ハイパーリンクの参照重要度に基づくページ品質の評価

山本　祐輔† 　手塚　太郎†† 　田中　克己††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 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
E-mail: †yamamoto@dl.kuis.kyoto-u.ac.jp, ††{tezuka,ktanaka}@i.kyoto-u.ac.jp

あらまし　現在の Web では爆発的に情報が増加しており、Google などに代表される検索エンジンを用いてもユーザが欲しい情報を獲得するのは難しい状況にある。また SEO と呼ばれる検索エンジン最適化手法を用いて、故意にページランクを上げようとする Web コンテンツ作成者の出現により、検索エンジンのランキング上位に現れているページでもコンテンツの信頼性の観点からすると上位に値しない可能性が起こっている。そこで本稿では、ページランクの計算に用いられるリンクに着目し、コンテンツ作成者がどのような意図で他の Web ページにリンクを張ったのかを解析し、それらに評価値を与える。そしてその評価値を用いてページの品質評価を行う。

キーワード　情報検索、信頼性、PageRank

Quality Evaluation of Web Pages by Referential Importance of Hyperlinks

Yusuke YAMAMOTO†, Taro TEZUKA††, and Katsumi TANAKA††

† School of Informatics, Kyoto University　Yosidahonmati, Sakyou-ku, Kyoto,606-8501 Japan
†† Department of Social Informatics,Graduate School of Informatics,Kyoto University
Yosidahonmati,Sakyou-ku,Kyoto,606-8501 Japan
E-mail: †yamamoto@dl.kuis.kyoto-u.ac.jp, ††{tezuka,ktanaka}@i.kyoto-u.ac.jp

Abstract　Todays, Information increases explosively in the Web, and even if the search engine, for example, Google etc. is used, it is difficulet for the users to acquire information. The possibility of not worth the high rank judging from the viewpoint of the reliability of contents has happened to the page that appears to the high rank of the ranking of the search engine by using the search engine optimization technique that is called SEO because of the appearance of the Web content authors who try to raise PageRank by intention. Then, we focus on the link used to calculate PageRank, whether by what intention they put the link on other Web pages is analyzed, and we give the evaluation value to them in this text. And, we calculate page qualities judging an unjustified link and the link without the meanings from the link value.

Key words　Information Retrieval、Trust、PageRank

1. はじめに

近年、インターネットの普及により多くの人が自由に様々な情報を発信、収集、閲覧できるようになった。誰でも自由に情報を発信できることもあって、Web 上の情報は爆発的に増加している。そのような大量の情報の中から効率良くユーザの欲しい情報を収集するのが Google [1] などの検索エンジンである。ユーザは検索エンジンに自分の欲しい情報に関するキーワードをクエリとして与え、キーワードを受け取った検索エンジンはそれらに関連するページを Web から収集、ランキングし、それらをユーザに検索結果として返す。

検索エンジンの多くは Web 情報の特徴的な構造であるハイパーリンクの構造を用いてランキングしている。代表的な検索エンジンである Google はリンクを一種の投票と見なし、「多くの良質なページからリンクされているページは、やはり良質なページである」という再帰的な関係を用いて全てのページの重要度を求めている。

このようにリンクを投票と考えてページを評価するランキングアルゴリズムはインターネット普及し始めた時期には非常に効果を発揮したが、近年の Web 情報の爆発的な増加によって、大量の Web ページの中から良質なページを投票対象として選びきれなくなったため、その結果、必ずしも良質なページ

がランキング上位に現れるとは限らなくなっている。また近年、SEO [2] と呼ばれる検索エンジンのアルゴリズムを逆手にとってランキングを不当に向上させるという技術も出現している。そもそもリンク構造を用いたランキングアルゴリズムは「たくさんのリンクを集めたページが価値の高いページになる」ことから「人気があるサイト」がランキング上位に来ても「信頼できるサイト」がランキング上位に来るとは限らない。現在のリンク解析に基づくランキングの問題点として様々な理由が考えられるが、その 1 つとしてリンク自体の意味は考慮されていないことが挙げられる。上で例に挙げた Google の採用するアルゴリズムでは、リンクを張るという行為は「リンク先のページを良質だと認めた」ということが前提であり、あるページの評価値はリンク元のページの評価値とリンク数に依存しているにすぎない。リンクを張る意図としては、他のページを良質だと評価するため以外にも、関連サイトを紹介するもの、友人のサイトを紹介するもの、引用目的のもの、批判対象を指定するもの、など様々なものが考えられる。また仮に他のページを評価しているリンクでも、リンク元の本文と関係が深い場合と薄い場合にはリンクの重要度は異なる。このようなに、実際の評価にはリンク元ページの評価値だけでなく、リンク自体の重要度も考慮しなければならないと考えれる。別の問題としては、リンクは一種の投票であり、投票対象とならなければ評価されることがないページも多く存在する。リンクを投票と捉えて Web ページを評価するには限界がある。

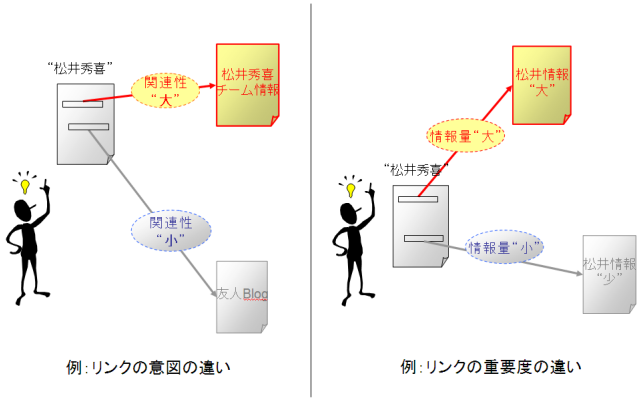

図 1　リンクの意図、重要度の違い

また別の問題点として、現在のリンク解析に基づくランキングアルゴリズムは、あるページの評価を、どれくらい良質なページからリンクを「張られている」か、つまりオーソリティ的な観点からしか行っていない。ページの評価を投票形式に基づく考えに従えば妥当な方法であるが、上でも述べたように、Web ページが肥大化していることで、投票形式では適切な評価が行えず、人気のないページでも良いコンテンツを作成しようとしているコンテンツ作成者の努力が全く考慮されていないことが挙げられる。

そこで本研究では、コンテンツ作成者がどのような意図で他のページへリンクを張ったのかを分析し、リンク元のコンテンツに対してリンク先のページがどれくらい価値があるのかを評価することで、リンクの参照重要度を求める。そして、それらを用いることで、コンテンツ作成者が自コンテンツをどれだけ充実させているかという従来のリンクを投票と捉える評価方法とは異なる観点で Web ページの品質評価を行う手法を考える。

2. 関 連 研 究

2.1 PageRank

Page [3] らは今日 Google などの代表的な検索エンジンに用いられている PageRank を提案した。PageRank の基本的なアルゴリズムでは、「多くの良質なページからリンクされているページは、やはり良質なページである」という再帰的な関係のもとに、全てのページの重要度を判定する。リンクを支持投票と見なして、より重要度の高いページによって投じられた票（リンク）が多ければ多いほどページとしての価値が高くなる。したがって単純に入力リンク数が多いだけではページとしての評価は決定されず、入力リンクの価値も重要な要素となる。

あるページの重要度が、入力リンクの重みの和で定義されることは「良質なページにたくさんリンクされているページは良質である」という考えを反映している。またリンクの重みを求める方法として、リンクを出しているページの重要度を出力リンク数で割っている。これは厳選されたリンクほど良いリンクである、という判断を行っている。

しかし、良質なページが張るリンクが常に意味があるとは限らない。例えば、java に関して有名なページが java に関する他のページにリンクを張っているならばそのページは（java に関して）価値のあるサイトだと言えるが、関係のない音楽のサイトにリンクを張っている場合、そのリンクが価値のあるものであるとは言いがたい。

2.2 HITS

Kleinberg [4] らは、他の Web ページからの評価の高さ (オーソリティスコア) と、評価の高い Web ページへの参照度合い (ハブスコア) から、重要性の高い Web ページを抽出するアルゴリズムである HITS を提案した。

個々の Web ページは「リンクを張られることによって評価をされる」という側面と、「リンクを張ることによって他のページを評価する」という 2 つの側面を持っている。あるテーマに関して重要なページから多くリンクを張られているページはそのテーマに関して重要であると考えられる。また、あるテーマに関して重要なページにたくさんリンクを張るページは、あるテーマのページを紹介するページとしては重要であると考えられる。この考えに基づき、HITS ではあるキーワードを含むページ集合を取得して、その中で重要なハブとオーソリティを発見する。

HITS ではキーワードを含むページを抽出することで、あるコミュニティ内でのページの評価を求められる。PageRank と異なりオーソリティスコアだけでなく、ハブスコアも合わせて計算している点が特徴的である。しかし、HITS も既に述べた PageRank と Teoma と同じく、ページの評価をオーソリティ的な観点からしか評価をしていない。またオーソリティの評価

をハブの重要度から求めてるものの、やはり個々のリンク自体の意味、重要度は考慮されておらず、良質なハブからでたリンクは重要であると決められてしまっている。

また PageRank、HITS ともにリンクを投票と捉える考え方であるため、ランキング上位にあるページのみが投票対象になってしまい、「人気は無いが良質であるページ」の発見ができない問題が生じている。

2.3 Combating Web Spam with TrustRank

Gyongyi [5] らは Web スパム対策として TrustRank アルゴリズムを提案した。Web スパムとは検索エンジンのランキングアルゴリズムを最適化することで故意にランキングを向上させる目的の Web ページである。Web スパムの発見は機械的には行われておらず、専門家が Web スパムかどうかを判定することによってスパムを排除していた。手動でスパム排除を行う作業はコストが高い。そこで TrustRank アルゴリズムでは、できるだけ Web スパムの可能性のあるページをランキング計算対象から除いて、ランキングを行い、PageRank を修正した TrustRank を求めることで検索エンジンのランキング結果の信頼性を向上させる。

手順としてはまず専門家が Web スパムでないページの中から、信頼度の高いページをいくつか選択し、それらをシード（種）として定義する。次にシードからリンクを辿って、他の良質なページを見つけていく。この方法は、シードからある範囲で辿られるページは良質であり、Web スパムの可能性が低いという仮定に基づいている。

人為的に良質だと判定されたページを用いてページの評価を行うので、純粋に機械的に計算した評価値よりも信頼できると考えられる。ページが信頼できるものかどうかは、やはり人間の目で見なければ判断できないので、できるだけ人間の評価をランキングの計算に加味しようとする点は評価できるが、評価値はリンク元の評価値にのみ依存してしまっている。リンクの意図、リンクの重要度を考慮すれば、信頼値を高めることができ、また Web スパムの発見にも繋がることが考えられる。

3. Web ページの評価軸

インターネットの情報を閲覧する際に、閲覧している文章が信頼できるかどうかは極めて重要である。得られた情報に信頼性がなければユーザは誤った情報を取り込んでしまう。インターネットから情報を得る場合、誰でも容易に情報を取得することでき、また誰でも情報を発信することができる。誰でも容易に情報を取得できるというインターネットのは魅力的なものであるが、急速にインターネットが普及したため情報の信頼性の確保の対策が間に合っていないこと、あまりに容易に情報を取得できることから、ユーザの多くはインターネット情報が信頼できるかどうかを意識していることは少ない。これら状況から、Web 情報の信頼性を量ることは極めて重要なことであると考えられる。

そこで本章では Web 情報の信頼性を量るための評価基準について考察する。

3.1 3 つの評価軸

信頼性と言っても様々な基準があると考えられる。実世界での信頼性を例に考える。「例えば○○先生は△△の分野の権威であるから、○○先生が書かれたその分野の情報は信頼できる」と言ったように情報発信元がどの程度発信情報に詳しいかという基準がある。また書籍や商品の人気投票のように、多くの人から良質だと評価されたものが信頼性が高いと評価する基準もある。少し見方を変えた場合、出来るだけ客観的な意見を述べているか、という基準もある。自民党支持者が書いた記事は他の政党支持者から見た場合、支持できない場合があるように、思い込みや主観ができるだけ排除された情報であるか、ということも基準の 1 つになりうる。他にも、ある作品を作るのにどれだけ費用をかけたか、時間を費やしたかのように努力度も考慮できる。

このような例を分析すると信頼性を評価するには大きくは以下の 3 つの軸があることが分かる。

まずある対象が世間一般的にどのように受け入れられているかという社会的重要度が考えられる。Web の世界では社会的受容度の評価軸でコンテンツの評価を行っている場合が多い。既存の検索エンジンのランキングアルゴリズムはリンク解析に基づくものであり、リンクを一種の支持投票と見なして他のページを評価する。検索エンジン以外にも社会的受容度を用いた例として、Amazon に代表されるインターネットショッピングのレビューがある。Amazon では各商品に対して、ユーザがレビューを書き、点数を与えることで商品の評価が決まる。ユーザが自由に評価ができ、肯定的にも否定的にも評価をつけることができる。

2 つ目の評価軸として情報の客観性が挙げられる。社会的受容度はいわゆる多数決の論理で評価が決まってしまうが、ある対象を評価する場合、ある立場から良いと判断できるが、別の立場から評価すると悪いと判断されてしまう場合がある。具体的な視点としては、バイアスがかかっていない情報か、majority であるか minority であるか、などが考えられる。

3 つ目が努力度である。努力度の点で Web ページを評価する場合、まずはユーザ側からの評価とコンテンツ作成者側からの評価に分けられる。

ユーザ側の努力度を考えた場合、先に挙げたコンテキスト依存型ブックマークのように、あるページを見つけるのにどれくらい時間をかけ、他のページとどの程度比べて選んだかが考えられる。また同じユーザがあるサイトをどの程度訪れたか、すなわちあるページへの執着度なども考えられる。コンテンツ作成者側の努力度を考えた場合、Adam [6] らが提案したようにページの更新頻度を評価することが考えられる。他にもユーザの反応を良くするためにデザインにこだわっているか、使いやすいページを作ることを心がけているか、などが挙げられる。

3.2 本研究の位置づけ

ユーザは情報を得るためにインターネットにアクセスするが、たいていの場合、検索エンジンがユーザと情報をつなぐ架け橋になっている。簡単に制限なく情報を取得できるのがインターネットの利点であり、情報を求める、特に何かの事実を調査し

ているユーザにとっては効果的な情報がもっとも望ましい情報となる。従って、ユーザを満足させるような情報をもつページを評価することが重要となる。

その場合、ページの更新頻度や見た目、使いやすさのようにコンテンツとは関係のない要素でページを評価するのは妥当でない。またユーザが客観的な情報をいつも求めているとは限らない。インターネットにアクセスするユーザは多種多様であり、バイアスのかかった情報を求めるユーザの存在も否めない。これらを考慮することも重要ではあるが、多種多様なユーザのニーズに応えるためには十分ではない。

ユーザ自身が評価能力を持っていれば提示された情報を評価することが可能である。しかし、インターネットの検索の場合、ユーザのバックグラウンドは多種多様であり、検索対象となるページが膨大であることから、システム側でページを評価して、有用なページをユーザに提示する必要がある。

現実的に信頼性というものを考えた場合、「良質である、または専門的である人間」が「良質である」と評価したものが高い信頼値を得る。既存のリンク解析に基づいた Web ページの評価アルゴリズムはオーソリティとしてどの程度優れているか、ハブとしてどの程度優れているかについては考慮しているが、「どの程度良質である」という評価、すなわちリンクの重要度は考慮に入れていない。このような方法ではオーソリティ的な、またはハブ的な評価も適切に行えない。逆にあるページから出るリンクの重要度が評価できれば、その評価値を用いることで他のページをより厳密な観点から評価できる。また自コンテンツにとって意味のあるリンクをたくさん持つページはコンテンツを充実させている良質なページと評価することも可能となる。

そこで本研究では、コンテンツの充実させるリンクの参照重要度を求める手法を提案し、その応用例としてハブ的な観点から Web ページを評価する手法を考える。

4. リンクの参照重要度の分析

4.1 リンクの重要度を考慮する意味

Web ページのコンテンツはページ内のコンテンツと、そこから張られたリンクが指すページのコンテンツによって決まる。コンテンツ作成者は基本的に自分のコンテンツにある主張を書き込むが、何らかの理由で外のページにリンクを張ることがある。リンクの張る意図は時と場合によって異なる。代表的なリンクとしては、情報源として参照するリンク、批判対象へのリンク、他のサイトの紹介のリンク、仲間のサイトへのリンク、リンク集、広告リンクなどが挙げられる。

このように様々なリンクが考えられるが、コンテンツ作成者にとっても閲覧ユーザにとっても効果的なリンクとそうでないリンクに分けられる。例えば、情報源として参照するリンクは、コンテンツ作成者側にとっては自分のコンテンツを補完し、閲覧ユーザにとっては理解の促進につながる。しかし、仲間サイトへのリンクは、コンテンツ作成者にとっては友人のサイトを紹介するだけで、自分のコンテンツの内容的補完にならず、またある情報を探しているユーザがこのサイトに訪れた場合、仲間のサイトへのリンクは重要でない。

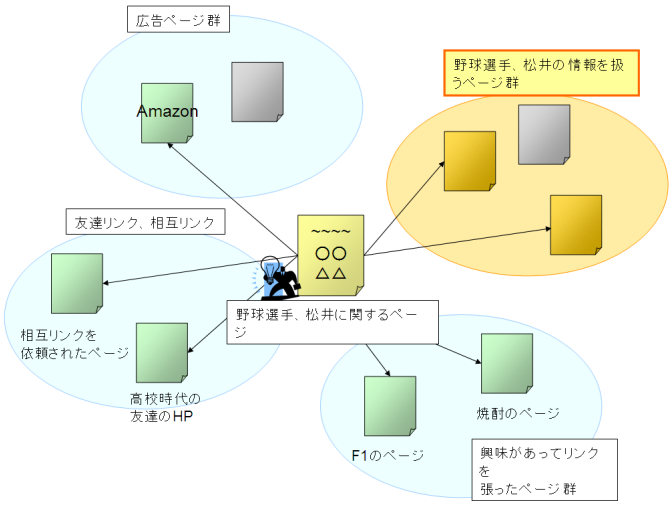

図 2　様々なリンクとその重要度

このような事情から、リンクの価値を等価に扱うべきではなく、リンクの使われ方によってリンクの価値に差をつけなければならない。

4.2 良質なコンテンツとは

一般的にコンテンツ作成者はあるテーマを持ってコンテンツ作成に取り組む。テーマ無しに作ったコンテンツは書かれている内容が発散し、情報の密度も小さいため、ユーザにとっても非常に分かりにくいものであり、コンテンツとしての評価も非常に低くなる。

既存の Web ページ評価の手法ではリンクによる投票によってコンテンツの評価が行われてきたが、どのようなコンテンツがユーザにとって満足できるものかは、あるテーマに従って作成されたコンテンツが内容的に充実しているかによって決まる。

コンテンツ作成者はこのようなユーザの要求を満足させるためにも、コンテンツを作成する場合には

「テーマに基づいて内容を充実させることによってのみコンテンツは良質になり、良質なページのみがユーザに認められる」

ことを意識してコンテンツを作成しなければならない。

リンクを「投票」ではなくコンテンツを充実させるための「道具」として捉えると、コンテンツを充実させることに寄与しないリンクは価値のないリンクとなる。

4.3 リンクの参照重要度

リンク先のコンテンツを用いて自コンテンツを補完する時、自分の扱うテーマと関連のあるサイトを参照する場合と、自コンテンツの内容の一部を補完する目的で他のサイトを参照する場合がある。前者の例としては、Google の新卒採用に関する内容を語るコンテンツから、同じように Google の新卒採用について語るページを参照しているケースが挙げられる。後者の例としては、話題の展開する上で必要な事実、用語を説明するために外部にリンクを張るケースがある。

リンク先のコンテンツがリンク元コンテンツにとって内容的な補完になっているかどうかは扱っているテーマが似ているか

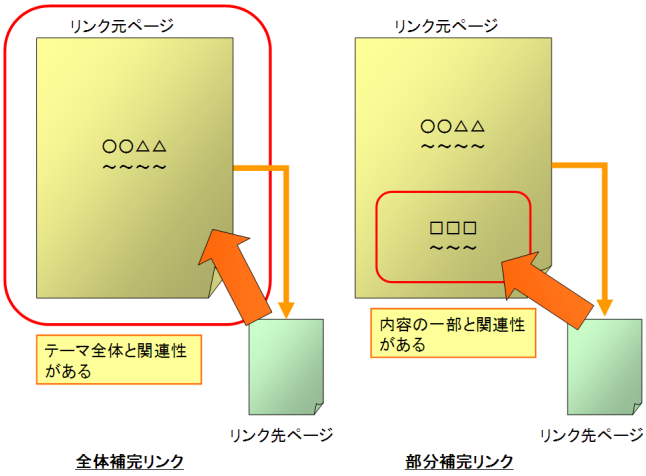

図 3　関連補完と引用補完の違い

どうかで判断できる。従って、リンク元とリンク先のコンテンツの類似度を求めることで内容的補完度が量ることが考えられる。文書全体を特徴ベクトルで表し、特徴ベクトル同士の類似度を計算することで扱うテーマの関連性を量ることができる。

しかし、この方法ではテーマ同士の類似度を計算するためテーマ全体を補完するようなリンクの重要度は高くなるが、部分的に情報を引用したいような目的で用いたリンクの重要度は低くなってしまう。コンテンツの一部を補完するようなリンクもリンク先コンテンツにとっては重要であるので、部分補完を行うリンクの重要度を求めるためには類似度のみを扱う方法は適さない。

また類似度が低いからコンテンツを補完していないわけではない場合もある。リンク先の内容を経て自分の主張を展開する場合、特徴ベクトルから求める類似度は低くなる。コンテンツの類似度ではなくテーマとの関連性を量る必要がある。

ここで、類似度とは別の尺度としてリンク先のページがリンク元に与えた影響度を量る「リンク先引用率」を定義する。リンク先引用率はリンク先の内容が自コンテンツの中でどれくらい扱われているかを表すものである。

文書類似度はリンク元からリンク先からの評価であり、リンク先引用率はリンク先からリンク元コンテンツの評価として捉えることができる。この 2 つの評価軸を用いてリンクの参照重要度を求める。

4.3.1　文書類似度

Web ページ中の文章は n 次元の特徴ベクトルで表現する。特徴ベクトルの次元数は文書群から抽出された索引語の総数とする。

ベクトルの各要素としては、語の出現頻度を用いる、単純に出現を 1、非出現を 0 とする方法が挙げられるが、ここでは語の重み付けの代表的な出現法である tf/idf 法を用いる。計算式は以下の通りである。まず準備として、

$$idf_j = log\frac{N}{df_j}$$

を用意する。N は文書総数、df_j は語 t_j が出現する文書数である。これにより文書 D_i の語 t_j の重み w_ij は以下のように定義される。

$$w(p_i) = tf_{i,j} * idf_j$$

但し、$tf_{i,j}$ は語 t_j の文書 D_i での出現頻度である。

各文章を特徴ベクトルで表現できると、文書 D_α と文書 D_β の文書類似度 $sim(D_\alpha, D_\beta)$ が求まる。類似度の計算手法として以下のコサイン類似度を用いる。

$$sim(D_\alpha, D_\beta) = \frac{V_\alpha * V_\beta}{|V_\alpha||V_\beta|}$$

4.3.2　リンク先テーマ引用率

前章で定義した文書類似度はある Web ページの内容全体の類似性を調べるには適するが、テーマの関連性を計ることは出来ない。類似度はできるだけ文章中で用いられる単語が似ていれば高くなるが、そもそも書く人が異なれば用いられる単語も多種多様になり類似度だけで関連性を計るのは難しい。リンクを参照目的で使った場合、そこからコンテンツ作成者の主張がすることも類似度を下げる要素となる。テーマを全体とは関連性が低いが、コンテンツの部分的な補完を行うようなリンクは文書類似度では評価地が低くなる。

あるページとページが、「内容は異なるがテーマは同じ」になる状況を考えると、文書を構成する語の大半は異なるがテーマを決定付ける重要な語は一致していると考えられる。このことより、リンク元ページがリンク先のコンテンツの内容をどの程度引用しているかは、リンク先のページの重要な語をリンク元ページがどの程度含んでいるかを計ればテーマの関連性を調べられる。この評価値をリンクテーマ先引用率と定義する。リンク先引用率を用いると、文書間の類似度が低い場合でもテーマが似ていれば評価値は高くなる。またコンテンツの部分的な補完を行うために張ったリンクも正しく評価できる。

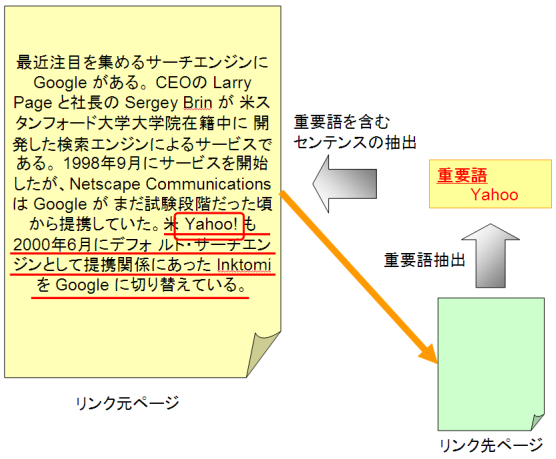

図 4　リンク先のテーマの引用範囲

具体的な計算は以下の手順で行われる。まずリンク先のテーマを表していると考えられる重要語 $Keyword$ を抽出する。重要語の定義は 2 つの手法が考えられる。1 つはリンク先ページ

のタイトル中に含まれる語と HTML 文書のメタタグ中に含まれるキーワードを合せたものである。タイトルはコンテンツのテーマを表しており、またメタタグ中のキーワードはコンテンツ作成者が文章を特徴づけるために用いたと考えられるからである。しかし実際の Web ページを観察すると、タイトルが必ずしもテーマを表しているとは限らない。また、メタタグ中のキーワードは検索エンジンに発見されやすいように意図的に入れる場合が多く、意味の無いものも多い。

そこで別の重要語抽出の方法として、tf/idf 法で求められた語の重みが大きいものを重要語と定義する。

抽出された重要語 $Keyword$ を用いてページ i のページ j に対すリンク先引用度 $V_{link(i,j)}$ を以下のように定義する。

$$V_{link(i,j)} = \frac{(\text{重要語を含むセンテンス数})}{(\text{ページ } i \text{ 中に存在するセンテンス総数})}$$

4.3.3 ページ品質の評価

参照重要度が高いリンクをたくさん持っているページは、自コンテンツの質を効果的に高めようとしている考えられる。あるページにあるリンクは全て補完するためのリンクでなく、相互リンクやリンク集、広告リンクもある。これらを考慮して、Web ページ i の品質 $Q(i)$ を以下のように定義する。

$$Q(i) = \frac{\sum V_{link(i,j)}}{N_{ValuableLinks}}$$

但し、$V_{link(i,j)}$ はページ i からページ j へ張られたリンクの参照重要度、$N_{ValuableLinks}$ は参照重要度の上位数件のリンクの数とする。

この式はリンクの参照重要度の上位数件の平均値を表している。

5. 実　　験

文書類似度とリンク先引用度がリンク元ページとリンク先ページの関連性をどの程度反映しているかを評価する。その結果を元にリンクの参照重要度を決定し、ページ品質の評価を行う。

5.1 文書類似度とリンク先テーマ引用度の検証

ページ間の関連性を評価するうえでの文書類似度の効果の検討、前章で定義した 2 種類のリンク先テーマ引用率の効果の比較を行う。また文書類似度とリンク先テーマ引用率の相関関係についても検討する。

5.1.1 実験の流れ

（1） リンクの参照重要度を測るリンク元ページをピックアップし、そこからリンクを張られたページ群を HTML 文書として取得する。

（2） リンク元ページとリンク先ページの HTML 文書から body タグで囲まれた部分をを抽出する。その際にリンクアンカーは除去する。これは本研究ではコンテンツの中心内容にのみ焦点を当てており、Web ページのテーマとは関係のないコンテンツの評価を行わないためである。またリンクアンカーの除去はアンカーテキストにはリンク先のページのタイトルが書かれることが多く、内容の類似度を量る場合に望ましくない影響が出るためである。

（3） 抽出された各文章を形態素解析し、「名詞」、「形容詞」を索引語とし、tf/idf 法を用いて特徴ベクトルを作る。

（4） 各ページの特徴ベクトル間のコサイン類似度を求め、文書類似度とする。

（5） リンク先ページの HTML 文書からメタタグ中に記述されたキーワードを抽出し重要語 1 とする。またタイトルを形態素解析し、得られた語を重要語 1 に加える。

（6） 各ページの特徴ベクトルの索引語のうち、tf/idf 法によって求めた重みの上位 20 件を各ページの重要語 2 とする。

（7） 各リンク先ページに対して、リンク先テーマ引用率を求める。

5.1.2 結果と考察

実験結果を以下に記す。図 7、9 はリンク先テーマ引用率の計算にキーワードとタイトル語を用いて求めたもの、図 8、10 は tf/idf 法を用いて求めたものである。参照されたページが価値のあるページかの判定は実際にリンク先のページがリンク元ページと関連があるかを手動で判断し、強い関連があれば○、一部関連があれば△、全く関連が無い場合を×として評価した。

文書類似度、リンク先テーマ引用率が単独でページ間の関連性にどの程度影響を与えているかを評価するために図 7、8 を記す。

また文書類似度とリンク先テーマ引用率の積とページ間の関連性への影響を評価するために図 9、10 を記す

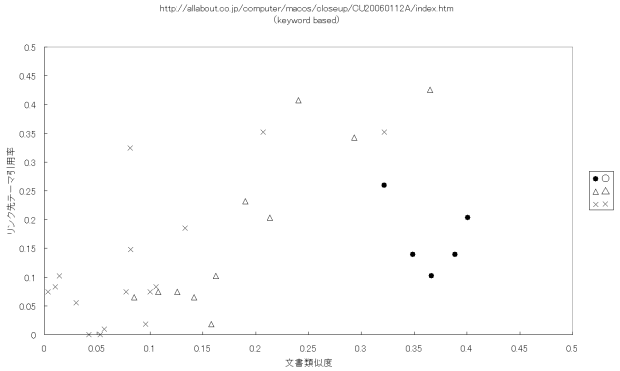

図 5　文章類似度とリンク先テーマ引用率の相関関係（キーワードベース）

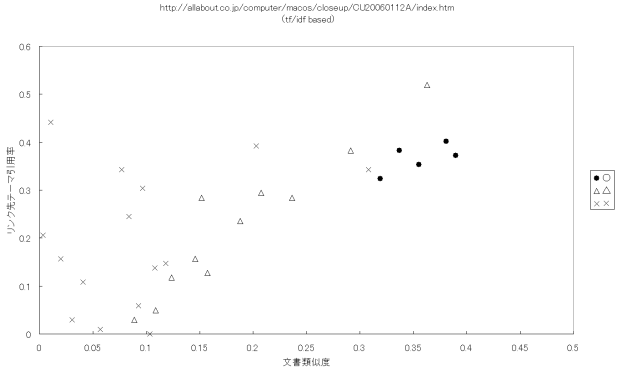

図 6　文章類似度とリンク先テーマ引用率の相関関係（tf/idf 法ベース）

文書類似度

テーマが似ているが、内容は異なるページを見つけるためには、文書類似度だけでは不十分で実際に内容まで似ているページを見つけてしまう可能性があると思われた。

しかし実際には、全く似ている、つまり類似度が極端に 1 に近くなるようなページが存在するはずが無い。実験によると、類似度が 0.5 程度のページはテーマが似ていて、かつ内容をかなり補完していると判断された。手動で「関連がある」と「一部関連がある」と判定されたページの類似度を比べると、「関連がある」ページの類似度のほうが高く、類似度が下がっていくに従って関連度が減少していく傾向が見られた。よって、類似度によってでページの補完度、関連性が計ることが可能であることが分かった。

リンク先テーマ引用率

テーマが同じで、かつ内容が異なる文章、部分的に内容が類似している文章を発見するために提案したリンク先テーマ引用率であった。

今回の実験では図 7、9 の縦軸を見れば分かるように、キーワードに基づく評価法、rf/idf 法に基づく評価法共に、リンク先テーマ引用率がリンク参照重要度に与える影響は見られなかった。原因として以下のことが考えられる。

キーワードとタイトル語を用いて引用率を求める場合の問題を考える。多くの Web ページではタイトル、キーワードが必ずしもコンテンツのテーマを表しているとは限らない。キーワードの場合はページランクを意図的に向上させる目的で不必要に挿入されるケースも見受けられる。またコンテンツ作成者によっては、キーワードは HTML に書き込まれない場合もある。このような理由から、キーワードとタイトル語がページのテーマを表しているとは考えられない。

tf/idf 法を用いる場合の問題を考える。この手法の場合、キーワードとタイトル語を用いた手法と比べると、文書を特徴付ける語をうまく抽出できる。しかし、抽出された語は文書を特徴付けているとは言えるが、テーマを特徴付けているかは分からない。

文書類似度とリンク先テーマ引用率の相関関係

文書類似度とリンク先テーマ引用率の 2 軸によってリンクの参照重要度が決まる、という仮定を設けた。文書類似度は文章の関連度を評価し、リンク先テーマ引用率はテーマの引用率を評価しているので、共に評価値が高くなることが望ましい。よって、文書類似度とリンク先テーマ引用率の積をとり、その値とページの関連性の評価を図 9、10 に記した。

図 9 から分かるようにキーワードに基づくテーマ引用率の定義では文書類似度とリンク先テーマ引用率の積の評価値とページの関連性は見られなかった。一方、tf/idf 法に基づくテーマ引用率の定義を用いた場合、図 10 から積の値が大きいほどページの関連性は高くなっている。しかし図 10 から実用的には文書類似度のみでページ間の関連性を計った方が結果が良いことが分かる。

以上の議論より、ページ品質の評価では文書類似度をリンク

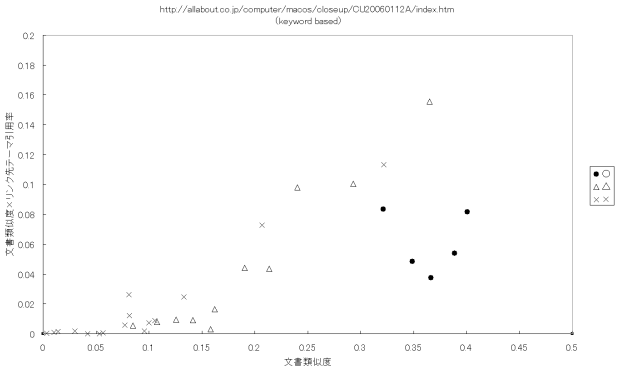

図 7　文章類似度とリンク先テーマ引用率の積による関連ページ判定（キーワードベース）

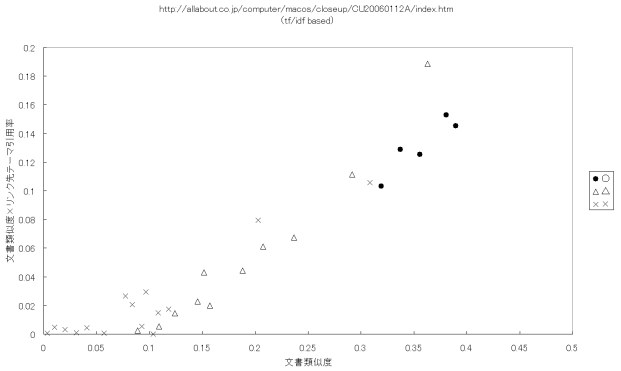

図 8　文章類似度とリンク先テーマ引用率の積による関連ページ判定（tf/idf 法ベース）

の参照重要度として実験する。

5.2　ページ品質の評価

5.2.1　実験の流れ

（1）　キーワードクエリを Yahoo に投げ、検索結果上位 40 件を取得する。

（2）　取得ページ各々から張られるリンクの参照重要度を求める。なおリンクの参照重要度は文書類似度とする。

（3）　得られた参照重要度を用いて、ページの品質を求める。

（4）　得られたページ品質評価値を元に検索結果をリランキングする。

5.2.2　結果と考察

クエリとして「ipod、nano、音質」をシステムに投げて得られた結果を図 11 に記す。元のランキングと比較できるよう Yahoo 検索でのランキングも付けた。

上位 2 位と 25 位以下のページ品質値がそれぞれ 1、0 になっているが、本システムでは基本的なテキスト形式の文書の計算にのみに対応しており、うまく処理がなされなかったため異常な値を示してしまっている。よってそれ以外の値を用いて結果を検証する。

リンク先のテーマがリンク元コンテンツと関連が高いほどページ品質の値が高くなることから、順位の高いページはコンテンツ内容と関連するリンクを持っており、ページを閲覧するユーザにとっては理解の助けになるページであった。今回の場合、ipod に関する blog やレビューページなどが上位に現れた

[クエリ]ipod nano　音質	提案手法ランキング	ページ品質	Yahoo検索順位
http://nanolog.sabaku.jp/?eid=82622	1	1	15
http://nanolog.sabaku.jp/?eid=82615	1	1	7
http://ipodnavi.livedoor.biz/archives/50126682.html	3	0.83410163	22
http://www.kakaku.com/bbs/Main.asp?PrdKey=013095	4	0.680327931	19
http://www.rbbtoday.com/news/20050329/21832.htr	5	0.599364797	18
http://d.hatena.ne.jp/roundtable/20050909	6	0.457103831	25
http://bubble4.2ch.net/test/read.cgi/wm/1126511954	7	0.440156734	14
http://galaxyquest.seesaa.net/article/6980342.html	8	0.437230361	11
http://sheepman.at.webry.info/200509/article_7.html	9	0.380397737	28
http://allabout.co.jp/computer/notepc/closeup/CU20	10	0.368794819	17
http://plusd.itmedia.co.jp/lifestyle/articles/0509/08/r	11	0.347790155	3
http://www.watch.impress.co.jp/av/docs/20050909/d	12	0.320831023	16
http://www.e-trend.co.jp/pcaux/46/139/product_720	13	0.316609707	23
http://www.ipod.cc/news+article.storyid+1048.htm	14	0.31341737	26
http://plusd.itmedia.co.jp/lifestyle/articles/0509/09/r	15	0.300919772	4
http://www.amazon.co.jp/exec/obidos/ASIN/B0007Y	16	0.277920399	1
http://www.amazon.co.jp/exec/obidos/ASIN/B0007Y	17	0.275008488	2
http://www.tuhanden.com/box/1860/324853	18	0.206916944	9
http://netsense.jp/blog/ipod音質.html	19	0.201595823	13
http://www.ipod.cc	20	0.159861199	5
http://www.drk7.jp/MT/archives/000910.html	21	0.143203673	6
http://www.drk7.jp/MT/archives/000922.html	22	0.106650281	8
http://arena.nikkeibp.co.jp/news/20050926/113641	23	0.08617473	12
http://arena.nikkeibp.co.jp/rev/portable	24	0.08388014	21
http://store-mix.com/ko-bai/product.php?afid=54070	25	0	10
http://bubble4.2ch.net/wm/subback.html	25	0	20
http://review.japan.zdnet.com/player/apple-ipod-nanc	25	0	24
http://www7a.biglobe.ne.jp/~ham700/report/nano2.ht	25	0	27
http://nexus.at.webry.info/200509/article_2.html	25	0	29
http://www.mm-style.net/ipod	25	0	30
http://yaplog.jp/osusumeprice/archive/44	25	0	31
http://kakaku.com/bbs/Main.asp?PrdKey=013095111	25	0	32
http://www.rakuten.co.jp/selection-echigo/589167/5	25	0	33
http://www.rakuten.co.jp/plusyu/423607/464505/45	25	0	34
http://find.2ch.net/?BBS=ALL&TYPE=TITLE&amp	25	0	35
http://yaplog.jp/osusumeprice/archive/276	25	0	36
http://allabout.co.jp/computer/av/closeup/CU20050	25	0	37
http://page6.auctions.yahoo.co.jp/jp/auction/f393430	25	0	38
http://inano.seesaa.net/article/8476259.html	25	0	39
http://www.selection-echigoya.com/pd/submenu_ipod	25	0	40

図 9　検索結果のリランキング

が、このようなページは同じ情報を扱う Blog、レビューページにリンクを多く持つ傾向がある。このような情報はリンク元ページの情報を補完し、ユーザの理解を助ける。

ページ品質が低いページは関連リンクが少ない、もしくは自コンテンツ内容が乏しいリンク集ページが見られた。リンク集ページの評価が低くなったのは、内容が少ないページからリンクをたくさん張っても自コンテンツの充実は図っていないことで評価を落としている。これは本研究の動機には適っている。関連リンクが少ないページが評価が低くなったのも、ユーザの理解の補助となるリンクが少ない、という観点からは評価値が低くなることは理解できる。

しかし、ランキングされた結果の順にページ品質の高いページが並んでいるとは考えにくい。本研究で定義したページ品質値は、リンク先ページとの文書類似度の平均値である。リンクの数が多く、その参照重要度が大きければ、ページを補完するリンクをうまく精選しているページであるとは言えるが、リンクの数が極端に少ない場合、少ないリンクでページの品質を評価するのは難しい。また、計算手法がページのテキスト内容に依存するので、トップページにリンクを張った場合はそれぞれの内容が薄いため、参照重要度が低くなってしまう。実際にリンクを参照として用いるときにはトップページに張ることがしばしば見受けられる。そもそもページを補完するリンクを張っているページが良質なページである、という仮定には無理があり、内容を補完するためにリンクを張らなくても、自コンテンツの内容のみで十分ユーザを満足させることが出来るページも存在する。提案手法で評価できるページは「良質なページ」ではなく「良質な参考リンクを持つページ」であろう。

6. おわりに

文書類似度とリンク先テーマ引用率でリンクの参照重要度を求めようとしたが、本研究で定義したリンク先テーマ引用率ではリンクの参照重要度は計れないことが分かった。類似度である程度リンクの参照重要度の判定はできるが、文書類似度のみを用いた方法では部分的な内容補足を評価することが出来ないので、より効果的なリンク先テーマ引用率を求める必要がある。

提案したリンクの参照重要度は、自コンテンツに対してリンク先コンテンツがどれだけ補完しているかを表し、その値が高ければ高いほどコンテンツ作成者は自コンテンツに効果的なリンクを張っていると言える。従って参照重要度の高いリンクを多く張っているページは、良質なページを作成しようとしているという観点で信頼性が高い。

しかし、この方法で Web ページの信頼性を計ろうとする場合、あるコンテンツがリンクを外向けに張っていることが前提になる。現在の Web を観察してみると、規模が小さいページではそれほど多くのリンクを張っていない。また、内容の補完の意味で用いられるリンクの数はそもそも多くない。Web 情報の信頼性を計る場合、そのように情報の少ないリンクの参照重要度のみで計算するのは妥当ではない。今回は「良質なリンクを多く張るページは良質である」というハブ的な視点でページを評価しようと試みたが、自コンテンツとの関係でリンクを評価すると、他のページとの相対的な品質評価ができない。また正確な評価はリンクの数に依存してしまうため、やはり妥当な方法ではない。

今後はリンクの参照重要度の定義を見直し、リンク以外の情報、特に内容を考慮したうえでコンテンツの信頼性を計る手法を検討していく。

7. 謝　　辞

本研究の一部は，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」、文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検索・統合ソフトウェア開発」（代表：田中克己），および，平成 17 年度科研費特定領域研究 (2)「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247，代表：田中克己）によるものです．ここに記して謝意を表すものとします．

文　　献

[1] Google
http://www.google.com/

[2] SEO 検索エンジン最適化
http:// www.searchengineoptimization.jp/concept-of-seo/index.html

[3] The PageRank Citation Ranking: Bringing Order to the Web
Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, 1998

[4] HITS
http://www.searchengineoptimization.jp/seo-foundation/ indexing/linkanalysys/hits-algorithm.html

[5] Combating Web Spam with TrustRank.
Zoltan Gyongyi, Hector Garcia-Molina, Jan Pedersen. VLDB 2004

[6] Temporal Ranking of Search Engine Results.
Adam Jatowt, Yukiko Kawai, Katsumi Tanaka.16th International Conference on Web Information Systems Engineering. WISE2005, 2005