in *Proc. 9th ACM Conference on Hypertext and Hypermedia*, pp. 225–234 (1998).
- [31] Flake, G. W., Lawrence, S. and Giles, C. L.: Efficient Identification of Web Communities, in *Proc. of ACM 6th International Conference on Knowledge Discovery and Data Mining*, pp. 150–160 (2000).
- [32] 村田剛志：参照の共起性に基づく Web コミュニティの発見, 人工知能学会論文誌, Vol. 16, No. 3, pp. 316–323 (2001).
- [33] Flake, G. W., Lawrence, S., Giles, C. L. and Coetzee, F. M.: Self-Organization and

- Identification of Web Communities, *IEEE Computer*, pp. 66–71 (2002).
- [34] 野村早恵子, 小山聡, 早水哲雄, 石田亨: WEB コミュニティ発見のための HITS アルゴリズムの分析と改善, 電子情報通信学会論文誌, Vol. J85-D-I, No. 8, pp. 741–750 (2002).
- [35] 河瀬基公子, 川原稔, 岩下武史, 河野浩之, 金澤正憲: Web コミュニティ発見のための大規模 Web グラフに対するデータ圧縮方, データベースと Web 情報システムに関するシンポジウム論文集, pp. 423–430 (2002).
- [36] 今藤紀子, 喜連川優: 可変辺容量に基づく最大フローアルゴリズムを利用したウェブコミュニティの抽出, データベースと Web 情報システムに関するシンポジウム論文集, pp. 391–397 (2002).
- [37] Haveliwala, T. H., Gionis, A., Klein, D. and Indyk, P.: Evaluating Strategies for Similarity Search on the Web, in *Proc. of the 11th International World Wide Web Conference*, pp. 432–442 (2002).
- [38] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J.: Graph structure in the Web, in *Proc. of the 9th International World Wide Web Conference*, pp. 309–320 (2000).
- [39] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tompkins, A. and Upfal, E.: The Web as a graph, in *Proc. of the 9th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database system*, pp. 1–10 (2000).
- [40] D.Davison, B.: Topical locality in the Web, in *Proc. of the 23rd annual international ACM SIGIR Conference*, pp. 272–279 (2000).
- [41] D.Davison, B.: Recognizing Nepotistic Links on the Web, in *Artificial Intelligence for Web Search, Technical Report WS-00-01*, pp. 23–28, AAAI Press (2000).

- [42] 原田昌紀, 風間一洋, 佐藤進也: 参照共起分析の Web ディレクトリへの適用, 情報処理学会研究報告 IF-61, pp. 45–52 (2001).
- [43] Tomlin, J. A.: A New Paradigm for Ranking Pages on the World Wide Web, in *Proc. of the 12th International World Wide Web Conference*, pp. 350–366 (2003).
- [44] Chuang, T.-R., Low, B. C. and Wu, in C.: The Two Webs: Towards a Web-based System for Social Network Analysis, in *Proc. of 3rd Workshop on Information Technologies and Social Transformations*, Vol. 1, pp. 75–85 (1999).
- [45] Adamic, L. A. and Adar, E.: Friends and Neighbors on the Web (2000), <http://www.hpl.hp.com/shl/papers/fnn/>.
- [46] Amento, B., Terveen, L. and Hill, W.: Experiments in Social Data Mining: The TopicShop System, *ACM Transactions on Computer-Human Interaction*, Vol. 10, No. 1, pp. 54–85 (2003).
- [47] Lempel, R. and Moran, S.: SALSA: the stochastic approach for link-structure analysis, *ACM Transactions on Information Systems*, Vol. 19, No. 2, pp. 131–160 (2001).
- [48] Cohn, D. and Chang, H.: Learning to Probabilistically Identify Authoritative Documents, in *Proc. of 17th International Conference on Machine Learning*, pp. 167–174 (2000).
- [49] Cohn, D. and Hofmann, T.: The missing link: a probabilistic model of document content and hypertext connectivity, *Neural Information Processing Systems*, Vol. 13, pp. 430–436 (2001).
- [50] Google:<http://www.google.com/>.
- [51] Brin, L., S. and Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, in *Proc. of the 7th International World Wide Web Conference* (1998),

- <http://dbpubs.stanford.edu/pub/1998-8>.
- [52] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford University (1999), Digital Library <http://dbpubs.stanford.edu:8090/pub/1999-66>.
- [53] Haveliwala, H.: Efficient Computation of PageRank, Technical report, Stanford University (1999), <http://dbpubs.stanford.edu/pub/1999-31>.
- [54] Haveliwala, T. H.: Topic-sensitive PageRank, in *Proc. of the 11th International World Wide Web Conference*, pp. 517–526 (2002).
- [55] Richardson, M. and Domingos, P.: The intelligent surfer: Probabilistic combination of link and content information in PageRank, in *Proc. of Advances in Neural Information Processing Systems 14*, pp. 1441–1448 (2002).
- [56] Amento, B., Terveen, L. and Hill, W.: Does authority mean quality? predicting expert quality ratings of Web documents, in *Proc. of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 296–303 (2000).
- [57] Diligenti, M., Gori, M. and Maggini, M.: Web Page Scoring Systems for Horizontal and Vertical Search, in *Proc. of the 11th International World Wide Web Conference*, pp. 508–516 (2002).
- [58] Farahat, A., Nunberg, G. and Chen, F.: AuGEAS (AAuthoritativeness Grading, Estimation, and Sorting), in *Proc. of the Eleventh International Conference on Information and Knowledge Management*, pp. 194–202 (2002).
- [59] Pável Calado, B. R.-N., Ziviani, N. and Edleno Moura, I. S.: Local Versus Global Link Information in the Web, *ACM Transactions on Information Systems*, Vol. 21, No. 1, pp. 42–63 (2003).

- [60] Tsoi, A. C., Morini, G., Scarselli, F., Hagenbuchner, M. and Maggini, M.: Adaptive Ranking of Web Pages, in *Proc. of the 12th International World Wide Web Conference*, pp. 356–365 (2003).
- [61] Hirabayashi, M., Ohtsuki, K., Kiyomitsu, H., Morishita, J., and Kitamura, S.: Scoring Web Pages with Social Credibility in a Hypertext Environment, in *Proc. of 5th International Conference on Enterprise Information Systems*, pp. 326–333 (2003).
- [62] 平林真実, 大月一弘, 清光英成, 森下淳也, 北村新三: リンクに基づいた明示的 Web ページ評価法, 情報処理学会論文誌: データベース No.SIG12(TOD16), Vol. 43, pp. 92–102 (2002).
- [63] 堤智也, 平林真実, 大月一弘: 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, 情報処理学会インタラクシオン 2000 論文集, pp. 65–66 (2000).
- [64] 絹川達也, 平林真実, 清光英成, 大月一弘, 北村新三: 作成者責任の概念を用いた Web ページ評価法のサイトへの適用実験, 情報処理学会研究報告 DBS-125, pp. 281–288 (2001).
- [65] 平林真実, 堤智也, 絹川達也, 大月一弘, 北村新三: 大規模サイトにおける WWW ページ評価手法の検討, 情報処理学会情報学シンポジウム講演論文集, pp. 75–82 (2001).
- [66] 平林真実, 大月一弘, 清光英成, 北村新三: WWW におけるハイパーリンクの構造解析を用いた良いページと悪いページの判定方法の検討, 神戸大学自然科学研究科紀要 (2004), 掲載予定.
- [67] 社団法人マルチメディア・タイトル制作者連盟 (AMD): FLMASK 事件に見るインターネットと公序良俗, ネットワーク・コンテンツ白書'98 (1998), <http://www.amd.or.jp/ncwp98/index.html>.
- [68] Marchiori, M.: The quest for correct information on the web: Hyper search engines, in

- Proc. of 6th International World Wide Web Conference* (1997).
- [69] Stewart, K. J.: Transference As A Means Of Building Trust In World Wide Web Sites, in *Proc. of the 20th international conference on Information Systems*, pp. 459–464 (1999).
- [70] DMOZ: <http://www.dmoz.org/>.
- [71] 廣川左千男, 池田大輔, 田口剛史 : リンク情報からの知識網構成, bit 別冊発見科学とデータマイニング, pp. 272–281, 共立出版 (2000).
- [72] Kahn, P. and Lenk, K.: *mapping web sites*, RotoVison (2001).
- [73] Dodge, M. and Kitchin, R.: *Atlas of Cyberspace*, Addison-Wesley (2001).
- [74] 原田昌紀, 佐藤進也, 風間一洋 : WWWページ間の階層構造の推定と検索システムへの応用, 情報処理学会研究報告 DBS-118, pp. 105–112 (1999).
- [75] Xue, G.-R., Zeng, H.-J., Ma, Z. C. W.-Y., Zhang, H.-J. and Lu, C.-J.: Implicit link analysis for small web search, in *Proc. of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 56–63 (2003).
- [76] Eirinak, M. and Vazirgiannis, M.: Web Mining for Web Personalization, *ACM Transactions on Internet Technology*, Vol. 3, No. 1, pp. 1–27 (2003).
- [77] Namazu プロジェクト : 日本語全文検索システム namazu : <http://www.namazu.org>.
- [78] C.P.Schnorr, : An Algorithm for Transitive Closure with Liner Expected Time, in *SIAM J. Computing*, Vol. 7, pp. 127–133 (1978).
- [79] Lessig, L.: *CODE*, 翔泳社 (2001), 山形浩生 , 柏木亮二訳.
- [80] Lessig, L.: *コモンズ*, 翔泳社 (2002), 山形浩生訳.
- [81] 情報ネットワーク法学会 (編): 情報ネットワーク・ローレビュー, 第 1 巻, 情報ネットワーク法学会 (2003).

[82] 情報ネットワーク法学会（編）：情報ネットワーク・ローレビュー，第2巻，情報ネットワーク法学会（2003）。

[83] 水谷雅彦，土屋俊，越智貢（編）：情報倫理学，ナカニシヤ出版（2000）。