自然言語処理 135－1
（2000．1．27）

参照重要度に基づくWWW検索

大野 潮満 黄瀬 浩一 松本 啓之亮
大阪府立大学 工学部 情報工学科

WWWを対象にした検索には，ページの評価を類似度で求める手法とWWWのリンク構造から求める手法がある．類似度に基づく検索は検索質問が特殊な場合に適しており，リンク構造に基づく検索は検索質問が一般的な場合に適している．しかし検索質問が一般的でもなく特殊でもない場合には，これらの手法では十分な結果が得られない．このような検索質問にも対応するには，類似度とリンク構造の両方を考慮してページを評価する必要がある．そこで本研究では，類似度とリンク構造に基づくページの評価指標として参照重要度を提案し，参照重要度に基づく検索システムを構築する．参照重要度は，反復計算によって類似度がリンクを通して伝搬することで求められる．実験より，本手法は従来法に比べよい結果を示すことが分かった．

WWW Retrieval Based on the Degree of Reference Importance

Shiomi Ohno, Koichi Kise, Keinosuke Matsumoto
Dept. of Computer and Systems Sciences, College of Eng., Osaka Prefecture University

From the viewpoint of evaluating web pages, methods for WWW retrieval can be classified into two types: methods based on the similarity to a query, and those based on the link structure. In general, the former are suitable for retrieval with specific queries, while the latter are fit for that with general queries. Those methods are, however, inappropriate for processing queries of intermediate specificity. In the report, we propose as another evaluation method the degree of reference importance which takes into account both the similarity and the link structure. The degree of reference importance is calculated by iterative propagation of the similarity through links. From the experimental results, we have confirmed that the proposed method is effective for all types of queries.

1 はじめに

WWWには30億ものページが存在し，日に100万づつ増えているといわれている [1]．このような膨大な情報の中から必要な情報を探し出すためには，情報検索は必要不可欠な技術となっている．

現在WWWの情報を検索する場合，検索エンジンが最も多く用いられている．検索エンジンは，ユーザから入力された検索質問とページの索引語を比較して類似度を求め，これをページの評価として結果を出力する．しかし，ユーザは検索結果に満足できないことが多い．これは，ページの評価をそのページの単語のみを用いて求めているためである．そこで近年，WWWのリンク構造に着目してページの評価を求める手法が提案されている [2][3]．しかし，ページの類似度をあまり考慮せずにリンク構造のみでそのページを評価しているため，ユーザは満足する結果を得られないことがある．

そこで本研究では，新たなページの評価指標として参照重要度を提案し，参照重要度に基づく検索システムを構築する．参照重要度は，類似度とリンク構造の両方に基づいて求められる．本稿ではまず，従来の検索システムと検索質問について述べる．次に，本システムについて述べ，従来の検索システムと本システムとの比較実験について述べる．

2 従来の検索システムと検索質問

2.1 検索質問

検索質問は，文章ではなく，論理演算子によって組み合わされたいくつかの単語で表現されることが

多い. 以下では, 検索質問は単語の論理積で表現されるとする.

検索質問には, 一般的なものから特殊なものまで様々なものがある. 検索質問がどの程度一般的あるいは特殊であるかは, 一般的に論じることが困難である. しかし, 検索エンジンの結果から間接的に判断することができる. つまり, 検索質問を検索エンジンに入力し, 得られた検索結果の件数が多いなら一般的な検索質問, 逆に少なければ特殊な検索質問といえる.

一般的な検索質問には, 例えば「linux」などが挙げられる. 一般的な検索質問を入力する場合, ユーザは検索質問自体の説明が書かれたページや, 検索質問に関して多くの人が参照する公式ページを求めることが多い. 一方, 特殊な検索質問には, 例えば「slackware ppxp ip dns」などが挙げられる. 特殊な検索質問を入力する場合, ユーザは検索質問に関して詳細に説明されていて, 余計な情報が書かれていないページを求めることが多い. さらに検索質問には, 一般的ではないが特殊でもない中間的なものもある. 例えば, 「mule perl-mode」などが挙げられる. 中間的な検索質問を入力する場合, ユーザは検索質問に関して詳細に説明されていて, かつ多くの人が参照するページを求めることが多い.

検索システムは, これら全ての検索質問に対応できなければならないといえる.

2.2 類似度に基づく検索

WWW では, Goo(http://www.goo.ne.jp/) などの検索エンジンがよく用いられている. 検索エンジンで用いられるデータベースは, 主に Web ロボットが WWW のリンクを辿ってページを収集し, ページに索引語をつけることで生成される. 索引語を付けるときページの内容を反映するように, タグの種類 (例えば, <H1> や <TITLE>) に応じて索引語に重みを付ける. そして検索エンジンは, 検索質問の単語と各ページの索引語を比較し, ページの評価として類似度を計算する. 検索結果は, 得られた類似度をもとに順序付けて出力される.

検索質問が特殊な場合, 検索結果はユーザが閲覧可能な量となる. また, 検索質問が特殊であるため分野が特定され, ユーザの満足する検索結果が得られることが多い. しかし検索質問が特殊でない場合, 検索結果は全てを閲覧することが不可能な量となる. また, 検索エンジンは類似度によって順位付けを行っているが, 精度が十分ではないため, ユーザは満足する検索結果を得られないことが多い. 以上のことから類似度に基づく検索は, 検索質問が特殊な場合には適しているが, それ以外の場合には適していないといえる.

●… 検索エンジンで検索されたページ
⊙… ● のリンク先のページ

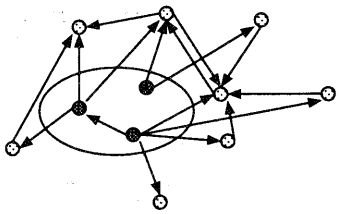

図 1: ページとリンク

類似度に基づく検索に上述の問題が発生するのは, ページの索引語のみを用いてページを評価することに原因がある. ページ単体だけでページを評価するのは困難であり, その他の特徴を用いる必要があるといえる.

2.3 リンク構造に基づく検索

類似度に基づいてページの評価を求める方法に対し, HITS[2] を初めとして, WWW のリンク構造を用いてページの評価を求める方法が提案されている. リンクは他のページを参照するために使用され, より詳しく書いてあるページを引用して文書を補完する場合や, 他のページと関連付ける場合に用いられることが多い. WWW のほぼ全てのページはリンクによって連結されており, このリンク構造の特徴から以下のことが導かれる.

- あるページが多くリンクされる場合, そのページの信頼性が高い. つまり, 多くのホームページ作成者があるページを参照することは, そのページに権威があるといえる.
- 信頼性の高いページへリンクしているページは, ユーザにとって道標となるページである.

以上のことからリンク構造に基づく検索では, 多くリンクされる信頼性の高いページは評価が高く, また, 信頼性の高いページへリンクしているページも評価が高いと考える.

具体的には次のようにしてページの評価を求めている. 図1に示すように, **2.2**の検索エンジンに検索質問を入力しページとリンクを得る. ページは検索エンジンで得られたページとそのページのリンク先のページであるため, 入力された検索質問に関連したページといえる. そして, これらのページの中で多くリンクされるページと多くリンクされているページへリンクしているページに高い評価を与える. このようにリンク構造に基づく検索は, 検索

質問に関連したページの多くの作成者がリンクするページと，それらのページへリンクしているページに高い評価を与える．

リンク構造に基づく検索では，検索質問が一般的な場合，図1のページとリンクが十分に得られる．よってページの信頼性が適切に求められ，ユーザの満足する検索結果が得られることが多い．しかし検索質問が一般的でない場合，検索結果には求める情報よりも一般的な情報しか得られないことがある．これは，リンクの使われ方に原因がある．ページ作成者は，例えば「mule と tex」について書く場合，「linux」などの単語を使用し，その単語にリンクを用いることがある．このようにリンクは，検索質問とはあまり関係のない，より一般的な情報に対して用いられることがある．よって一般的な情報のページが高い評価を得ることになり，検索質問とあまり関係のない情報しか得られなくなる．以上のことからリンク構造に基づく検索は，検索質問が一般的な場合には適しているが，一般的でない場合には適していないといえる．

リンク構造に基づく検索においてこのようなの問題が発生するのは，ページの評価をリンク構造のみを用いて求めることに原因がある．よってページを評価するには，リンク構造だけでなくページの類似度を考慮する必要があると考えられる．

3 本システムについて

3.1 類似度とリンク構造

前章で説明したように，類似度に基づく検索はページの索引語から，リンク構造に基づく検索はリンク構造からページの評価を求めている．類似度に基づく検索は，検索質問が特殊でない場合には適しておらず，リンク構造に基づく検索は，検索質問が一般的でない場合には適していない．そのため，検索質問が中間的な場合には，これら従来の検索システムでは，ユーザの満足する結果が得られないといえる．このような検索質問に対してもページに適切な評価を与えるためには，類似度とリンク構造を個別に用いるのではなく，その両方を用いる必要があると考えられる．

そこで本稿では，新たなページの評価指標として参照重要度を提案する．参照重要度は，類似度とリンク構造に基づいて求められる．類似度とリンク構造の両方を用いることで，内容がユーザの求める情報に類似し，かつ信頼性のあるページを得ることが期待できる．以下では，まず参照重要度を用いた本システムの概要について述べ，次に参照重要度について説明する．

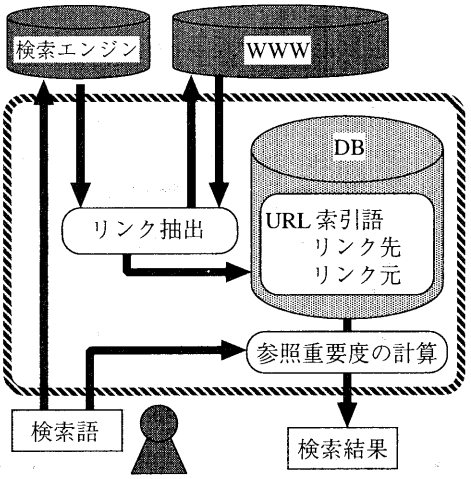

図 2: システムの概要

3.2 システムの概要

本システムの構成は図2のようになっている．まずユーザが検索質問を入力する．そして，得られた検索質問を検索エンジンに入力し，検索結果であるリンクのリストを得る．このリンク先のページと，さらにそのページに含まれるリンク先のページを WWW から取得し，DB に格納する．この DB には，得られたページの URL，そのページの索引語，リンク先の URL，リンク元の URL が保存される．この DB のページとリンクについては，**2.3**と同じである．ここで，ページの索引語は以下の方法で作成する．

日本語 Juman3.6[4] を用い，名詞を抽出する．ただし，平仮名だけからなる名詞は除く．

英語 wordnet-1.6[5] で用いられる名詞の辞書とコンピュータ用語辞典の辞書 [6] を用いて，名詞を抽出する．

本システムは得られた DB を用いて参照重要度を計算し，参照重要度の高い順にページをユーザへ提示する．

3.3 参照重要度

類似度とリンク構造の両方に基づいて求まるページの評価値は，以下のような性質を持つことが望ましいと考えられる．

- 検索質問との類似度が大きいほど高い値
- 評価値の高いページへリンクするほど高い値

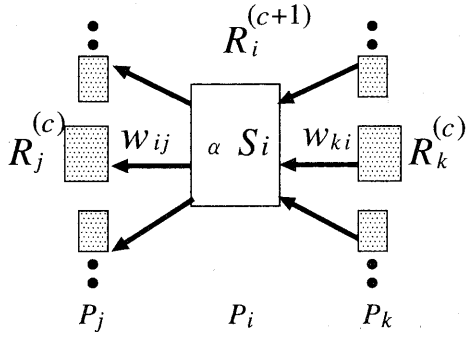

図 3: 参照重要度の計算

- 評価値の高いページからリンクされるほど高い値

そこで本研究では, これらの性質を考慮し, 参照重要度を次式で定義する (図 3参照).

$$R_i^{(c+1)} \leftarrow \alpha S_i + \sum_j w_{ij} R_j^{(c)} + \sum_k w_{ki} R_k^{(c)} \quad (1)$$

ただし, R_i はページ P_i の参照重要度である. P_j は P_i からリンクされているページであり, P_k は P_i へリンクしているページである. S_i は P_i の索引語と検索質問の類似度であり, α は S_i の重みである. ここで, 検索質問の単語と P_i の索引語のマッチ数を M, P_i の索引語の数を N とすると, S_i は以下の式で求める.

$$S_i = \frac{M}{N} \quad (2)$$

w_{ij} は, P_i から P_j へのリンクの重みである. 式 (1) を DB に格納されたページ数分集め列ベクトルで表現すると, 式 (1) は次式のように書ける.

$$\boldsymbol{R}^{(c+1)} \leftarrow \alpha \boldsymbol{S} + (\boldsymbol{W} + \boldsymbol{W}^T)\boldsymbol{R}^{(c)} \quad (3)$$

ただし $\boldsymbol{W}$ は w_{ij} からなるリンクの重み行列, $\boldsymbol{R}$ は R_i からなる参照重要度の列ベクトル, $\boldsymbol{S}$ は S_i からなる類似度の列ベクトルである.

式 (3) は, サイクル数 c を用いて反復計算で求める. ただし, $\boldsymbol{R}^{(0)} = \boldsymbol{O}$ である. $\boldsymbol{R}$ は, 式 (3) によってそのまま求めると発散してしまう. そこで, $\boldsymbol{R}$ が収束するように各サイクルごとに正規化を行う. すなわち,

$$\hat{\boldsymbol{R}}^{(c+1)} \leftarrow \alpha \boldsymbol{S} + (\boldsymbol{W} + \boldsymbol{W}^T)\boldsymbol{R}^{(c)} \quad (4)$$

$$\boldsymbol{R}^{(c+1)} = \frac{\hat{\boldsymbol{R}}^{(c+1)}}{\|\hat{\boldsymbol{R}}^{(c+1)}\|} \quad (5)$$

以上のように, 参照重要度は類似度とリンク構造から反復計算によって求められる. 反復計算を用いることで, 類似度がリンクによって伝搬し, 類似度とリンク構造の両方を考慮したページの評価が得られる. そして, α を大きくすることで類似度を重視し, 逆に小さくすることでリンク構造を重視したページの評価が得られる.

3.4 リンクの重み

リンクの重みについては, 文献 [1] で以下のことが述べられている. まず, WWW には, 例えばトップページへ戻るなど, ナビゲーションを目的としたリンクがある. そのため, あるサイトで作られているページの中でトップページが多くリンクされることになる. このようなリンクは一般に, リンク元のページとリンク先のページが同じドメイン名であることが多い. 次に, あるドメインにおいて多くのページが同一のページへリンクしていることがある. これはホームページ作成のために, 雛型ページが用意されているためである. 雛型ページには, あるページへのリンクがページの最上部や最下部に記述されている. この雛型ページが用いられると, あるページが多くリンクされることになる. 雛型ページはあるドメインの中で繰り返し用いられることが多い.

ナビゲーションのリンクや雛型ページによるリンクが用いられると, トップページやあるページが多くリンクされる. そのため, これらのページの信頼性が不当に高くなってしまう. したがって, ナビゲーションのリンクや雛型ページによるリンクには小さい重みを与え, ページの信頼性が不当に高くなるのを防ぐ必要がある.

以上の考え方から本システムでは, リンクの重み w_{ki} を次式で表す.

$$w_{ki} = \frac{log(1 + L_{ki})}{log(2 + C_k)} \quad (6)$$

ここで, L_{ki} は P_k から P_i へリンクが存在するなら 1, 存在しないなら 0 である. ただし, P_k と P_i が同一のドメイン名である場合はナビゲーションのリンクであると考え, L_{ki}=0 とする. C_k は P_i へリンクしているページのうち, P_k とドメイン名が一致するページの数である. C_k は, 雛型ページが用いられる場合に大きな値となる.

3.5 HITS の計算式との比較

この節では, **4.2**の実験で本システムと比較する HITS について計算式を説明し, 本手法の計算式との比較を行う.

HITS は, 以下の式によってページの評価を求めている.

$$R_x^{(c+1)} \leftarrow W'^T W' R_x^{(c)} \quad (7)$$

$$R_y^{(c+1)} \leftarrow W' W'^T R_y^{(c)} \quad (8)$$

W' と W'^T はリンクの重み行列であり, R_x と R_y はページの評価である. HITS ではページの評価について, リンクされることで高くなるページの評価を R_x, リンクすることで高くなるページの評価を R_y としている.

本手法では, ページの評価に W と W^T の和を用いている. よって, HITS では $W'^T W'$ と $W' W'^T$ による評価が別々であるが, 本手法では同時に評価することになる.

4 実験と考察

以下では, まず実験条件について述べ, 次に本システムの有効性を検討するための比較実験について述べる. 最後に, 参照重要度の効果を調べた実験について述べる.

4.1 実験条件

実験は 1999 年 10 月現在の WWW を対象に行った. 用意した検索質問を検索エンジンである Goo に入力して, 得られたページのリストのうち上位 1000 件を用いて DB を作成した. 検索質問には, 一般的な検索質問, 特殊な検索質問, 中間的な検索質問の 3 種類を用意した.

具体的には, 以下の 3 つの検索質問を用いた.

検索質問 1 linux

検索質問 2 linux redhat sendmail nomx

検索質問 3 windows mule tex xdvi install

各検索質問を用いたときの Goo から得られた検索結果の件数, その結果を用いて作成した DB に含まれるページ数, DB 内のページ同士を結ぶリンク数を表 1に示す. 検索質問 1 は, 一般的な検索質問であり, ユーザが公式ページを求める場合を想定した. 検索質問 2 は, 特殊な検索質問であり, 詳細に述べられているページをユーザが求める場合を想定した. また, 検索質問 3 は検索質問 1 ほど一般的ではないが検索質問 2 ほど特殊ではない検索質問であり, 詳細なページをユーザが求める場合を想定した.

4.2 比較実験の結果と考察

比較実験では, 類似度に基づく検索である Goo, リンク構造に基づく検索である HITS, そして本システムを比較した. まず, 検索質問を Goo に入力し,

表 1: Goo の検索結果の件数と DB の詳細

	件数	ページ数	リンク数
検索質問 1	232939	5349	8174
検索質問 2	55	312	47
検索質問 3	226	6270	15432

表 2: パラメータ

	検索質問 1	検索質問 2	検索質問 3
α	1.0×10^{-10}	10	20

得られたページのリストの上位 20 件を Goo の検索結果とした. 次に, HITS と本システムは DB を用いて計算し上位 20 件を出力した. ただし HITS については, 式 (7), (8) で示したように 2 種類の結果があるので, それぞれ上位 10 件づつを検索結果とした. また, HITS のリンクの重み W',W'^T には, それぞれ本システムの W,W^T と同じ値を使用した. HITS と本システムの計算では, c=200 まで反復計算を行った. これは, ページの評価の順位が決まるのに十分な回数である.

本システムでは, α を 0 から 100 まで 5 ずつ変化させた. ここで実際には, 本システムでは α=0 のとき参照重要度は全て 0 となってしまうため, 0 の代わりに $\alpha = 1.0 \times 10^{-10}$ を, また 100 の代わりに極限として $\alpha = 1.0 \times 10^{10}$ を用いた. それぞれの質問において最もよい結果であると判断した α を表 2に示す.

(1) 検索質問 1 の結果

Goo で得られた代表的なページを表3に示す. また本システムと HITS で得られたページは, 20 件中 15 件が一致した. この一致したページのうち代表的なページを表 4に示す. 検索質問 1 を用いた場合, 3 つの検索システムの検索結果には, 主に以下のページが含まれた.

Goo 企業の Linux 関連の情報, Linux の協議会の案内, HOWTO, リンク集, メールのリストのページ

HITS と本システム 有名な公式ページとそれらのページへのリンク集のページ

以上のことから, 検索質問 1 のような一般的な検索質問の場合, 次のことがいえる.

Goo 類似度のみでページを評価したため, 単に検索質問に用いられる単語が含まれるページを出力

HITS と本システム リンク構造を考慮しているため, Goo よりも信頼性の高いページを出力

表 3: 「検索質問 1 」Goo の代表的なページ

http://www.ntl.co.jp/index.html
http://iw98.linux.or.jp/index.html
http://www.teu.ac.jp/hiha/~amano/old/link/links/index.html
http://www.v6.linux.or.jp/linux6bone.html
http://iw98.linux.or.jp/reception.html
http://www.linet.gr.jp/ML/plagia/19990223/index.html
http://pc1.peanuts.gr.jp/1999/0227/Network/Multicast/index.html
http://www.hero-island.ne.jp/linux/index.html
http://cnn.co.jp/TECH/computing/9905/14/lxrun.idg/index.html
http://linux.oracle.co.jp/ols_dl.html

表 4: 「検索質問 1 」本システムと HITS に共通の代表的なページ

http://www.linux.or.jp/index.html
http://www.redhat.com/index.html
http://www.debian.org/index.html
http://www.slackware.com/index.html
http://www.pje.linux.or.jp/index.html
http://www.debian.or.jp/index.html
http://netvillage.www.linux.or.jp/index.html

すなわち, 本システムはリンク構造を重視することで, 従来のリンク構造に基づく検索と同様に, 公式ページなどの信頼性のあるページを得ることができたといえる.

(2) 検索質問 2 の結果

HITS の得られた代表的なページを表5に, また Goo と本システムで得られた代表的なページを表6に示す. 検索質問 2 を用いた場合, 3 つの検索システムの検索結果には, 主に以下のページが含まれた.

HITS 検索質問とは関係のないソフトウエアの情報, ソフトウエアの日本語化などのページ

Goo と本システム 検索質問に関する問題や回答が述べられているメーリングリストのアーカイブ, sendmail についてまとめてあるページ

以上のことから, 検索質問 2 のような特殊な検索質問の場合は以下のことがいえる.

HITS リンク構造のみでページを評価するため, 検索質問とはあまり関係のないページを出力

Goo と本システム 類似度を考慮しているため, HITS よりも内容が検索質問に類似したページを出力

すなわち, 本システムでは類似度を重視することで, 従来の類似度に基づく検索と同様に, よい検索結果を得ることができたといえる. しかし表6に示すように, Goo, 本システムともよいページが得られた

表 5: 「検索質問 2 」HITS の代表的なページ

http://www.bekkoame.or.jp/egota/~NetscapeV.html
http://skk.kuis.kyoto-u.ac.jp/skk/index.html
http://ac3.aimcom.co.jp/~macpgp/index.html
http://jf.linux.or.jp/index.html
http://www.omronsoft.co.jp/index.html
http://www.lesstif.org/index.html
http://www-nagao.kuee.kyoto-u.ac.jp/member/tsuchiya/sdic/index.html
http://okux.matsusaka-u.ac.jp/index.html
http://www.pje.linux.or.jp/Howto/PJE-0.1.5cm/html/PJE-HOWTO-6.html
http://okux.matsusaka-u.ac.jp/~okumura/linux/egcs.html

表 6: 「検索質問 2 」Goo と本システムの代表的なページ

Goo	http://www.ryukyu.ne.jp/~sheemer/oia/benkyoukai/19980912/index.html
	http://www.onsen.ne.jp/howto/sendmail.html
	http://www.mizuno.riec.tohoku.ac.jp/~yasu/ml/linux/users/3500/3508.html
	http://fish.miracle.ne.jp/s-makoto/makoto/ocn.html
本システム	http://okux.matsusaka-u.ac.jp/~okumura/networking/sendmail.html
	http://www.mizuno.riec.tohoku.ac.jp/~ml/linux-users/3400/3476.html
	http://www.mizuno.riec.tohoku.ac.jp/~ml/linux-users/3400/3483.html
	http://www.onsen.ne.jp/howto/sendmail.html

のは, 2 0件中 4 件と少なかった. よって類似度を用いた Goo と本システムは, 十分な結果が得られているとはいえない. このことから現在の類似度の求め方は十分ではなく, 改良する余地があるといえる.

(3) 検索質問 3 の結果

Goo, HITS で得られた代表的なページをそれぞれ表 7, 表 8に示す. また, HITS と Goo では得られなかったが, 本システムでは得られた代表的なページを表 9に示す. 検索質問 3 を用いた場合, 3 つの検索システムの検索結果には, 主に以下のページが含まれた.

Goo 様々な分野へのリンクを集めたページ, 様々なインストールの記録をメモ程度に記述したページ

HITS 検索質問よりも一般的な情報の公式ページとそれらのページへのリンク集のページ

本システム 検索質問に関して詳しく説明されている本の紹介のページ, ある大学でまとめられているページ, それらのページへリンクしているページ

表 7:「検索質問 3 」Goo の代表的なページ

http://www.nak.ics.keio.ac.jp/~satoken/link.html
http://www-a2i.is.tokushima-u.ac.jp/~hino/Install/97.html
http://www.ee.aoyama.ac.jp/~y-abe/lab/my_install.html
http://www.ioa.s.u-tokyo.ac.jp/~ykanya/computer/index.html
http://www.jks.is.tsukuba.ac.jp/~e961449/link/index.html
http://benz.ise.eng.osaka-u.ac.jp/~waka/mybookmarks.html
http://www.annie.ne.jp/~imaishi/nikki.htm

表 8:「検索質問 3 」HITS の代表的なページ

http://samba.bento.ad.jp/index.html
http://www.linux.or.jp/index.html
http://www.apache.or.jp/index.html
http://www.jp.freebsd.org/index.html
http://www.isse.kuis.kyoto-u.ac.jp/usr/yy/links/Software.html
http://www.nucl.phys.tohoku.ac.jp/~m_isihar/server.html
http://kk-www.ss.titech.ac.jp/~shaba/link/links.html
http://www-a2k.is.tokushima-u.ac.jp/member/hiroki/link/link.html

以上のことから, 検索質問 3 のような中間的な検索質問の場合, 以下のことがいえる.

Goo 類似度のみでページを評価したため, 単に検索質問に用いられる単語が含まれるページを出力

HITS リンク構造のみでページを評価したため, 検索質問とはあまり関連がなく, 単に多くリンクされているあるいはリンクしているページを出力

本システム 類似度とリンク構造の両方を用いることで, 互いが不足している部分を補い合い, ユーザの求める情報に類似しかつ信頼性のあるページを出力

よって, Goo や HITS に比べ本システムはよい検索結果が得られたといえる.

以上のことから, 検索質問 1 のような一般的な検索質問の場合, そして検索質問 2 のような特殊な検索質問の場合, 本システムは α を設定することで類似度を重視あるいはリンク構造を重視した検索が可能であり, 従来の検索システムと同程度の検索結果が得られると分かった. また, 検索質問 3 のような中間的な検索質問の場合, 本システムは類似度とリンク構造の両方を考慮した検索が実現でき, 従来の検索システムでは得られないような検索結果が得られたといえる.

表 9:「検索質問 3 」本システムのみ検索できた代表的なページ

http://www.matsusaka-u.ac.jp/~okumura/texfaq/index.html
http://www.ce.kochi-ct.ac.jp/~isami/home/int_texj.html
http://www.cutt.co.jp/book/mulewin.html
http://home.impress.co.jp/catalog/impress/bk4652/index.htm
http://www.maekawa.is.uec.ac.jp/research/eunice/internal/bookmarks/80.html

課題としては, まず類似度の改良がある. 今回の実験より, リンク構造はページの評価を十分に改善しているものの, 類似度については改良の余地があることが分かった. また, 本システムは, α を大きくすることで類似度重視, 小さくすることでリンク構造重視の評価を得ることができた. しかし現在, α はユーザが設定する必要があり, これはユーザにとって大きな負担となる. そこで α の自動設定が必要であるといえる.

4.3 内分法との比較

類似度とリンク構造の両方を考慮する最も簡便な手法は, 類似度とリンク構造の各々によって得られる評価値の加重和によってページを評価する手法 (以下, 内分法) である. そこで参照重要度の反復計算による効果を調べるため, 内分法と本手法を比較した. この比較実験には, 検索質問 3 から作成した DB を用いた.

本手法では, α を **4.2**と同様に変化させて実験を行った. リンク構造のみの評価 ($\alpha = 1.0\times10^{-10}$) の上位 20 ページ, 類似度のみの評価 ($\alpha = 1.0\times10^{10}$) の上位 20 ページでは得られないページが α=20 で 5 ページ得ることができた. 代表的なページについて, α と R の関係を図 4に示す. ここで, A, B, C は α=20 の上位 20 ページのうち, α=1.0×10^{-10},1.0×10^{10} の上位 20 ページに含まれないページである. また, X は類似度のみで評価されるページ, Y はリンク構造のみで評価されるページである.

内分法では, まず類似度のみの評価として式 (3) の $\boldsymbol{S}$ を用いた. つぎにリンク構造のみの評価として, 式 (3) を用いて α=0, $\boldsymbol{R}^{(0)}$ の各要素を 1 とし $c = 200$ まで反復計算を行った. 計算結果の値は $\boldsymbol{W}+\boldsymbol{W}^T$ の最大固有値の固有ベクトル $EV(\boldsymbol{W}+\boldsymbol{W}^T)$ に収束する [7]. そして, 内分法のページの評価 R_N として次式を用いた.

$$R_N = \beta\boldsymbol{S} + (1-\beta)EV(\boldsymbol{W}+\boldsymbol{W}^T) \qquad (9)$$

β を 0.0 から 1.0 まで 0.5 ずつ変化させたところ, リンク構造のみの評価 (β=0.0) の上位 20 ページ, 類似度のみの評価 (β=1.0) の上位 20 ページでは得ら

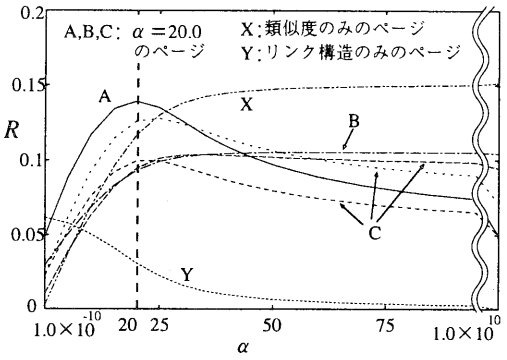

図 4: R と α

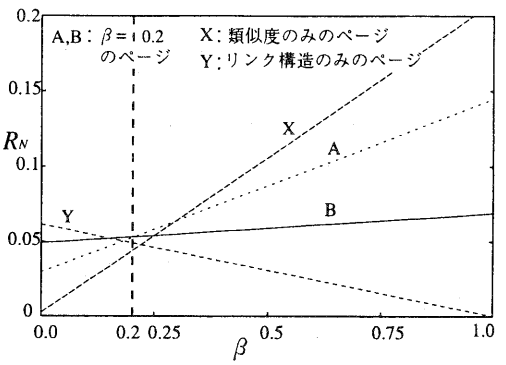

図 5: R_N と β

れないページが β=0.2 で 2 ページ得ることができた. 代表的なページについて, β と R_N の関係を図5に示す. ここで, A, B, X, Y は図 4のものと同じページである. また, A, B は β=0.2 の上位 20 ページのうち, β=0.0,1.0 の上位 20 ページに含まれないページである.

図 4と図 5に基づき, 本手法と内分法を比較する. まずグラフ全体について, 内分法はリンク構造と類似度の各々の評価の加重和であるため, ページの評価値は直線的に変化する. 一方, 本手法は類似度がリンクを通して反復計算によって伝搬するため直線とはならない. 次に X, Y に着目すると, 内分法, 本手法のいずれの場合も X は単調増加, Y は単調減少となり大きな差はないといえる. しかし A, B, C から, 本手法は内分法より類似度とリンク構造の両方を重視したページをより多く取り出せることが分かる. 特に図 4の C のページは, $\alpha = 1.0 \times 10^{-10}$, 1.0×10^{10} で B のページの下位にあるが, α=20 では B より上位にある. つまり C は内分法では得られないページであり, 本手法は内分法では得られない効果を得ることができたといえる. また C の 3 つのページは, 表 9の 1 行目のミラーページが 2 つと 2 行目のページからなり, 本手法でしか得られない効果は検索結果を向上させることができたといえる.

以上のことから, 単に類似度の評価とリンク構造の評価の内分を取るのではなく, 反復計算を用いることで, 本手法は類似度とリンク構造の両方を重視したページにより大きな評価値を与えることが可能であることが分かった.

5 おわりに

WWW 検索において, ページを評価するためには, 類似度とリンク構造を個別に用いるのではなく, 両方を考慮する必要があるという観点から, 本稿では参照重要度を提案し, 参照重要度に基づく検索システムを構築した. 本手法の特徴は, 類似度とリンク構造の影響をパラメータを用いて調節し, 類似度重視やリンク構造重視の検索, そして両方を重視した検索を行うことである. 本実験の結果から, 従来法よりよい検索結果が得られることが分かった. 今後の課題としては, 類似度の改良と α の自動設定がある.

参考文献

[1] Chakrabarti,S., Dom,B.E., Kumar,S.R., Raghavan,P., Rajagopalan,S., Tomkins,A., Gibson,D. and Kleinberg,J. : Mining the Web's Link Structure, *IEEE Computer*, August, pp60–67(1999).

[2] Kleinberg,J.M.: Authoritative Sources in a Hyperlinked Environment, *Proc.9th ACM-SIAM Symp. Discrete Algorithms*, ACM Press, New York and SIAM Press, Philadelphia, pp.668–677(1998).

[3] Savoy,J. : Ranking Schemes in Hybrid Boolean Systems: A New Approach, *JASIS*, vol.48, no.3, pp235–253(1997).

[4] 黒橋 禎夫, 長尾 眞 : "日本語形態素解析システム JUMAN version 3.6", 京都大学工学部大学院工学研究科 (1999).

[5] wordnet-1.6 : <http://www.cogsci.princeton.edu/~wn>

[6] ASCII GROSSARY HELP : <http://www.ascii.co.jp/ghelp/ab.html>

[7] Golb,G. and Van Loan,C.F.: *Matrix Computations*, Johns Hopkins University Press, Baltimore(1989).