ブックマークされているページのData Qualityを考慮したPersonalized PageRank手法の提案

谷津　昌宏†　　有次　正義††

† 群馬大学大学院工学研究科情報工学専攻　〒376-8515 群馬県桐生市天神町1-5-1
†† 群馬大学工学部情報工学科　〒376-8515 群馬県桐生市天神町1-5-1
E-mail: †masahiro@dbms.cs.gunma-u.ac.jp, ††aritsugi@cs.gunma-u.ac.jp

あらまし　Webページの重要度を判定する手法としてPageRankがある．これを基にユーザの嗜好を反映したPersonalized PageRank手法が盛んに研究されている．本稿ではユーザの嗜好を表すデータとして,ユーザのブックマークされているWebページを利用したPersonalized PageRank手法を提案する．具体的には，ブックマークされているWebページを,Data Qualityを扱う分野で研究されている *timeliness*，*currency*，*completeness*，*consistency*，*accuracy*，*accessibility*，*believablity* の複数の指標で評価し，その評価値を考慮して重要度を判定するPersonaliozed PageRank手法を提案する．

キーワード　Data Quality，PageRank，Personalized，ブックマーク

A Personalized PageRank Method with Applying Data Quality Measurement to Bookmarked Pages : a Proposal

Masahiro YATSU† and Masayoshi ARITSUGI††

† Department of Computer Science, Graduate School of Engineering, Gunma University
1-5-1 Tenjin-cho, Kiryu, Gunma 376-8515, Japan
†† Department of Computer Science, Faculty of Engineering, GunmaUniversity
1-5-1 Tenjin-cho, Kiryu, Gunma 376-8515, Japan
E-mail: †masahiro@dbms.cs.gunma-u.ac.jp, ††aritsugi@cs.gunma-u.ac.jp

Abstract　The PageRank method evalutes the importance of Web pages. There has been a growing interest in effective processing on personalized PageRank methods reflecting user's interests based on the PageRank method. In this paper, we propose a personalized PageRank method applys bookmarked pages that represent user's interests. Specifically, we evalute bookmarked pages using multi measurements of timeliness, currency, completeness, consistency, accuracy, accessibility and believablity that are investigated in the area of data quality. Considering these evalution resultsCwe measure the importance of Web pages using personalized PageRank mothod.

Key words　Data Quality, PageRank, Personalized, Bookmark

1. はじめに

近年，インターネットの普及に伴い，WWW空間上においてWebページの数は爆発的に増加している．この膨大なWebページの中からユーザが必要とする情報を抽出することは困難である．その解決策の1つに，WWW空間上の各Webページの重要度を判定し，検索するWeb検索がある．

その代表的な手法にPageRank[1]というアルゴリズムがある．これはWebページのリンク構造を解析し,「多くの良質なWebページからリンクされているWebページは，やはり良質なWebページである」という再帰的な関係をもとに，Webページの重要度を計算している．

しかしこのPageRankでは，各Webページに対し汎用的な重要度しか判定できず，この結果が個人の満足を得ているとは言い難い．そこでよりユーザの嗜好を考慮したWebページの重要度を判定するPersonalized PageRank[2]〜[5]の研究が盛んになってきている．Personalized PageRankのアルゴリズムではユーザの嗜好を表わすデータが必要となり，その選出と評価がユーザの満足する結果を得るために重要となる．

ユーザの嗜好を表わすデータの1つとしてユーザにブック

マークされている Web ページが挙げられる [4], [5]．ブックマークされている Web ページはユーザが何度も訪れるサイトやユーザが興味を持っている内容についてのサイトを含むので，十分にユーザの嗜好を含むと言える．[4], [5] ではブックマークされている Web ページを使い，その Web ページ内のリンク情報からその周辺の Web ページを求めている．その Web ページ集合の PageRank のスコアからユーザの嗜好を表わす Web ページ集合を求める．その Web ページ集合に一定の重みを割り当てることで Personalized PageRank のアルゴリズムの入力値としている．

本稿では，ユーザにブックマークされている Web ページとその外部リンク 1 回で辿ることの出来る Web ページを 1 つの集合として，集合内の各 Web ページに対して複数の側面から評価する手法を提案する．その Web ページ集合に対し，データのユーザにとっての使いやすさを評価する Data Quality [6]〜[10] の分野で研究されている *timeliness*，*currency*，*completeness*，*consistency*，*accuracy*，*accessibility*，*believability* の複数の指標で重要度を評価する．7 つの指標それぞれの評価値をユーザがパラメータで調節し，ユーザの嗜好を表わす重みとして割り当てる．その重みを Personalized PageRank のアルゴリズムにおけるユーザの嗜好を表わす入力値とすることで，ユーザの様々な嗜好をより反映した Web ページの重要度を判定できないかを考える．

本稿の構成は次の通りである．まず 2. で，本研究の関連研究について述べる．3. では Data Quality について説明し、4. で提案手法について述べる．5. では提案手法を用いた評価実験について述べる．最後に 6. で，まとめと今後の課題とする．

2. 関 連 研 究

Jeh と Widom はユーザにブックマークされている Web ページを使った Personalized PageRank を提案している [4]．Page らが提案する PageRank は汎用的な重要度を判定するもので，個人の嗜好は考慮していない．より精製した検索結果を出すために，ユーザの好みの Web ページを集合 P とし，WWW 空間上での“personalized view”の重要度を考える．具体的には，ランダムサーファーモデル内でサーファーはそれぞれのステップにおいて確率 c で集合 P 内の Web ページにジャンプして戻る．そして確率 $1-c$ でリンクを辿り前に進む．このモデル内のサーファーの最終的な配置をパーソナライズした集合 P 上の *personalozed PageRank vector*（*PPV*）とする．*PPV* は WWW 空間上の Web ページの重要度の personalized view と言える．この *PPV* を使い．ランキングのスコアを出す．*PPV* の計算をするときに，ユーザの嗜好を表わすデータとして Personalized PageRank ユーザプロファイル（以後，*PPR* ユーザプロファイル）を入力値としている．このとき，*PPR* ユーザプロファイルとなる Web ページ集合 P は集合 H の部分集合としている．集合 H とは *hub* ページのことを指し，具体的には *Yahoo!* [11] や *OpenDirectory* [12] といった，人が構築したディレクトリ内の Web ページや高い PageRank のスコアを持つ Web ページなどを指す．その集合 H の中から，ユーザの好みに合った Web ページをユーザ自身が選出している．そして，Personalized PageRank の計算時に集合 P 内の各 Web ページに対して，(1) 式の $u(p)$ のように集合 P 内の (1/Web ページ数) という一定の重みを割り当てている．各 Web ページのリンク情報を A で表わす．ある Web ページ i をのスコアを計算するときにその Web ページ i の被リンクを持つ Web ページの集合の 1 つを j とする．このとき，Web ページ j のリンク数を $O(j)$ として，Web ページ j のスコアを A の値分だけ Web ページ i に割り当てる．ユーザの嗜好を表わす PPR ユーザープロファイルを u，ジャンプの確率を c とすると，PPV を求める式 v は式 (1) のようになる．

$$\begin{cases} p \in P \text{ のとき}, u(p) = \frac{1}{|P|} \\ p \notin P \text{ のとき}, u(p) = 0 \end{cases}$$

$$\begin{cases} j \text{ が } i \text{ へリンクを持つとき}, A_{i,j} = \frac{1}{|O(j)|} \\ \text{それ以外のとき}, A_{i,j} = 0 \end{cases}$$

$$\boldsymbol{v} = (1-c)\boldsymbol{A}\boldsymbol{v} + c\boldsymbol{u} \tag{1}$$

Chirita，Olmedilla と Nejdl は Jeh らの PPV アルゴリズムを使い，パーソナライズのランキングをするための *PROS*（*A Personalized Ranking Platform*）というプラットフォームを構築した．

Chirita らは Jeh らの提案したユーザの嗜好を表わす Web ページ集合の選出の仕方と選出した Web ページ集合への重みを割り当て方ついて問題点を指摘し，改善策を提案している [5]．Jeh らの提案ではその Web ページ集合の選出をユーザが行うために自動化が出来ず，時間も掛かるという問題があった．そこでユーザのブックマークとプロキシサーバを使い，ユーザの直接入力を必要としない Web ページ集合を選出するアルゴリズムと，選出した Web ページ集合への重みの割り当て方を提案している．以下に挙げた 2 つの Web ページ集合に対してそのアルゴリズムを適用する．

- 最も立ち寄ったページ

ユーザが訪問した Web ページをプロキシを使い追跡する．プロキシは各 Web ページを見ている継続時間と，その Web ページに戻ってくる頻度を記録する

- ユーザのブックマーク

ユーザがブックマークしている Web ページを使う

このアルゴリズムでは，Web ページのリンク情報からリンク，被リンクで辿れる Web ページのうち，6 回のリンクで辿れる Web ページを関連する Web ページの集合としている．

これはサーチエンジンに送ったクエリのトップの結果を抽出し，それらの周りのWWW空間上のサブグラフを構築する．その時，トップの Web ページのうち，その Web ページ自身が被リンクを持ち，さらに多くても d 個の Web ページへのリンクを持っている Web ページを全て加えることで，この基本集合を拡張する．より多くの Web ページの収集のために何度か拡張する．この結果，出力の集合が膨大になるので，それぞれの中間ステップでの後刈り込みが必要となる．その時，刈り込みの基準として *HubRank* で Web ページのスコアを付ける．こ

れは Jeh らのアルゴリズムがページランクの高い Web ページを入力として必要とし，*HubRank* はページランクのバイアスとしてデザインされているからである．以上のステップから *hub* 集合を構築する．このとき得られた集合の Web ページ数を NP とすると，得られた集合内の各 Web ページに対して，式 (2) のように一定の重みを割り当てている．

$$\begin{cases} p \in P \text{ のとき}, u(p) = \frac{1}{NP} \\ p \notin P \text{ のとき}, u(p) = 0 \end{cases} \quad (2)$$

この式 (2) の $u(p)$ を PPR ユーザープロファイルの集合 P の値として式 (1) に適用している．

3. Data Quality

本研究ではユーザにブックマークされている Web ページに対して，ユーザにとってのデータの使いやすさを評価する Data Quality の複数の指標で評価している．ここでは Data Quality について説明する．

3.1 概　　要

近年，ソフトウェアの発達やマルチユーザ環境の広がりに伴い，1 つのデータが複数のソフトウェアや複数のユーザで使用されるようになった．その結果，様々なソフトウェアやユーザ環境でデータが交換可能にする技術が発達した．それらソフトウェア間，ユーザ間でのデータの質の保証が重要となり，Data Quality という研究分野が発達した [7]．Redman は質の悪いデータが経済，企業，政治，学問の枠を超えた全ての分野に影響を与える普遍的な問題を引き起こすものであり，その認識が必要だと述べている [6]．Data Quality という言葉は “Fitness for use” として定義され，それはデータの質のことを指す．データの質とはどのくらいユーザの必要性に合うデータであるかということである．あるユーザにとっては質が高いデータであっても他のユーザではそうとは限らない．それはユーザによってデータに求める要求が異なるからである．Data Quality はユーザの異なる視点の要求を全て評価することで，データの評価をしている．この概念を出発点とし，Wand らは理論的，実験的結果から，データの質を評価するための複数の側面を考えた．これが Data Quality Dimensions である．

3.2 Data Quality Dimensions

現在あらゆる側面からデータの質を評価するためにたくさんの Data Quality Dimensions が定義されている [6]．しかし，本研究では Web ページソースを評価するために使用する 7 つの Dimension についてのみ紹介する．他にも文献 [6] では *Interpretability*, *Portability*, などの Dimension が存在するがこれらの Dimension が Web ページの Data Quality の評価に適用できるかについては現在検討中である．

- *Timeliness*

データの中にはデータが新しく生成されるものや，データが変更・更新されるものがある．そのデータにおける生成や変更，更新の程度で表わす．

- *Currency*

データから得られる情報には，その情報が有効である期限が付いているものがある．その期限が有効である程度で表わす．またデータベースでは最後に更新された時間で表わす．

- *Completeness*

実世界で動いている具体的なシステムにおけるいかなる重要な状況においても，そのデータを表現できる程度で表わす．

- *Consistency*

2 つのデータセットが互いに矛盾せず，データが互いに一貫している程度で表わす．

- *Accuracy*

データがユーザ，もしくはシステムから要求された精度で識別されたソースと一致する程度で表わす．

- *Accessibility*

データがユーザにとって使いやすい，もしくは簡単に見つけられる程度で表わす．

- *Believability*

データが真実であり，信憑性のあるものだと見なされる程度で表わす．

3.3 Web アプリケーションへの適用

近年，1 つのシステム内において，複数のチャネルを持つマルチチャネルシステムが出現したことや，特にインターネットや web ベースの技術が発達したことが，システム内のデータに対する Data Quality での評価の適用に拍車を掛けている [13]〜[15]．Cappiello らは Web ページの質を評価するために、データベースで適応されている Data Quality の指標を Web ページで適応するために拡張することを提案している [13]．Gregg は Web ページのメタ情報を利用した Data Quality の評価を提案している [16]．Bouzenghoub らはデータの古さに関係する Data Quality について分析している [17]．Velayathan らは Data Quality と直接関係していないが，Web ページの相対信頼度について研究している [18]．

現在，Data quality を評価する指標の具体的なメジャのほとんどが，具体的な問題を解決するためにアドホックに発達している．この問題は定義者の主観性が評価に大きく影響する．その為，Web ページの Data Quality を評価する指標の中には，定型的なメジャの定義がまだ存在しない指標や，中にはメジャが定義できないデータが存在する．[14] 本研究では [13], [16], [17] で提案されている Web ページの Data Quality を評価する Data Quality Dimension の定義から，具体的なメジャを以下のように考え，検討する．なお, [18] で提案している評価方法は一部をそのまま利用する．

- *Timeliness*

ソース内でどのくらい頻繁にデータが変更・更新されるか，又はどのくらい頻繁に新しいデータが生成されるかを集めることで表わす．以上の定義から，本稿では Web ページの更新頻度で評価する．

- *Currency*

データが更新される時からデータが使われる時までの時間間隔の範囲で表わす．以上の定義から，本稿では Web ページの最

終更新時間から，ユーザがその Web ページを閲覧するまでの経過時間で評価する．しかし実際にはユーザが閲覧する時間は予想できないので，スコアの計算時を閲覧する時間に置き換えて計算する．

- *Completeness*

1 つの Web ページが関連する全ての情報を含んでいるかの程度として表わす．もし Web ページがドキュメントを含むなら，ドキュメントのメタデータから質を保証される．もし Web ページが要素のリストを含むなら，データソースを持つリストを比較することで保証される．以上の定義から，本稿では Web ページの情報量と，もしその Web ページが外部リンクを持つなら，その外部リンク先の情報量の和がそのページの内容と関連する情報の総量をどの程度含んでいるかで評価する．

- *Consistency*

データ内における値の一貫性は，そのデータが全てのデータ表現において同じ値を持つことである．

representation consistency（表現の一貫性）は Web ページの一貫性を評価するのに最も利用しやすい．それは与えられたセクション内に置かれたオブジェクトのフォーマットと，期待したフォーマット間における一致の程度として表わす．Web ページ内の文章のフォーマットは html タグや dtd 情報や css のメタ情報によって定義される．以上のことから，本稿では次の 2 つのメジャで検討する．

- css（Cascading Style Sheets）の定義の有無
- dtd（Document Type Definition）の定義の有無．

- *Accuracy*

1 つの Web ページ内のそれぞれの情報オブジェクトと一致する正しいコピーを比較する．これにより値の不一致や誤った結果を識別する．正しいと考えられる他の値 v' と 1 つのデータの値 v の近似のメジャとして表わされる．以上の定義から，本稿では Web ページの内容と関連する Web ページの集合において，その関連する内容の特徴を表わすキーワードとなる単語群が存在する．正しいと考えられる他の値 v' を関連する Web ページの集合と考えて，その Web ページが関連する Web ページ群の単語群のうちどのくらいの単語をその Web ページが含んでいるかの程度を測る．その程度から，その Web ページの内容が正しい情報であるかを評価する．

- *Accessibility*

Web ページはページタイプと主題の領域に関連する具体的なデータを提供しなければならない．ユーザが検索リクエストを行ったときに、自律エージェントがすぐ適用できる一貫したフォーマットを持たなければならない [16]．以上の定義から，本稿では各メタタグの content の要素 keyword，description 内に含まれる単語が，その Web ページの特徴を表わすキーワードと一致しているかの程度で評価する．

- *Believability*

本稿では以下の 3 つのメジャで検討する．

- コンテキスト内における copyright 情報の有無．
- Web ページの内容と関連する Web ページから受ける被リンク数．
- Web ページの *Top Level Domain* が信頼できるドメイン $nTLD, gTLD, iTLD$ に属しているか否か．
  (*.ne.jp*,*.co.jp*,*.jp*,*.ac.jp*,*.com*,*.edu*,*.net*,*.gov*,*.mil*)

4. 提 案 手 法

ここでは，3 章で述べた Data Quality Dimensions の定義とそれを基に提案したメジャからどのように Web ページの重要度を実際に評価するかについて述べる．さらにその評価値を *PPV* アルゴリズムにどのように適用するかについて述べる．

4.1 問 題 定 義

Jeh らの Web ページの選出の自動化の問題点を Chirita らの *PROS* は解決策を提案しているが, 選出した Web ページ集合への重みの割り当て方についてはどちらの提案においても検討されておらず，選出した各 Web ページに対して一定の重みを割り当てている．しかし，ブックマークされている Web ページはユーザの主観で選ばれているので，その中のどの Web ページがユーザにとって本当に満足する Web ページであるかはわからない．さらに *PROS* ではブックマークを入力値として使っているが，その周辺の Web ページや PageRank，*HubRank* のスコアの高いものを選出し，その集合の全てに一定の重みを割り当てている．それではブックマークされている Web ページから得られるユーザの嗜好がぼやけてしまう可能性がある．

我々は，選出した Web ページ集合に対し，客観的な評価をすることが重要と考える．そして，ブックマークされている各 Web ページとその周辺ページに対し，3 章で述べた Data Quality Dimensions の定義とそれを基に提案したメジャを使い，7 つの側面から Web ページ毎の質を評価し，その評価値を Web ページの重要度として u の値とする．このように評価した値をユーザがパラメータで調節することで，ユーザの様々な嗜好をより反映した Personalized PageRank のスコアを求めることを考える．

4.2 Data Quality の評価

- *Timeliness* の計算

定期的に Web サーバに最終更新時間の問い合わせをし，最初の問い合わせからの経過時間を *lapsedtime*，その経過時間内での更新回数を *updatenum* とすると，更新頻度 *updatefreq* は以下の式で求められる．

$$updatefreq = \frac{lapsedtime}{updatenum}$$

更新頻度 *updatefreq* の値を求め，表 1 から *timeDQ* を式 (3) で表す．

$$timeDQ = \{0.0, 0.1, \cdots, 0.9, 1.0\} \quad (3)$$

ただし，Web ページの中には過去の事実など一度公開されたら二度と更新されない Web ページもある．よって，最初に最終更新時間を問い合わせてからの経過時間が 1440 時間を越えたものは，*timeDQ* の指標は考慮しないものとする．

- *Currency* の計算

最終更新時間から *Currency* の計算をするまでの時間を *lctime*

表 1　更新頻度の間隔と評価値
Table 1　Interval and evaluation value of update frequency

1.0 - 6 時間以内	0.9 - 12 時間以内	0.8 - 24 時間以内
0.7 - 48 時間以内	0.6 - 72 時間以内	0.5 - 168 時間以内
0.4 - 336 時間以内	0.3 - 504 時間以内	0.2 - 720 時間以内
0.1 - 1440 時間以内	0.0 - 1440 時間超	

とすると，更新頻度 $updatefreq$ の値と $lctime$ の値から，$Currency$ の値 $currDQ$ を式 (4) で表わす．

$$\begin{cases} currDQ \geqq 0 \text{ のとき}, currDQ = \frac{updatefreq - lctime}{updatefreq} \\ currDQ < 0 \text{ のとき}, currDQ = 0 \end{cases} \quad (4)$$

- $Completeness$ の計算

まず，各 Web ページの特徴を表わす単語をそれぞれ抽出する．各 Web ページのコンテンツにおいて，形態素解析により，その Web ページに含まれる名詞の単語のみを抽出する．抽出した各単語に対して，tf-idf アルゴリズムを使い，その Web ページ内でのその各単語の重要度を数値化する．その計算値の内，数値の高い上位 80 個の単語を，その Web ページの特徴を表わす単語とし，$keywords$ という単語リストで表わす．各 Web ページから外部リンク 1 回で辿れる Web ページをその Web ページの補佐 Web ページ群とし，その補佐 Web ページ内における数値の高い上位 80 個の単語を，その補佐 Web ページ群の特徴を表わす単語とし，$aidwords$ という単語リストで表わす．また，各 Web ページから外部リンク 2 回のリンクで辿れる Web ページをその Web ページの周辺 Web ページ群とし，その周辺 Web ページ全てにおける数値の高い上位 80 個の単語を，その周辺 Web ページ群の特徴を表わす単語とし，$arowords$ という単語リストで表わす．

このとき，$compDQ$ を式 (5) で表わす．

$$compDQ = \frac{\mid arowords \cap (keywords \cup aidwords) \mid}{\mid arowords \mid} \quad (5)$$

- $Consistency$ の計算

各 Web ページのソースから css，dtd を含むタグを抜き出す．このとき，$consDQ$ を式 (6) で表す．

$$\begin{cases} css \text{ と } dtd \text{ の両方含んでいるとき}, consDQ = 1.0 \\ \text{どちらか一方のみ含んでいるとき}, consDQ = 0.5 \\ css \text{ と } dtd \text{ 両方とも含まないとき}, consDQ = 0.0 \end{cases} \quad (6)$$

- $Accuracy$ の計算

正しいと考えられる値 $\vec{v}$ を関連する Web ページ群と考えて，ここでは関連する Web ページ群 $arowords$ という単語リストで表わす．その単語群うちどのくらいの単語をその Web ページがカバーしているかを求める．このとき，$accuDQ$ を式 (7) で表す．

$$accuDQ = \frac{\mid arowords \cap keywords \mid}{\mid arowords \mid} \quad (7)$$

- $Accessibility$ の計算

各 Web ページのソースから，メタ情報の $keywords$，$description$ 部分のタグを抜き出し，その中に含まれる単語をそれぞれ $kwords$，$deswords$ という単語リストで表わす．さらに，各 Web ページのソースからタイトルを含むタグを抜き出し，その中に含まれる単語を $titwords$ という単語リストで表わす．これらの単語が各 Web ページの $keywords$ に含まれているかを求める．このとき，$acceDQ$ を式 (8) で表す．3 つの要素はパラメータ $\alpha_k, \alpha_d, \alpha_t$ で調節できる．

$$\begin{aligned} acceDQ \;=\; & \alpha_k \frac{\mid kwords \cap keywords \mid}{\mid kwords \mid} \\ & + \alpha_d \frac{\mid deswords \cap keywords \mid}{\mid deswords \mid} \\ & + \alpha_t \frac{\mid titwords \cap keywords \mid}{\mid titwords \mid} \\ \alpha_k \;+\; & \alpha_d + \alpha_t = 1 \end{aligned} \quad (8)$$

- $Believability$ の計算

各 Web ページから URL，著作権に関する情報を抜き出す．このとき URL の $TLD(Top\ Level\ Domain)$ や著作権（$copyright$）の情報は信憑性を表わすものと言える．[18] より TLD が $.ne.jp$, $.co.jp$,$.jp$,$.ac.jp$,$.com$,$.edu$,$.net$,$.gov$,$.mil$ ならばその URL は信憑性が高いと言える．この TLD を持つ URL を $beliefurl$ とする．各 Web ページに対して，リンクを張っている Web ページ群をその Web ページの被リンク Web ページ群とし，被リンク Web ページ内における数値の高い上位 80 個の単語を，その被リンク Web ページ群の特徴を表わす単語とし，$inlinkwords$ という単語リストで表わす．被リンク先の Web ページの内，内容が関連しているかを以下の $inlinkvalue$ で表わす．

$$inlinkvalue = \frac{\mid inlinkwords \cap keywords \mid}{\mid keywords \mid}$$

この inlinkvalue の値が 0.2 以上のものを関連しているとみなし，満たしている被リンク先の Web ページの数を $inlinknumber$ とする．このとき，次の 3 つの要素を考える．

- urlparts

$$inlinkparts = \frac{inlinknumber}{\text{被リンク数}}$$

- copyparts

$$\begin{cases} Web \text{ ページが } copyright \text{ を持つとき}, copyparts = 1.0 \\ Web \text{ ページが } copyright \text{ を持たないとき}, copyparts = 0 \end{cases}$$

- inlinkparts

$$\begin{cases} url \text{ が } beliefurl \text{ のとき}, urlparts = 1.0 \\ url \text{ が } beliefurl \text{ でないのとき}, urlparts = 0 \end{cases}$$

これら 3 つの要素から，$beliDQ$ を式 (9) で表わす．3 つの要素はパラメータ $\alpha_u, \alpha_c, \alpha_i$ で調節できる．

$$\begin{aligned} beliDQ = \alpha_u urlparts + \alpha_c copyparts + \alpha_i inlinkparts \\ \alpha_u + \alpha_c + \alpha_i = 1 \end{aligned} \quad (9)$$

- Data Quality の計算

以上の 7 つの Data Quality Dimensions で計算し，その値から各 Web ページの総合的な Data Quality の評価値 $TotalDQ$ を求める．$TotalDQ$ を式 (10) で表わす．7 つの Data Quality Dimensions はパラメータ α_1からα_7 で調節できる．

$$\begin{aligned} TotalDQ = \ & \alpha_1 timeDQ + \alpha_2 cuurDQ + \alpha_3 compDQ \\ & + \alpha_4 consDQ + \alpha_5 accuDQ + \alpha_6 acceDQ \\ & + \alpha_7 beliDQ \qquad (10) \\ & \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6 + \alpha_7 = 1 \end{aligned}$$

求めた TotalDQ の値は $0 \leqq TotalDQ \leqq 1$ である．

4.3 Personalized PageRank への適用

ここでは 4.2 で定義した式で計算した各 Web ページの評価値をユーザがパラメータで調節し，自分の嗜好に合った各 Web ページの重要度を割り当てる．各 Web ページに割り当てた重要度を u の値として利用する．その為には各 Web ページの $TotalDQ$ の値を正規化する必要がある．ここでは正規化と $u(p)$ の入力について述べる．このとき，$bookmarkedpages$ はユーザにブックマークされている Web ページとその Web ページから外部リンク 1 回で辿ることに出来る Web ページを指す．WWW 空間上の各 Web ページ p の重み u と $TotalDQ$ の正規化は式 (11) のようにする．

$$\begin{cases} u(p) = \frac{TotalDQ(p)}{\sum TotalDQ(p)}, & (p = bookmarkedpages) \\ u(p) = 0, & (p \neq bookmarkedpages) \end{cases} \qquad (11)$$

Personalized PageRank の計算アルゴリズムを以下に示す．アルゴリズム内では，まず WWW 空間上の全ての Web ページに (1/WWW 空間上の全ての Web ページ) の重みを初期値として割り当てる．次に関連研究で紹介した式 (1) を適用させる．i 回目の Personalized PageRank の計算値の合計の値 v_i を i+1 回目の Personalized PageRank の計算値の合計の値 v_{i+1} から引いたノルムが閾値を下回るまで再帰計算を続け，閾値を下回ったときの各 Web ページに与えられている値がその Web ページの Personalized PageRank のスコアとなる．

$v_0 \leftarrow S$//初期値の設定

$loop$:

$v_{i+1} \leftarrow (1-c)Av_i + cu$

$diff \leftarrow ||v_{i+1} - v_i||_1$

$while \ \ diff > \epsilon$

5. 評価実験

提案手法のメジャ有効性を示すため，プロトタイプシステムを実装し，評価実験を行った．

5.1 実験環境

本システムに使用した計算機環境は表 2 の通りである．本システムの構成は図 1 に示す．

（1） 実験用 WWW 空間の構築

Wget [19] コマンドで $Web\ Server$ から目的 URL の Web ページソースをディレクトリごと取得する．目的 URL の Web ページ内の外部リンクで繋がっている Web ページへリンクを辿り，辿った先の Web ページソースを取得する．このとき，その Web ページソースが置かれている $Web\ Server$ が異なっていても，サーバの枠を越えて取りに行く．この操作を繰り返し，WWW 空間における目的 URL を中心とした Web グラフのサブグラフを構築する．この部分が図 1 の $Web\ Page\ Source$ である．

表 2　実験に用いた計算機環境

Table 2　Experiment environment

CPU	Intel Pentium(R)4 3.00GHz
メモリ	1.0GB RAM
OS	Windows XP Professional Service Pack 2
開発言語	Perl(5.8.7)

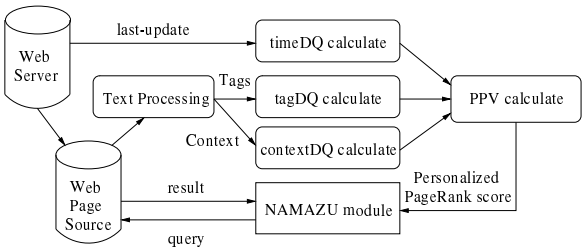

図 1　プロトタイプシステムの構成

Fig. 1　Prototype system architecture

（2） 検索部分の構築

取得した各 Web ページソースを全文検索システム namazu [20] でインデックス付けをする．

（3） Data Quality の計算部分の構築

以下の手順で Data Quality の計算を行う．

（a） *Text Processing*

取得した Web ページソースの URL に非負の整数 ID を付ける．今回の実験では 2250 の Web ページソースを取得した．また，各 Web ページソースを html タグとコンテンツに分ける．

（b） *tagDQ calculate*

各 Web ページソースの html タグから発リンク先の URL を抽出し，その URL が取得ページに含まれるときはその URL に対応する ID を返し，それ以外の URL は新たに ID を付ける．各 Web ページソースの *html* タグからメタ情報，*dtd*，*css* を表わすタグをそれぞれ抽出し，4.2 節で述べた指標で評価する．

（c） *contextDQ calculate*

各 Web ページソースのコンテンツ部分を形態素解析で名詞のみを取り出し，tf-idf アルゴリズムにより，その Web ページでの重要な単語を計算する．その単語からコンテンツの Data Quality を評価値する．

（d） *timeDQ calculate*

各 Web ページの最終更新時間のヘッダー情報を定期的に取得し，そのデータからその Web ページの更新頻度を求める．

（4） Personalized PqageRank の計算部分

(b), (c), (d) で求めた Data Quality の評価値を入力値として，Personalized PageRank アルゴリズムで各 Web ページ

表 3　実験用データ
Table 3　Experiment data

Web ページ数	2250
目的 URL	http://www.townpita.com
ブックマークされている Web ページ数	3
集合 u ページ数	88

表 4　Data Quality のパラメータ
Table 4　Data Quality Parameter

acceDQ	$\alpha_k = \alpha_d = \alpha_t = \frac{1}{3}$
acceDQ	$\alpha_u = \alpha_c = \alpha_i = \frac{1}{3}$
TotalDQ	$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = \alpha_7 = \frac{1}{7}$

の Personalized PageRank のスコアを求め，そのスコアを *namazu module* のソート部分に関連付けする．

5.2 実　　験

スコアの計算にかかる時間が莫大になってしまう．論文投稿に間に合わせるために，今回の実験ではとても小さい WWW 空間を構築し利用した．表 3 は本実験で使用したデータである．このとき 2250 の Web ページの中から 3 つの Web ページを選び，ブックマークされている Web ページとした．さらに，その 3 つの Web ページとその Web ページから 1 回で辿れる Web ページ（以後，隣接 Web ページ）の集合を Data Quality の評価する集合を PPV アルゴリズムの入力値 u とした．この集合 u 内の Web ページ全てを 4.2 節の評価式で評価する．

本実験では 88 個の Web ページを持つ周辺 Web ページ集合 u に対し，各 Dimension の評価値のみで重み付けしたもの，7 つの Dimension 全ての評価値で重み付けしたもの，一定の重み付けをした PPV の 7 パターンの重み付けをし，そのランキングのスコアを計算した．その 7 種類のランキングのスコアから，Data Quality で評価した集合 u の重み付けが，ユーザの嗜好を考慮したランキングに反映されているかについて考察する．さらに，Data Quality の各 Dimension の評価値とランキングスコアの結果から，Data Quality のどの Dimension がランキングに影響するかについて考察する．

今回の実験では集合 u の各 Web ページの Data Quality を計算するときに全てのメジャを同様に評価するために各 Data Quality のパラメータを表 4 のように設定した．以上の設定で 2250 の Web ページの Personalized PageRank のスコアを求めた．このとき周辺 Web ページにジャンプする確率 c=0.15，計算の反復回数を決める閾値 $\epsilon = 0.05$ とした．

5.3 実 験 結 果

PPV，Data Quality の各 Dimension だけで評価したランキングのスコアトップ 30 を表 5 に示す．ここで，周辺 Web ページの集合 u に一定の重みを割り当てたものを PPV，式 (9) で求めた重みを割り当てたものを TDQ，*timeliness*，*completeness*，*consistency*，*accuracy*，*believability* で評価した重みのみを割り当てたものをそれぞれ DQ1, DQ2, DQ3, DQ4, DQ5 とする．

まず，Data Quality の各 Dimension のみ評価したランキング DQ1 から DQ5 の考察をする．

- DQ1(*timeliness* のみ) のランキング

DQ1 では，*timeliness* の評価のみで重みを割り当てた．表 5 から，1，4 の Web ページが高いランキングであると言える．1 の Web ページは全国の都市のお得情報を提供するポータルサイトで，4 の Web ページはグルメのポータルサイトであった．このページは更新が頻繁であるために，ランキングの上位に来たと考えられる．

- DQ2(*completeness* のみ) のランキング

DQ2 では，*completeness* の評価のみで重みを割り当てた．表 5 から，1191，1190，1200，1197，1196，1193，1186，1203，1199，1192，1187，1202 が他のランキングと比べ上位に上がったと考えられる．これらの Web ページは全て同じ音楽サイトの Web ページで，これらのサイト間で相互リンクが張られていることから，*completeness* の評価値が上がり，ランキングの上位に来たと考えられる．

- DQ3(*consistency* のみ) のランキング

DQ3 では，*consistency* の評価のみで重みを割り当てた．表 5 から，PPV とあまり変わらない Web ページが上位を占める結果となった．これは周辺 Web ページの集合のほとんどの Web ページで css，dtd の定義がされていたために，評価値に差が出ず，結果周辺 Web ページの集合に一定の重みを割り当てた PPV と似たランキング結果となった．

- DQ4(*accuracy* のみ) のランキング

DQ4 では，*accuracy* の評価のみで重みを割り当てた．表 5 から，1206，1187，1193，1184，1204，1194，1202，1189，1191，1185，1205 が他のランキングと比べ，上位に来たと考えられる．これは 1184，1187，1193，1194，1204 の Web ページが更新後すぐに *accuracy* の計算を行なったため，こっらの Web ページの評価値が高く，ランキングの上位に上がり，これらの Web ページと元のサイトが同じ Web ページである 1205，1206，1202，1189，1191 のランキングが上がったと考えられる．

- DQ5(*believability* のみ) のランキング

DQ5 では，*believability* の評価のみで重みを割り当てた．表 5 から，PPV とあまり変わらない Web ページが上位を占める結果となった．これは周辺 Web ページの集合内のほとんどの Web ページが信頼性の評価値が低かったためである．その結果，評価値に差が出なかったこが原因だと言える．また，1 は特に評価値が低かった．1 はポータルサイトなので，様々な分野のサイトからリンクを受けている．そのため，関連する単語が膨大になり，*believability* の要素のうち，*inlinkparts* の評価値が低くなったためと言える．

6. お わ り に

本稿では，ユーザの嗜好を表わすデータとして，ユーザにブックマークしているページを利用した Personalized PageRank 手法を提案した．提案手法では Web ページに対し，データのユーザにとっての使いやすさを評価する Data Quality の複数の指標で Web ページの重要度を求め，その評価値 u をユーザの嗜好の値とした．

実験では，小規模な WWW 空間において，周辺 Web ページ

表 5　PageRank，PPV，Data Quality の各指標でのランキングのスコアトップ 30

Ranking	PPV	TDQ	DQ1	DQ2	DQ3	DQ4	DQ5
1	1	1	1	1	1	1	2
2	4	4	4	4	4	4	4
3	66	66	66	2	66	66	1
4	2	2	2	66	2	2	66
5	3	3	3	3	3	3	3
6	162	162	162	162	162	162	72
7	161	161	161	161	160	161	160
8	160	160	160	160	703	160	162
9	703	703	703	703	78	1206	161
10	196	196	196	196	196	1187	416
11	1503	1503	1503	78	1503	703	415
12	68	409	409	1503	409	1193	413
13	409	5	78	281	5	1184	414
14	78	78	5	409	376	1204	418
15	5	376	376	5	411	196	417
16	376	411	411	376	281	1194	419
17	411	281	410	411	410	1503	68
18	410	410	82	1191	82	281	86
19	82	68	281	1190	394	409	78
20	394	82	394	1200	72	5	703
21	72	394	72	1197	79	376	82
22	79	72	79	1196	68	411	376
23	81	79	68	1193	81	1202	281
24	73	81	81	1186	73	1189	71
25	281	73	73	1203	80	410	69
26	80	80	80	1199	69	1191	73
27	69	69	69	1192	67	1185	74
28	67	67	67	410	88	82	67
29	88	88	88	1187	74	1205	412
30	74	74	74	1202	1624	394	75

の集合内の各 Web ページソースの html タグとコンテンツのそれぞれから，ユーザにブックマークされている各 Web ページの重要度を評価し，入力値 u とした．Data Quality で評価した u を Personalized PageRank のアルゴリズム PPV にユーザの嗜好を表わすデータとして適用させた．さらに u 内の Web ページの重みを一定にして計算したランキングスコア，7 つの指標のうち，*timeliness*，*completeness*，*consistency*，*accuracy*，*believability* のみで u を評価し，計算したランキングスコア，7 つ全てで u を評価し，計算したランキングスコアでランキングの変動を比較した．比較結果から，本稿で提案した手法のうち，*timeliness*，*completeness*，*accuracy* のみで u を評価し，計算したランキングでは，提案したメジャ，評価式の妥当性を確認できた．しかし，*consistency*，*believability* のみで u を評価し，計算したランキングでは，そのメジャ，評価式の十分な妥当性を確認できなかった．よって本実験では，Data Quality の各 Dimension で Web ページを評価することによって，その Dimension の質を満たした Web ページの重要度を判定することが出来る Dimension があることが明らかになった．

今後の課題として以下が考えられる．本研究で提案した手法の有効性を示す実験を行うことが必要である．本実験で構築した WWW 空間はとても小規模であった．より実世界に近い環境での実験が必要であることから，より大きな WWW 空間で実験することが必要である．そして綿密な実験から，よりユーザの嗜好を考慮したランキングのスコアを計算できるような Data Quality の各 Dimeonsin の具体的なメジャ，評価式を定義することが必要である．さらに現在検討している Data Quality の他の Dimension についても，Web ページを評価する指標に適用させることが必要である．今後，各 Dimension の適切な評価ができるようになれば，ユーザがどの嗜好を考慮したいかをパラメータで調節することで，ユーザの様々な側面の嗜好をより考慮した Web ページの重要度を判定できると考える．

文　献

[1] L. Page, S. Brin, R. Motwani and T. Winograd: “The pagerank citation ranking: Bringing order to the web”, Technical report, Stanford Digital Library Technologies Project (1999). Available at http://dbpubs.stanford.edu/pub/1999-66.

[2] T. H. Haveliwala: “Efficient computation of pageRank”, Technical Report 1999–31, Stanford U. (1999).

[3] T. H. Haveliwala: “Topic-sensitive pagerank”, Proceedings of the Eleventh International World Wide Web Conference, Honolulu, Hawaii, pp. 517–526 (2002).

[4] G. Jeh and J. Widom: “Scaling personalized web search”, Proc. WWW, pp. 271–279 (2003).

[5] P.-A. Chirita, D. Olmedilla and W. Nejdl: “PROS: A personalized ranking platform for web search”, Proc. AH, Vol. 3137 of Lecture Notes in Computer Science, Springer, pp. 34–43 (2004).

[6] T. C. Redman: “Data Quality for the Information Age”, Artech House (1996).

[7] G. K. Tayi and D. P. Ballou: “Examining data quality”, CACM, **41**, 2, pp. 54–57 (1998).

[8] R. Y. Wang: “A product perspective on total data quality management”, CACM, **41**, 2, pp. 58–65 (1998).

[9] D. Kaplan, R. Krishnan, R. Padman and J. Peters: “Assessing data quality in accounting information systems”, Communications of the ACM, **41**, 2, pp. 72–78 (1998).

[10] T. C. Redman: “The impact of poor data quality on the typical enterprise”, CACM, **41**, 2, pp. 79–82 (1998).

[11] “Yahoo!”. http://www.yahoo.com/.

[12] “Open directory project”. http://dmoz.org/.

[13] C. Cappiello, C. Francalanci and B. Pernici: “Preserving web sites: A data quality approach”, Proceedings of the 2003 International Conference on Information Quality, pp. 331–343 (2003).

[14] C. Cappiello, C. Francalanci and B. Pernici: “Data quality assessment from the users perspective”, Proc. IQIS, pp. 68–73 (2004).

[15] C. Cappiello, C. Francalanci and B. Pernici: “Time related factors of data accuracy, completeness, and currency in multi-channel information systems”, CAiSE Short Paper Proceedings (2003).

[16] D. G. Gregg: “Using Distributed Meta-information Systems to Maintain Web Data Quality”, Doctoral dissertation, Arizona State University (2000).

[17] M. Bouzeghoub and V. Peralta: “A framework for analysis of data freshness”, Proc. 2004 Int. Workshop on Information Quality in Information Systems, pp. 59–67 (2004).

[18] ヴェラヤサン, 山田：“Web ページの相対信頼度”, 第 18 回人工知能学会全国大会, pp. 3F1–05 (2004).

[19] “Gnu wget”. http://www.gnu.org/software/wget/.

[20] “全文検索システム namazu”. http://www.namazu.org/.