リンク参照元ページ群の話題多様性に基づく被参照ページの品質推定

荘司　慶行†　　田中　克己†

† 京都大学 大学院 情報学研究科 社会情報学専攻〒 606–8501 京都府 京都市 左京区 吉田本町
E-mail: †{shoji,tanka}@dl.kuis.kyoto-u.ac.jp

あらまし　現在広く使われている PageRank は，ウェブページを人気度に基づいてランキングする．人気度は，“人気のあるページから多くリンクされたページは人気である” という仮説に基づき計算される．この際，PageRank はリンクの本数だけに注目し，リンク元ページの内容を考慮しない．しかし，ある偏った層からのみ支持されるページと，幅広く多様な層に支持されたページでは，人気の性質が異なると考えられる．本論文では，あるページを複数のページ群がリンクしていた際，リンク元のそれぞれのページ中に出現する語のトピックがどれだけ多様であるかによってリンク先ページの内容の専門性を推定可能な事を明らかにし，専門度に基づく検索ランキングについて評価を行う．

キーワード　情報検索，多様性，専門性

1. はじめに

コンピュータの低価格化やスマートフォンの普及に伴い，ウェブは一部の限られた専門家のためのものではなく，より多様な人々が利用するものになりつつある．また，ウェブ上のリソースの内容とその用途についても，従来のコンテンツや文書だけにとどまらず，コミュニケーションや商業など多岐にわたるようになりつつある．このような状況の下で，専門的な知識を有する利用者もそうでない一般の利用者も，それぞれが望むコンテンツを探す必要が生じており，より柔軟な検索エンジンが求められている．

現在広く用いられている検索エンジン Google では，ウェブページを PageRank アルゴリズムによってランキングしている．PageRank では，ウェブページから他のウェブページへのリンクをソーシャルな投票と見なし，ウェブページ間のリンクを解析することで，それぞれのウェブページの人気度を算出している．人気度の高いウェブページは重要であると考えられ，ランキングの上位に配置される．人気度は，“人気のあるページから多くリンクされたページは人気である” という仮説に基づき計算される．

しかしながら PageRank では，あるページを複数のページが参照していた際，リンクの本数に関しては考慮するが，そのリンク元ページ群がどのような内容かを考慮しない．そのため人気度の高いページが，誰に人気なのか，どう人気なのかが考慮されておらず，ランキングには特定の層に人気なページと，幅広く人気なページが入り混じった状態で出現する．例として，“協調フィルタリング” という，情報学分野における手法名をクエリとした検索結果を図 1 に示す．この検索結果には，“協調フィルタリング” の定義や解説などの情報学に詳しくない人向けの入門的な記事と，学術論文や近年の動向など，情報学に詳しい人向けの記事が含まれる．利用者が専門的な知識を有する人であった場合，彼らが欲しいのは専門的なページであり，検索結果中に含まれる入門的な記事は彼らの検索意図を満足させるものではない．一方で，情報学についてまったく知識のない

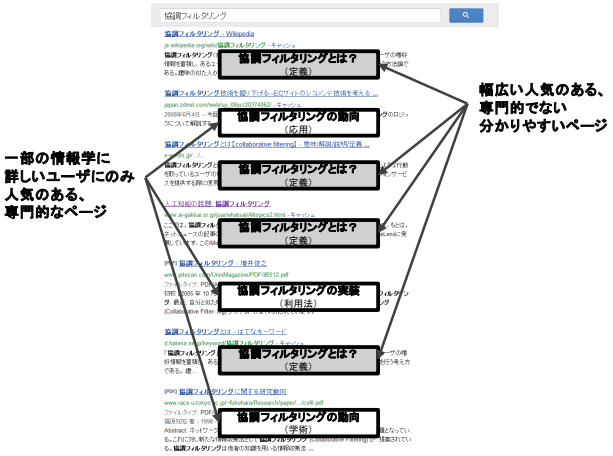

図 1　従来の人気度に基づく Web 検索ランキング

利用者にとっては，専門的な記事は読んでも理解できないため，不必要なページである．このように，人気度が高いページが常にだれにとっても有用であるわけではなく，そのページがどのような層に対して人気なページであるかを考慮することが重要である．

本研究では，ある文書を参照する複数の文書について，それぞれがどの程度独立で多様であるかを計算することで，参照先の文書の性質を判定する手法について提案する．複数の文書があった際，それらがどれだけ多様であるかは，様々な要素によって定義可能である．例として，

- ページのトピック
- センチメント
- 意見の肯定否定
- 著者の年齢などメタデータ
- 地理的情報
- 時間的情報

等があり，これらに基づく多様性を計算することで，“賛否両論ある政策と，誰もが賛同する政策（意見の肯定否定の多様性）” や，`` 北海道民にも沖縄県民にも受け入れられるレシピと，関

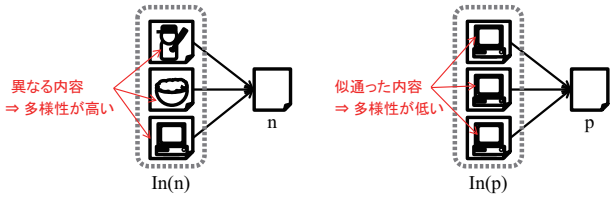

図 2　ページのトピックと参照構造

東圏でしか受け入れられえないレシピ（地理的情報に基づく多様性）’’ など，様々な性質のコンテンツが検索可能になると考えられる．

本論文では，これらのうち，ある文書をリンクしている複数の文書についてそのコンテンツの内容（ページのトピック）の多様性に限定し，実験を行う．本研究では，基本的な仮説として，**多様な内容の文書群から参照された文書は万人向けの文書であり，画一的な内容の文書群から参照された文書は限定された読者向け**であると考える．例として，図 2 のように，参照されている文書の個数が等しく人気度の等しい 2 つのページページ p とページ n があった場合を考える．図中の各文書の図柄をその文書の扱っているトピックとすると，ページ p は 3 件のページか参照されているが，参照しているそれぞれのページは，コンピュータに関するページである．あるトピックに関するページの著者はそのトピックに興味を持った人物であり，リンク元ページの著者はリンク先ページの内容を読んだ上でリンクを貼ると考えると，ページ p は，コンピュータに興味を持つ人だけにとりわけ強く興味をもたれる内容のページであると推測できる．逆に，ページ n はリンク元のそれぞれのページが異なる内容である．この場合，ページ n は幅広い読者に支持されているページであり，その内容は万人受けするものであると推測できる．このように，同じ本数のリンクが貼られているページであっても，その参照元のページがどれだけ多様であるかによって人気の質は異なるものであると考えられ，とりわけページの内容に注目した場合にはその文書がどれだけ万人受けする内容であるかに差が出ると考えられる．

本稿では，

- 別々の内容を取り扱っている多様な集団からリンクされたコンテンツは万人受けする内容を含む
- 同じ内容を取り扱っているページ群から偏ってリンクされるコンテンツは専門的な内容を含む

という仮説に基づき，文書の専門性に基づくウェブ検索ランキング手法を提案し，実際に作成したランキングについて評価実験を行うことで参照元ページの多様性が参照先ページの専門性へ関係するかを明らかにする．

以下に，本論文の構成を記す．2. 節では関連する研究について，3. 節では提案する手法についてそれぞれ述べ，4. 節では提案手法を評価する実験について述べる．

2. 関 連 研 究

本研究は，ウェブページ間のリンク構造に着目し，リンク元ページの多様性を考慮することでページのランキングを行い，専門性に基づいた検索結果を作成することを目的とする．本章では，これらに関連する先行した研究として，2.1 項では情報の多様性について，2.2 項ではリンクに基づく他のランキングアルゴリズムについて，2.3 項では情報の専門性についてそれぞれ記す．

2.1 情報の多様性

情報科学の分野では，近年多様性（diversity）を対象とする研究が活発になってきている．特に，Web2.0 と呼ばれる双方向性をもったインターネット利用が盛んになるに伴い，集合知という概念が注目を集めている．Surowiecki [12] は自著の中で，多数の人間の意見が有用な集合知として機能する “群衆の英知” の成立条件として，集合知的データの 3 つの特徴

- 意見の多様性：Diversity of opinion
- 独立性：Independence
- 権力不在性：Decentralization

を挙げており，またこれら 3 つの条件の満たすデータがあった際，それを集約する適切なアルゴリズムとして

- 集約性：Aggregation

が必要であると述べている．近年のウェブは多種多様な利用者による情報の集積体であり，ページ間リンクもその一部である．そのため PageRank をはじめとするリンクを用いたランキングアルゴリズムにおいても，多様線の観点を持ち込むことは有効であると考えられる．

情報検索分野においても，多様性に関する研究は盛んである．もっとも活発な研究の一つとして，検索結果の多様化が挙げられる [15] [2]．情報検索における検索結果の多様化の研究では，表示件数の限られた検索結果を出力する際に，ランキングの上位に重複した内容のページが出現することを避け，なるべくそれぞれ異なるページを含む検索結果を作成する．例として，あるクエリに対する検索結果の上位が全て同じサブトピックに関するページであったり，また特定のサービスやウェブサイト内のページである場合，表示される検索結果のどれを選択してもほぼ同じ情報しか得られないものになる．Carbonell ら [3] らは多文書要約の手法 MMR によって検索結果の多様化を行っている．ランキングを上位から決定してゆく際，既にランキングに登場したページと類似した内容のページのスコアを下げることによって，ランキングに登場するそれぞれのページの内容に多様性を持たせている．

Minack ら [9] は，検索結果の多様化には以下の 2 つの段階があると指摘している．

- クエリの曖昧性による多様性
- 情報ソースによる多様性

前者は，クエリ自体が曖昧性を含むため，検索結果に含まれるページのトピックを多様化する必要が生じる場合を指す．複数の意味を持つクエリが入力された場合，検索結果はクエリのモツそれぞれ別の意味に関するページ 1 つずつが並ぶ方が，多様性の高い検索結果となる． 後者は，クエリ自体は明確だが，そのトピックについて複数の情報ソースが存在する場合に情報ソースを多様化することを指す．この場合，それぞれ別のウェブサイトのページや，異なった種類のコンテンツを検索結果の

上位に登場した方が，多様性の高い結果と言える．本研究においては，検索結果内のページではなく，あるページを参照するページ群について多様性を算出するため，根本的にこれらの研究とは異なる．

情報学分野以外に目を向けると，多様性に関する議論は，社会学，経済学，生命科学分野などで，古くからなされている．Stirling ら [11] は，多様性において以下の 3 つを重要な要素として挙げている．

- Variety
- Balance
- Disparity

例として，ある領域に複数のカテゴリに属するオブジェクトが多数存在する場合を考える．Variety はその領域内にあるカテゴリの数を指し，領域内のカテゴリの数が多い方がよりその領域のもつ多様性が高いと見なす．Balance はそれぞれのカテゴリに属すオブジェクトの数のばらつきを指し，それぞれのカテゴリに属すオブジェクトの数がどのカテゴリも等しいほど多様性が高いと見なす．Disparity は，それぞれのカテゴリがどれだけ綺麗に分化されているかを指し，カテゴリ内のオブジェクトが類似し，別カテゴリのオブジェクトが類似しない場合に多様性が高いと見なす．本研究における多様性は，主に Variety の考え方に基づいて計算される．

2.2 リンク構造に基づくウェブページのランキング

文書をランキングする際，本研究と同様に，文書内の単語ではなく文書間のリンクに基づいた計算を行う手法は多い．代表的な手法として PageRank が挙げられる．PageRank ではページ間のリンクを投票と見なすことで，人気度に基づくランキングを作成する．このアルゴリズムでは，ページ閲覧者をランダムサーファーとしてモデル化し，リンクをたどり隣り合った文書に移動する Random Walk，グラフ中のどれかの文書に移動する Random Jump をさせることで，それぞれのページでの存在確率を計算し人気度として扱う．PageRank をトピックを考慮した形に拡張したアルゴリズムとして，topic sensitive PageRank [6] がある．topic sensitive PageRank では，クエリと文書のトピックを考慮し，ランダムサーファーが Random Jump を行う際，グラフ全体ではなくそのトピックに関連するページにジャンプする．また，TrustRank [5] は PageRank のアイデアをスパム発見に利用している．TrustRank では，スパムページは良質なページと悪質なページの両方にリンクを貼るのに対して，良質なページは良質なページにしかリンクを貼らないという仮説に則り，特定の信頼できるページ群のみに Random Jump を行うランダムサーファーを用いることで，良質なページを発見する．また，Takahashi [13] らは，PageRank に時空間要素を加えることで，歴史的イベントのインパクトを算出している．

PageRank とは別のリンク解析に基づく手法として，HITS アルゴリズム [7] がある．HITS アルゴリズムでは，ウェブグラフを 2 部グラフとして見なし収束計算を行うことでページの重要度を計算する．HITS アルゴリズムは Hubs と Authorities というお互いにお互いを決定しあう 2 つの値を取り扱う．Authorities はあるページがどれだけ重要な情報を含むかを表し，そのページがどれだけ Hubs の高いページにリンクされているかによって計算される．Hubs は優れたリンク集や検索結果ページなど，そのページがどれだけ重要な情報を発信しているページへ案内しているかを表す値である．HITS アルゴリズムに基づくアルゴリズムとして，Generalized Co-HITS アルゴリズム [4] がある．これは HITS アルゴリズムをウェブリンクからより一般的な 2 部グラフに拡張したものであり，論文とその著者などに適用可能である．その他の HITS アルゴリズムの拡張として，SALSA [8]，Fast random walk with restart [14] などが存在する．

これらの従来のリンク解析に基づくアルゴリズムでは，グラフ中のリンクの本数や流量を主眼とした計算を行っており，それぞれのリンクがどれだけ独立であるかや，リンク元のノードがどれだけ多様であるかについては考慮していない．

2.3 情報の専門性

本研究では，多様な読者から支持を受ける文書を万人向けする文書とみなしたランキングアルゴリズムを提案している．これは従来行われている専門性の研究と近い．すなわち，万人向けする文書は専門性が低く，逆に狭い層に熱狂的に支持される文書を専門性が高いということができる．本手法では，リンク解析的なアプローチに基づいてコンテンツの性質を推定したが，一方で本文を解析することで文書の専門性を推定する研究が多数行われている．中でも，文書中に出現するそれぞれの語の専門性を推定する研究が盛んである．

自然言語処理の分野では文書から専門語を自動抽出する研究がなされている．合原ら [16] は医療生物分野において，機械学習を用いてコーパスからの専門語の自動抽出，分類を行っている．中川ら [17] はコーパス中の連接名詞を HITS アルゴリズムでスコアリングする FLR 法を提案している．Nakatani ら [10] は Wikipedia のカテゴリ情報を用いて，リンク解析的に専門語の抽出を行っている．これらの研究ではある特定分野の大規模コーパスや Wikipedia のカテゴリ情報を用いることで，語の専門度を判定している．本手法では，一般的なウェブ全体を対象とし，ページ単位での専門性を推定するためこれらの手法とは異なるが，これらの手法は相補的に適用可能である．

3. リンク参照元ページ群の話題多様性に基づく被参照ページの品質推定手法

与えられた文書群がどれだけ多様であるかについての一般的な考え方と，ウェブページの専門性を測る上で機能を特化させた 2 種類の拡張手法について説明する．

3.1 文書の特徴ベクトル化

はじめに，$In(n)$ 内の各文書を，それぞれを特徴づけるベクトルで表す．本手法では，リンク元文書のトピックがどれだけ異なっているかを多様性として取り扱うため，それぞれの文書中に出現する語から特徴ベクトルを作成する．語の出現頻度を文書の特徴ベクトルとして扱う際，全データセットに登場する全ての語を対象とすると，次元数が膨大になる．そこで，それぞれの語について，トピックモデルに基づき i 種類のトピック

のどれかに属す語であるとする．文書中に含まれる各語について，どのトピックに属すかを LDA を用いて推定する．各文書に，各トピックに属す単語が何度出現しているかを，文書の特徴ベクトルとして用いる．これにより，全ての文書を i 次元のベクトルで表すことが出来る．データセットに含まれるそれぞれの文書は長さが異なるため，特徴ベクトルをノルムで正規化する．

なお，文書群があった際，多様性は文書のトピックに限らず，否定や肯定の度合い，個人の立場，センチメントなど，様々な観点に基づき定義可能である [9]．本稿における手法では，文書群の多様性を，それぞれの文書の扱っているトピックに基づいて算出するため，文中語のトピックから特徴ベクトルを生成したが，文書を表す特徴ベクトルをどう作成するかによって，異なる観点に基づく多様性を算出可能であると考えられる．

3.2 参照元ページの多様性の計算

複数の文書 $In(n)$ があった際，$In(n)$ がどれだけ多様であるかを表す値 $d(In(n))$ を求める．$d(In(n))$ は文書 n を参照している文書群を指し，$d(In(n))$ の値が高ければ文書 n は幅広い文書から支持される文書であり専門性が低く，逆に $d(In(n))$ の値が低ければ文書 n は限られた層からのみ支持される専門性の高い文書であると見なす．

すべての個別の文書が内容語のトピックに基づく特徴ベクトルで表現される際，ある文書 $n = (n_1, n_2, n_3, n_4, \ldots, n_k)$ にリンクした文書群 In(n) がどの程度多様であるかを，以下の式で定義する．

$$d(In(n)) = \frac{1}{|In(n)|} \sum_{p \in In(n)} dist(p, mean(In(n))) \qquad (1)$$

ここで，mean(In(n)) は In(n) の平均ベクトルを，dist(a,b) はベクトル間の距離をそれぞれ表す．mean(In(n)) は，ベクトル集合が与えられた際，それらの平均ベクトルを返す関数である．すなわち，

$$mean(In(n)) = \frac{1}{|In(n)|} \sum_{p \in In(n)} p \qquad (2)$$

である．$dist(a, b)$ は任意の 2 ベクトル間の距離である．本手法ではユークリッド距離に基づいて，任意のページの特徴ベクトル間の差を計算した．それぞれのベクトルが $a = (a_1, a_2, a_3, \ldots, a_k)$, $b = (b_1, b_2, b_3, \ldots, b_k))$ と表される際，$dist(a, b)$ を以下のように計算する．

$$dist(a, b) = \sqrt{\sum_{i=1}^{k} (a_i - b_i)^2} \qquad (3)$$

本手法における $In(n)$ における多様性 $d(In(n))$ は，平均ベクトルからの差の総和を要素数で正規化したものである．これはスカラーにおける分散と同様の性質をもつ値であり，$In(n)$ 内の各文書のばらつきが大きいほど値は大きくなる．値の範囲は，各文書を表す特徴ベクトルのノルムが 1 に正規化されている場合 [0,1] である．$d(In(n))$ の値は，$In(n)$ 中の全文書の特徴ベクトルがまったく等しい場合に最大の 1 となり，逆に $In(n)$ 中の全文書の特徴ベクトルがまったく異なる場合に最低の 0 をとる．ここで，文書がどれだけ万人向けかに基づくランキングを作成する場合には文書 n のスコアを多様性の高さ $d(In(n))$ とし，逆に専門性に基づくランキングを作成する場合には文書 n のスコアを多様性の低さから

$$\mathrm{u(In(n))} = 1 - \mathrm{d(In(n))} \qquad (4)$$

とする．

3.3 多様性スコアの伝播を考慮した拡張

ここで，提案した多様性算出手法をもとに，文書の対象とする読者層の広さの推定に特化した拡張を考える．一部の人向けに書かれた文書群からリンクされた文書と，一般向けに書かれたから参照された文書では，文書の性質が違うと考える．文書の内容とリンクの関係については Akamatsu らの研究 [1] をはじめ議論されており，本研究の対象とする文書の専門性もリンクにより伝播する可能性がある．そこで，文書の専門性がリンク元文書の多様性によって推測可能であるとしたうえで，以下のような 4 つの場合を考える．

- 万人向け 文書から 多様に 参照されている
- 万人向け 文書から 画一的に 参照されている
- 専門的 文書から 多様に 参照されている
- 専門的 文書から 画一的に 参照されている

ここで万人向け文書から多様に参照されている度合い $dd(In(n))$ について，3.2 項の手法をもとに以下のように定義する．

$$dd(In(n)) = d(In(n)) \sum_{p \in In(n)} \frac{d(In(p))}{|In(n)|} \qquad (5)$$

ページ n を参照する文書群 $In(n)$ に含まれるページ $p \in In(n)$ について，3.2 項の手法でそれぞれの専門性を算出し，それぞれを足し合わせ，参照元文書の個数で正規化を行ったものを，参照元ページからの伝播として扱う．また，ページ n について，参照元文書群 $In(n)$ の多様性を考慮し，伝播した値と掛け合わせたものを万人向け文書から多様に参照されている度合いとして扱う．

同様に他の場合についても，万人向け文書から画一的に参照されている度合い $du(In(n))$ を

$$du(In(n)) = u(In(n)) \sum_{p \in In(n)} \frac{d(In(p))}{|In(n)|} \qquad (6)$$

専門的 文書から 多様に 参照されている度合い $ud(In(n))$ を

$$ud(In(n)) = d(In(n)) \sum_{p \in In(n)} \frac{u(In(p))}{|In(n)|} \qquad (7)$$

専門的 文書から 画一的に 参照されている度合い $uu(In(n))$ を

$$dd(In(n)) = u(In(n)) \sum_{p \in In(n)} \frac{u(In(p))}{|In(n)|} \qquad (8)$$

とそれぞれ定義する．

本拡張手法では，それぞれのページの参照元について，1 階層前だけを対象として拡張を行っている．この手法は再帰的に適用することが可能である．また，各文書から伝播する値に，文書長や人気度に基づく重みを設定することが考えられる．

3.4 参照元文書と参照先文書のトピックを考慮した拡張

また別の拡張として，リンク元ページとリンク先ページのトピックの近さに基づく手法が考えられる．3.2 項で提案した基本的な手法では専門性を推定する際，参照元の文書群の多様性の低さのみに注目していた．しかしながら，例として “コンピュータ” に関する文書があり，その文書を参照している文書が画一的で似通った内容であったとしても，それらの文書のトピックが “スポーツ” など “コンピュータ” とかけ離れている場合がある．このような場合，この文書はスポーツの分野で専門的な文書であるかもしれないが，コンピュータの分野で専門的な文書ではない可能性が高い．専門性の高い文書を探す際には，参照元文書の内容の多様性に加えて，参照元の文書が同じ領域のトピックに言及する文書であるかを考慮する必要がある．

3.2 項の手法をもとに，参照元文書群 $In(n)$ の多様性が低く，なおかつそれらの文書のトピックがどれだけ参照先文書 n と近いかを表す値 $tu(In(n))$ を以下のように定義する．

$$tu(In(n)) = 1 - \frac{1}{|In(n)|} \sum_{p \in In(n)} dist(p, n)) \qquad (9)$$

ここで，$dist(p, n)$ は先ほどと同じく，任意の 2 つのベクトルのユークリッド距離である．この拡張手法では，参照元の文書 $In(n)$ に含まれるそれぞれの文書と参照先の文書 n の距離の総和を文書数で正規化したものを文書のスコアとして扱う．

4. 評 価 実 験

3. 項で提案したそれぞれの手法について，実際に作成されたランキングがどれだけ多様な人に受け入れられるものになっているかどうか，実際のウェブページからなるテストセットを用いて，評価実験を行った．

予め用意された 10 件のクエリについて適合する文書を Web 検索エンジンを用いて取得し，またそのページを参照しているページを収集した．これらの収集したウェブページについて，検索エンジンの人気度ベースのランキングでの上位 30 件のページについて人手で各ページのページトピックがどれだけ読者層を限定するものであるか，ページの記述内容がどれだけ専門的であるかについて評価し，正解セットを作成した．このデータセットについて本論文で提案した各種法を用いて並び替えを行い，それぞれの手法を正解と照らし合わせることで評価を行った．

評価を行ったランキングは以下の 7 つの手法によるものである．3.2 項における手法で算出した

- 式 1 による多様性の高さ
- 式 4 による多様性の低さ

に基づくランキング，および，3.3 項においてリンクによるスコアの伝播を考慮した手法で算出した，

- 式 5 による万人向け文書から多様に参照されている度合い $du(In(n))$
- 式 6 による万人向け文書から画一的に参照されている度合い $du(In(n))$
- 式 7 による専門的な文書から多様に参照されている度合い $du(In(n))$
- 式 8 による専門的な文書から画一的に参照されている度合い $du(In(n))$

に基づく 4 つのランキング，3.4 項において提案した参照元と参照先のトピックの違いを考慮した

- 式 9 によるトピックを考慮した多様性の低さ $tu(In(n))$

およびベースラインとして

- 人気度（Web 検索結果の順位）

である．

本実験で対象とするデータセットは，Yahoo Japan Web 検索 API (注1) によって収集した．この API で得られる検索結果ランキングは Google のものと等しく，また各文書の参照元文書についても検索可能である．

各文書からのトピック推定には LDA のオープンソースの実装である GibbsLDA++を利用した．LDA のモデルの推定は，各クエリごとに行った．検索クエリで直接適合した文書と，それを参照する文書について MeCab を用いて形態素解析し，ストップワードを取り除いた上で使用した．この際，トピックの次元数は 100，gibbs サンプリングの繰り返し回数は 2000 回とした．

人手での正解セットの作成は，評価用に実装したウェブアプリケーション上で行われた．各評価者はウェブブラウザで，クエリとそれに関するウェブページの一覧を閲覧し，各ページの内容をスニペットおよびページ本文を確認したうえで評価した．評価項目はそれぞれ 5 段階で評価された．クエリはあらかじめ用意された 10 件を利用し，また検索結果はクエリごとに 30 件を評価の対象とした．正解セットは情報検索に関して十分な知識を持つ学生 1 名によって作成された．

実験に利用したクエリの一覧を表 1 に示す．それぞれの手法は，作成された正解データをもとに，適合率に基づく評価指標 $p@k$，ランキングの評価指標である MAP および $nDCG$ で評価された．

4.1 結　　果

全クエリを通しての各ページのとりあつかったトピックに基づく，対象とする読者層の狭さについての結果を表 2 に示

クエリ
京都 観光
AKB48
日露戦争
モーツァルト
原油
ニート
ハリーポッターと死の秘宝
大気汚染
フェリス女学院
チョコレート

表 1　実験クエリ

(注1)：http://developer.yahoo.co.jp/webapi/search/websearch/v2/websearch.html

す．ここで，map はデータセット 30 位までの Mean Average Precision を，p@K はそれぞれ K 位までの適合率を，nDCG は 30 位までの nDCG を表す．この際，各評価項目に関する適合度の判定について，5 段階で評価したうち 3 より大きいものを適合，そうでないものを不適合として扱った．

同様に，全クエリを通しての各ページの記述に基づく，対象とする読者層の狭さについての結果を表 3 に示す．

5. 考　　察

最も万人受けする順にランキングを作成したのは検索エンジンの返した人気度に基づくランキング手法であり，逆に最も万人受けしない順のランキングを作成したのは 3.2 におけるもっとも基本的な提案手法であった．

1 章で述べた仮説とは逆に，ページのトピック，内容の両評価項目において，参照元の多様性が高かった場合に万人受けしないという結果が出た．これは直観に反するが，実際のデータセットの中で万人受けすると評価されたページの多くは Wikipedia やオンライン辞書サイトのクエリに関する記事，ならびに機関の公式サイトのトップページなどが多かった．これらのサイトでは，外部のサイトからのリンクも多く集めるが，それ以上に自サイト内でのリンクが多く，同一サイト内であるため使用されている語彙が近かったため非参照元の内容が多様でないと判定されたことが一因である．一方で，評価者によって専門性が高く万人受けしないと評価されたサイトの多くは，個人の blog であったり，クエリから派生した語句に関する辞典，ニュースの記事やウェブサイトであった．これはクエリ自体の曖昧性により，粒度の異なる情報の万人受けを同時に評価したため起こった問題である．例としてクエリ「チョコレート」のタスクでは，チョコレートの定義に関するページは万人受けすると評価されたが，一般的なチョコレートの作り方のページや，個別のチョコレートメーカの公式サイトはチョコレートのサブカテゴリであるため，万人受けする内容ではないと評価された．

実際の例として，クエリ「ニート」の場合の正解セットと各手法におけるスコアについて表 4 に示す．各項目はそれぞれ，TS はトピックの専門性の高さ，CS はコンテンツの専門性の高さ，Div は 3.2 節における提案手法，DD, DU, UU, UD はそれぞれ 3.3 節における各手法，TOPIC は 3.4 節における手法，RANK は Yahoo! Japan Web 検索 API の検索順位，#P はそのページを参照している文書の数を表す．

評価者が万人受けするトピックではないと判定した上位 5 つのページは特定の誰かの blog，個別の事象に関して相談する内容の QA サイトのコンテンツ，ニートを題材とするマイナー映画の公式サイト，および特定の小さなニュースのニュース記事であった．この中で最も単純にリンク元の多様性が高かったのは映画の公式サイトであった．このサイトには個人 blog や映画を紹介するニュース等からリンクが貼られていたためスコアが高くなった．

逆に，日本語俗語辞典に含まれるコンテンツは，辞典形式で万人受けする内容であると判定されたにもかかわらず，各種の多様性に基づくスコアが低く算出された．これは辞典サイトは関連語などへのリンクを文中にたくさんふくむため同一サイト内でのリンクが多く，さらにクエリと対応する本文部分が定義の記述だけで短く，サイト内で共通したテンプレート部分の比率が大きいため，類似した内容のページから多数リンクされていると見なされたためである．

以上のように，実験の結果，仮説に反して参照元コンテンツの内容が多様であっても，各タスクにおけるクエリの粒度やページの性質などにより一様にページが万人受けするかどうかを推定することは難しいことが分かった．

6. まとめと今後の課題

本研究では，文書を複数の文書が参照していた際，参照元の文書群がどれだけ多様であるかにより文書の性質が推定可能であるとし，参照元文書群の多様性に基づく文書のランキング手法を提案した．また実際に，参照元のそれぞれの文書の本文中に登場する語がどれだけ多様であるかに着目し，ウェブページを文書をその文書がどれだけ万人向けの内容であるかによってランキングする手法について提案し，実際に評価実験を行った．

本稿では，多様性について最も基本的に取り扱うため，いくつかの重要な要素を省いている．

- 参照元文書の数
- 各文書の情報量
- 各文書の影響力

今後の研究ではこれらの要素も組み込んだ上でのモデルの提案ならびに評価を行う．

また本稿における実験では，ページの内容に着目することで，ウェブページを専門性に基づいてランキングする手法について提案を行った．しかしながら，ページの内容の他にも，本文中のセンチメントや意見などの要素，文書に付随するメタデータから書かれた年代や著者の年齢など，様々な観点からの多様性が考えられ，これらを考慮することで専門性に限らない手法の活用が期待でき，今後の課題として研究を続ける予定である．

謝　　辞

本研究は文科省科研費基盤 (A)「ウエブ検索の意図検出と多元的検索意図指標にもとづく検索方式の研究」(24240013，研究代表者：田中克己)，および JSPS 科研費 245417 の一部補助によって行われました．ここに記して謝意を表します．

文　　献

[1] K. Akamatsu, N. Pattanasri, A. Jatowt, and K. Tanaka. Measuring comprehensibility of web pages based on link analysis. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, pages 40–46, Washington, DC, USA, 2011. IEEE Computer Society.

[2] G. Capannini, F. M. Nardini, R. Perego, and F. Silvestri.

data	map	p@5	p@10	p@20	p@30	nDCG
提案手法に基づく多様性の高さ	0.472	0.320	0.312	0.248	0.202	0.861
提案手法に基づく多様性の低さ	0.353	0.208	0.228	0.216	0.201	0.820
トピックを考慮した多様性の低さ	0.356	0.211	0.230	0.216	0.200	0.821
万人向け文書から多様に参照	0.392	0.247	0.252	0.222	0.201	0.830
万人向け文書から画一的に参照	0.378	0.229	0.247	0.217	0.201	0.828
専門的文書から多様に参照	0.354	0.209	0.229	0.214	0.201	0.820
専門的文書から画一的に参照	0.370	0.222	0.240	0.219	0.201	0.825
RANK	0.341	0.195	0.217	0.212	0.200	0.810

表 2　ページトピックの専門性に基づく評価

data	map	p@5	p@10	p@20	p@30	nDCG
提案手法に基づく多様性の高さ	0.15	0.05	0.06	0.06	0.09	0.78
提案手法に基づく多様性の低さ	0.16	0.08	0.09	0.08	0.09	0.82
トピックを考慮した多様性の低さ	0.16	0.08	0.09	0.08	0.09	0.82
万人向け文書から多様に参照	0.15	0.06	0.07	0.07	0.09	0.80
万人向け文書から画一的に参照	0.15	0.07	0.07	0.07	0.09	0.80
専門的文書から多様に参照	0.16	0.07	0.08	0.08	0.09	0.81
専門的文書から画一的に参照	0.15	0.07	0.08	0.07	0.09	0.81
RANK	0.17	0.09	0.09	0.08	0.09	0.82

表 3　ページの記述内容に基づく評価

記事の概要	TS	CS	Div	DD	DU	UU	UD	TOPIC	RANK	#P
個人ニートの blog	4	2	0.55	0.13	0.10	0.35	0.42	0.05	4	8
ニートに関するニュース記事	4	2	0.41	0.14	0.21	0.38	0.26	0.59	12	40
個人ニートの blog	4	3	0.43	0.12	0.15	0.41	0.32	0.65	22	29
個人ニートの blog	4	1	0.40	0.17	0.25	0.35	0.23	0.58	23	4
ニートを題材にした映画の公式サイト	4	2	0.57	0.28	0.22	0.22	0.28	0.15	29	5
ニートをサポートする企業	3	3	0.18	0.02	0.08	0.74	0.16	0.78	5	4
個人ニートの blog	3	3	0.36	0.00	0.00	0.64	0.36	0.42	6	3
個人ニートの blog	3	3	0.50	0.20	0.20	0.30	0.30	0.41	10	14
ニートに関する CQA コンテンツ	3	4	0.38	0.16	0.27	0.35	0.22	0.24	11	4
ニートに関するニュース記事	3	2	0.50	0.00	0.00	0.50	0.50	0.22	13	3
Wikipedia「社内ニート」	3	3	0.20	0.07	0.30	0.51	0.13	0.81	14	38
ニートを題材にした漫画	3	2	0.50	0.17	0.17	0.33	0.33	0.17	15	3
ニートを題材にした漫画	3	2	0.32	0.00	0.00	0.68	0.32	0.09	16	4
ニートを題材にした書籍の公式サイト	3	3	0.51	0.12	0.12	0.37	0.39	0.29	17	14
ニートに関するまとめ記事	3	2	0.37	0.16	0.27	0.36	0.21	0.56	19	8
個人ニートの blog	3	2	0.63	0.24	0.14	0.23	0.39	0.17	21	35
ニートに関するニュース記事	3	2	0.57	0.21	0.16	0.27	0.36	0.17	27	5
ニートをサポートする企業	3	2	0.10	0.01	0.09	0.81	0.09	0.82	28	3
個人ニートの blog	3	2	0.54	0.18	0.16	0.30	0.36	0.35	30	4
ニートをサポートする企業	2	2	0.48	0.14	0.15	0.37	0.35	0.28	7	58
個人ニートの blog	2	3	0.46	0.16	0.19	0.35	0.29	0.54	9	40
語源由来辞典「ニート」	2	2	0.12	0.03	0.19	0.69	0.09	0.81	18	5
「ニート」という名前の車の公式サイト	2	2	0.54	0.20	0.17	0.29	0.34	0.22	20	6
日本語俗語辞典「大人ニート」	2	1	0.12	0.04	0.30	0.58	0.08	0.77	24	4
日本語俗語辞典「恋愛ニート」	2	1	0.18	0.06	0.26	0.55	0.12	0.69	25	6
Wikipedia「ニート」	1	3	0.60	0.23	0.15	0.25	0.37	0.11	1	41
はてなキーワード「ニート」	1	2	0.61	0.23	0.15	0.24	0.37	0.18	2	9
ニコニコ大百科「ニート」	1	3	0.22	0.07	0.26	0.52	0.14	0.48	3	22
アンサイクロペディア「ニート」	1	4	0.48	0.21	0.23	0.29	0.27	0.06	8	6
日本語俗語辞書「ニート」	1	1	0.12	0.04	0.31	0.57	0.08	0.76	26	9

表 4　クエリ「ニート」の正解および結果

Efficient diversification of web search results. *Proc. VLDB Endow.*, 4(7):451–459, Apr. 2011.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.

[4] H. Deng, M. R. Lyu, and I. King. A generalized co-hits algorithm and its application to bipartite graphs. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 239–248, New York, NY, USA, 2009. ACM.

[5] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, VLDB '04, pages 576–587. VLDB Endowment, 2004.

[6] T. Haveliwala. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784 – 796, july-aug. 2003.

[7] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, Sept. 1999.

[8] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Computer Networks*, 33(1 – 6):387–401, 2000.

[9] E. Minack, G. Demartini, and W. Nejdl. Current approaches to search result diversification. In *Proceedings of The First International Workshop on Living Web at the 8th International Semantic Web Conference (ISWC)*, Oct. 2009.

[10] M. Nakatani, A. Jatowt, H. Ohshima, and K. Tanaka. Quality evaluation of search results by typicality and speciality of terms extracted from wikipedia. In X. Zhou, H. Yokota, K. Deng, and Q. Liu, editors, *Database Systems for Advanced Applications*, volume 5463 of *Lecture Notes in Computer Science*, pages 570–584. Springer Berlin Heidelberg, 2009.

[11] A. Stirling. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15):707–719, 2007.

[12] J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[13] Y. Takahashi, H. Ohshima, M. Yamamoto, H. Iwasaki, S. Oyama, and K. Tanaka. Evaluating significance of historical entities based on tempo-spatial impacts analysis using wikipedia link structure. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 83–92, New York, NY, USA, 2011. ACM.

[14] H. Tong. Fast random walk with restart and its applications. In *In ICDM ' 06: Proceedings of the 6th IEEE International Conference on Data Mining*, pages 613–622. IEEE Computer Society, 2006.

[15] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 115–122, New York, NY, USA, 2009. ACM.

[16] 合原博, 宮田高志, and 松本裕治. 医学生物学分野からの専門用語の抽出・分類. **情報処理学会研究報告. 自然言語処理研究会報告**, 2000(11):41–48, jan 2000.

[17] 中川裕志, 湯本紘彰, and 森辰則. 出現頻度と連接頻度に基づく専門用語抽出. **自然言語処理**, 10(1)(2003-01):27–46, 2003.