Vol.2009-DBS-149 No.3
2009/11/20

隣接ページのクエリ尤度を考慮した文書特徴付け手法の実装とその評価

田村航弥[†1]　波多野賢治[†2]　宿久洋[†2]

検索エンジンを通してユーザの情報要求を満たす情報を返すためには，各々の文書の内容を正確に考慮した特徴付けを行い順位付けする必要がある．文書検索技術にはTF-IDF 法のような経験則的に得られた手法が用られてきたが，近年の研究では確率的言語モデルを用いた情報検索が主流となり，また検索精度も向上していることが確認されている．この検索モデルでは，文書に対してクエリが生成される確率をクエリ尤度として算出し，このクエリ尤度が各文書スコアとされる．本稿では，この検索モデルでは考慮されていない文書間に存在する文書内容の関連性を，隣接文書のクエリ尤度を用いて考慮することによって新たな文書の特徴付手法を提案し，Web 文書検索へ応用することでその有効性を示す．

Implementation and Evaluation of A Document Characterization Method Considering Query Likelihood of Neighbor Page

Koya Tamura,[†1] Kenji Hatano[†2]
and Hiroshi Yadohisa[†2]

For retrieving information satisfying user's information need from a search engine, we need to characterize document contents and rank them precisely. Previously, we have used the TF-IDF method in information retrieval techniques. In recently studies, however probabilistic language model for information retrieval is mainly used and helps to increase retrieval accuracies of the search engine. Probabilistic language model calculates probability of generated query for each document as query likelihood, and that regard as its document score. In this paper, we propose a new approach for characterizing documents based on relativity of document contents between query likelihood of neighbor documents, and evaluate effectiveness of our approach to apply Web document retrieval.

1. はじめに

今日の高度情報化社会において，世の中に存在しているデータは増加の一途を辿っている．その中でも，インターネットの普及に伴った Web 文書や電子テキストなどのデータは氾濫しており，その結果，ユーザは検索エンジンを通して必要な情報を取得することが困難となってきている．このような問題を受け，情報検索に関する研究の大きなタスクの一つとして，ユーザの情報要求を満たす情報を如何に正確に提示するか，すなわち検索結果の精度向上が求められる．ユーザは一般的に，検索エンジンによって返される検索結果の上位数十件しか閲覧しないということが報告されている[10]．よって検索精度の向上を図るためには，ユーザから与えられたクエリを元に各文書に対して正確な評価を行い，正解文書上位にランキングする必要がある．このように文書集合に対して検索を行う際に，従来の検索モデルは TF-IDF 法[14] による索引語の重み付け手法など経験則的に得られたものを用いていた．しかし，近年の情報検索の分野における研究において，上記のような手法に比べて，確率的言語モデルによる情報検索手法であるクエリ尤度モデルの有用性が高いことが報告されており[12]，近年の研究で多く用いられるようになってきている．

クエリ尤度モデルとは各文書に対して与えられている文書モデルをもとにクエリが文書に発生するもっともらしさ，すなわちクエリ尤度を算出する検索モデルである[12]．この検索モデルは，各文書における言語現象を推測することが一番の課題であり，その推測方法には様々な手法が研究されている[9),17]．また，零頻度問題を解消するための手法も複数研究されており[6),7]，各手法において TF-IDF 法による重み付け検索手法より高い検索精度を実現している．

このように新たなアプローチによって検索モデルが提案されているが，今日の Web 検索では前述した検索技術より他の情報を利用して検索を行っている事例も数多く存在する．特に Web 検索に用いられている有名な手法として，リンク構造を解析する PageRank アルゴリズム[5] や HITS アルゴリズム[3] が挙げられる．この手法は，ある Web 文書から他の Web 文書へリンクしている out-link と，他の Web 文書からリンクされている in-link の二種類の情報を利用し，Web 文書を特徴付けたランキングを行っている．このリンク情報

†1 同志社大学大学院文化情報学研究科
Graduate School of Culture and Information Science, Doshisha University
†2 同志社大学文化情報学部
Faculty of Culture and Information Science, Doshisha University

を用いた Web 文書の特徴付け手法においても高い検索精度を実現しているが，この手法ではリンク構造のみを用いて Web 文書に対して特徴付けを行っているため，その Web 文書に記述されている内容は考慮されていない．よってこの手法は Web 文書に対して正確な特徴付けを行えているとは言い難い．

以上のような問題点を考慮した上で本稿では，従来の TF-IDF 法より良い成果を挙げているクエリ尤度モデル用いて，このモデルでは考慮されていなかった文書間の内容の関連性を隣接文書のクエリ尤度によって考慮した新たな文書特徴付け手法を提案する．本提案手法は検索された Web 文書に対して，その文書に隣接している文書のクエリ尤度を検索対象文書のクエリ尤度に付加する手法である．また本提案手法を，隣接文書のクエリ尤度を考慮しない従来のクエリ尤度モデルによる検索手法と比較し，本提案手法の評価を行う．

以下 2 章ではクエリ尤度モデルに関する基本事項及び関連研究について述べる． 3 章では提案手法について詳述し，4 章では提案手法に対して評価実験を行い，結果と考察を述べる．最後に 5 章では，まとめと今後の課題について述べる．

2. 基本的事項及び関連研究

2.1 クエリ尤度モデルに関する基本的事項

確率的言語モデルを用いた情報検索モデルは，各検索対象文書に対して起こる言語現象を言語モデルを用いて確率，統計的に推測し，その中で与えられたクエリが生成される確率の算出を行った後に，その値を元に文書ごとにランキングする手法である．このことからクエリ尤度モデルとも呼ばれる．クエリ尤度モデルの一番の課題は各文書に対して言語現象を推測することである．この文書に対する言語現象を文書モデルと呼び，文書モデルを推測するにためには，各文書に対して分布を仮定する必要がある．過去の研究においてはこの分布に対して，二項分布[12]，多項分布[17]，ポアソン分布[9] を仮定した上で，各文書に対して文書モデルの推定を行っている．この文書モデルの推定方法には，グッド・チューリング推定法 (Good - Turing estimate)[8] や EM アルゴリズム (Expectation Maximization algorithm) ，最尤推定法 (Maximum Likelihood estimate) などが存在する． これらの方法で推測された文書モデルを用いて各文書に対してクエリ尤度を求める．

しかし，クエリキーワード Q が複数の索引語で表現されていた場合，クエリキーワードの中の一つの索引語が出現しない，すなわちその索引語の尤度が 0 であれば，たとえ他の索引語の尤度が 0 でなくてもその文書に対するクエリ尤度は 0 となる．これを零頻度問題と呼ぶ．この問題を対処するために文書モデルのスムージングが用いられる．このスムージングにおいても，線形補完法[7] やディリクレ・スムージング[6] など様々な手法が提案されている．本提案手法では各文書に対して多項分布を仮定し，クエリ尤度モデルによって各文書に対して特徴付けを行っている．この際に文書モデルの推定には最尤推定法を，スムージングには線形補完法を用いている．これらの計算過程は次章で述べる．

2.2 関 連 研 究

情報検索モデルにおける文書を特徴付ける方法は大きく二つに分けられると言われている[18]．

一つはある文書に索引語が含まれているか否かのテキストベースで特徴付ける手法である．この特徴付け手法を用いた検索モデルで代表的な手法は，文書に対してクエリキーワードの有無のみで検索を行うブーリアン検索モデル (Boolean retrieval model)[15] から始まり，各文書の各索引語を要素とするベクトルで表現するベクトル空間モデル (Vector space model)[16] ，各文書に対して索引語が出現する確率を言語モデルを用いて推測するクエリ尤度モデル[9),12),17] が挙げられる．しかし，テキストベースでの検索エンジンでは，文書を構成している要素，特に Web 文書などに見られる特徴であるハイパリンクや文書の構造情報，位置情報，さらにユーザの情報などの様々な要素を考慮していないため，ユーザの情報要求を十分に満たすことができなかった．

この問題を解消するためにもう一方の特徴付け手法として，リンク構造を用いた文書の特徴付け手法が提案され，主に Web 検索エンジンで用いられるようになった．代表的な例として挙げられるのは Google [⋆1] の検索エンジンに適用されている PageRank アルゴリズム[5] や，CLEVER プロジェクト[1] で適用されている HITS アルゴリズム (Hypertext Indeuced Topic Search)[3] がある．これらの手法は Web 文書の重要度を判定するスコアリングアルゴリズムであり，Web 文書に存在する in-link，out-link の情報のみを用いている．またこれらの発展系として，各 Web 文書を複数のトピックから構成されるものとし，Web 文書のあるトピックから他の Web 文書のあるトピックへの遷移と考えたユーザモデルから文書の重要度を算出する研究[11] や，ユーザの Web 文書間の遷移のログからユーザモデルを構築し重要度を算出する手法[4] などがある．これらのリンク構造解析による Web 文書の特徴付け手法によってテキストベースでの文書の特徴付け手法よりも良い検索精度を実現している．

しかし，前述したリンク構造を用いた Web 文書に対する特徴付け手法に対して杉山ら

⋆1 http://www.google.com

Vol.2009-DBS-149 No.3
2009/11/20

は，(1) Web 文書に対する重みが単に定義されているに過ぎない，(2) リンクによって結ばれた Web 文書間の内容の関連性が考慮されているわけではない，という問題点を挙げており，このような問題を解決するためには，Web 文書の検索を行う単位を Web 文書単体で行うのではなく，その周辺の Web 文書の内容も考慮する必要がある，としている[18]．このように周辺文書の内容を考慮した特徴付けを行った上で Web 文書の検索を行っている研究[18] や Web 文書の分類を行っている研究[13] では，ターゲットとなる文書単体で特徴付ける手法と比べて良い成果を挙げている．

このような先行研究によって対象となる Web 文書単体で特徴付けを行うより，対象文書の周辺の文書の内容を考慮して Web 文書を特徴付けを行うことより，正確にその文書を特徴付けすることが可能であることが分かる．ここで，これらの手法はいずれも Web 文書の特徴付けの際に TF-IDF 法[14] を用いて，隣接文書の特徴量を検索対象文書に対して付加しているが，本手法では検索が行われた後に算出されるクエリ尤度を隣接文書に対して考慮している点で従来手法と異なる．また TF-IDF 法より良い検索精度を実現しているクエリ尤度モデルを用いて隣接文書のクエリ尤度を考慮することで，検索精度の向上が期待できる．

3. 提 案 手 法

2.2 節で述べたように，Web 文書を特徴付ける際に周辺ページの特徴を考慮することは，その文書を正確に特徴付けることにつながると考えられる．そこで本章では各文書の隣接文書のクエリ尤度を考慮した文書の特徴付け手法を提案する．本提案手法は，ユーザが与えたクエリを用いて各文書のクエリ尤度を算出する際に，その文書の隣接文書のクエリ尤度を検索された文書に反映させる方法である．以下に本提案手法の処理手順を説明する．

(1)　検索クエリを入力
(2)　クエリ尤度モデルにもとづいて各 Web 文書に対してクエリ尤度を算出
(3)　各文書に対して，その文書と隣接している文書のクエリ尤度を付加
(4)　各文書のクエリ尤度の再計算及びその尤度にもとづいたランキング

図 1 は，上記の各手順を図式化したものである． 以下にその各手順について詳述する．

3.1 各文書のクエリ尤度の算出

本節では図 1 における手順 (1)，(2) の説明をする．まず，ユーザによって検索クエリ $Q = (t_1, t_2, ..., t_n)$ が与えられる．ここで $t_i (i = 1, 2, ..., n)$ は索引語であり， n はクエリを構成している索引語の数であり，このクエリ Q を用いて各文書のクエリ尤度を算出する．本提案手法では 2.1 節で述べたように，各文書に対して多項分布を仮定し，この分布をもとに文書モデルを推定する．ここでの推定方法は最もシンプルである最尤推定[12] を用いる．多項分布におけるパラメータに関する最尤推定の一般形を用いた文書モデルの推定は以下の式を用いて行う．

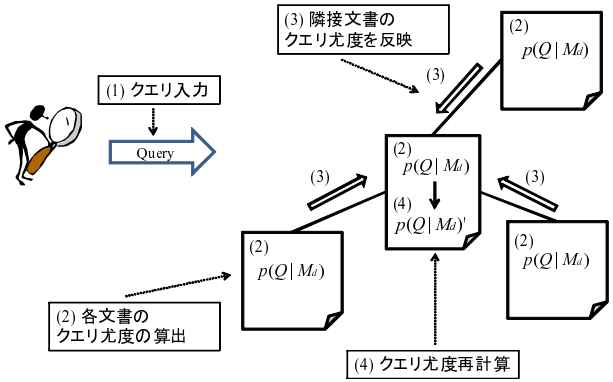

図 1　提案手法概要

$$\hat{P}_{mle}(t_i|M_d) = \frac{tf_{t_i,d}}{N_d} \tag{1}$$

ここで $tf_{t_i,d}$ は文書に出現する索引語 t_i の出現頻度であり，N_d は文書長である．また M_d は推測された文書モデルである．この式 (1) を用いて各単語の出現確率を算出する．また，ここで起こる零頻度問題を解消するために本提案手法では線形補完法によるスムージング[7),17)] を以下の式を用いて行う．

$$\hat{P}(t_i|M_d) = \omega \hat{P}_{mle}(t_i|M_d) + (1-\omega)\hat{P}_{mle}(t_i|M_c) \tag{2}$$

ここで M_c はコーパスモデルであり，検索対象文書全体での索引語の尤度を算出し，文書モデルで算出された尤度に加算することで，例え $\hat{P}_{mle}(t_i|M_d)$ の値が 0 であっても，式 (4) を用いた際のクエリ尤度が 0 になる零頻度問題を解消している．このコーパスモデル M_c は以下の式によって推定する．

$$\hat{P}_{mle}(t_i|M_c) = \frac{\sum_{d \in c} tf_{t_i,d}}{\sum_{d \in c} N_d} \tag{3}$$

また ω は $0 < \omega < 1$ の値をとる M_c によるスムージングの度合いを決定するパラメータである．ω に設定する値が大きければ，文書モデルの効果を強調し，小さくすればスムージ

ングの効果を大きくする．上記の方法で得られた索引語の尤度 $\hat{P}(t_i|M_d)$ を用い，下記の式 (4) から各文書のクエリ尤度を算出する．

$$\hat{P}(Q|M_d) = \prod_{t_i \in Q} p(t_i|M_d) \tag{4}$$

3.2 各文書に対する隣接文書のクエリ尤度の付加

本節では図 1 における手順 (3)，(4) の処理である，各文書に隣接する文書のクエリ尤度を考慮して Web 文書を特徴付ける方法を説明する．Web 文書同士がリンクによって接続されているということは，その Web 文書間に何らかの関連があると考えられる．そして，検索対象の文書の周辺にクエリ Q による尤度が高い文書が多く存在する場合，検索対象の文書に対してもクエリ Q の情報要求を満たすような文書であると考えられる．

図 2 にその具体例を示す．同じクエリ Q で検索された文書 a，b があり，それぞれ 図 2 のように隣接文書が存在するとする．また文書 a はクエリ Q によるクエリ尤度がそれぞれ高い文書と隣接しており，対して文書 b はクエリ Q によるクエリ尤度がそれぞれ低い文書と隣接している．このような状況で文書 a, b のクエリ尤度を考えた場合，文書 a はクエリ Q に関してクエリ尤度の高い文書と隣接していることから，文書 a はクエリ Q の内容に合致する文書であると考えられる．これに対して文書 b はクエリ Q に関してクエリ尤度の低い文書と隣接していることから，文書 b はクエリ Q の内容に合致しない文書であると考えられる．よってこれらの文書 a, b に対して算出されるクエリ尤度を何らかの形で調節することによって，隣接文書の内容を考慮した文書の特徴付けが行える．このような考えの元，隣接文書のクエリ尤度を考慮した特徴付けを行う．

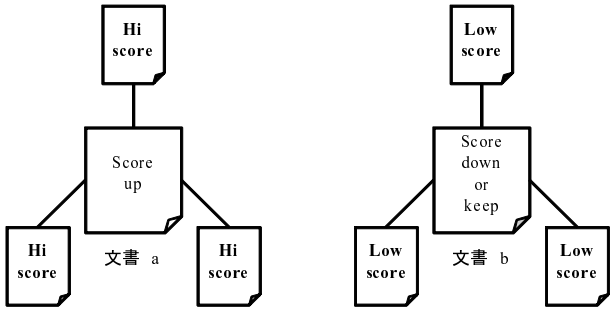

図 2 検索スコアによる隣接文書との関連度

まず，ある文書 d とリンクによって接続された文書群 $U_d = (u_1, u_2, ...u_{N_U})$ を抽出する．ここで文書群 U_d は，各文書間のリンクの方向は考慮せず，文書 d と単方向または双方向リンクで接続されている文書全てであり，N_U は U_d の総数とする．この U_d に含まれる文書のクエリ尤度を文書 d に反映することによって，周辺文書の特徴を考慮した文書 d の特徴付けとする．

この特徴付けの手法について本稿では以下の二つの事項を考慮する．

(1) 対象文書に隣接している文書の内容がクエリと合致している，すなわちクエリ尤度が高ければ対象文書のクエリ尤度に対して周囲の文書の特徴を反映させる必要がある．

(2) 上記のような状況下である Web 文書とそうでない文書，すなわち隣接文書のクエリ尤度が低い文書とでは，対象文書のクエリ尤度に差を付ける必要がある．

以上の考えをふまえ，対象文書に隣接している文書のクエリ尤度を対象文書に反映させることを手法を提案する．以下にその方法を詳述する．

3.2.1 隣接文書のクエリ尤度の付加手法 1

あるクエリに対して，検索された対象文書に隣接している文書のクエリ尤度が高ければ対象文書のクエリ尤度にそれを反映させる必要がある．検索クエリ Q によって検索された対象文書の周辺にもクエリ尤度が高い文書が多いということは，対象文書の周辺にも検索質問の内容を含んだ Web 文書が多いということであり，対象文書の内容は検索クエリの内容に合致する Web 文書であると考えられる．具体例として図 3 のようなクエリ Q によって検索された対象文書 1 とその文書とリンクによって隣接しているページがあったとする．ここで対象文書 1 の隣接文書のクエリ尤度を，対象文書に与えられているクエリ尤度 $P(Q|M_1)$ に付加する (図 4)．この隣接文書のクエリ尤度を検索対象文書に反映する方法を手法 sum1 とし，以下の式 (5) で表す．

$$P'_{sum1}(Q|M_d) = P(Q|M_d)(\sum_{u_j \in U_d}^{n} P(Q|M_{u_j})) \tag{5}$$

ここで $P'_{sum1}(Q|M_d)$ は 文書 d のクエリ尤度を隣接文書のクエリ尤度を用いて再計算した値である．式 (5) 右辺第二項では，検索クエリ Q によって検索された対象文書の隣接文書集合に対してクエリ尤度の総和を算出し，この値を対象文書のクエリ尤度 $P(Q|M_{d_a})$ に乗じて $P'_{sum1}(Q|M_1)$ を算出する．図 3 においては，対象文書 1 のクエリ尤度が 0.1, 隣接文書のクエリ尤度がそれぞれ 0.2, 0.3, 0.4 であったとすると，実際に対象文書 1 の再計算されたクエリ尤度 $P'_{sum1}(Q|M_1)$ は $0.1 * (0.2 + 0.3 + 0.4) = 0.07$ となる．

Vol.2009-DBS-149 No.3
2009/11/20

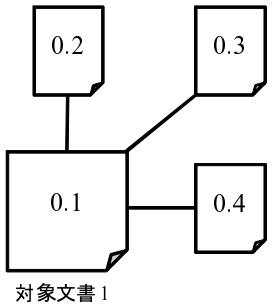

図 3 検索対象文書と隣接文書及びクエリ尤度

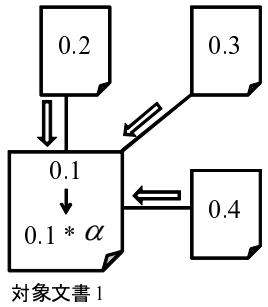

図 4 隣接文書のクエリ尤度による再計算

一方，隣接文書のクエリ尤度が高い文書と，隣接文書のクエリ尤度が低い文書とでは，対象文書のクエリ尤度に差を付ける必要がある．これは，検索対象文書の隣接文書が複数存在し，その中にクエリ尤度が高い文書が複数存在した場合，隣接文書の数が多ければ対象文書に対して与える効果も薄れると考えられるからである．例として図 5 のように，検索クエリ Q において検索された対象文書 1，2 と各々の隣接文書があったとする．ここで，対象文書 1 は隣接文書は 4 個存在し，その中でクエリ尤度が高いものは 2 個である．それに対して対象文書 2 は隣接文書は 2 個存在し，2 個ともクエリ尤度が高い状態である．このような場合，対象ページ 2 の方が 1 よりもクエリ尤度が高い隣接文書を含んでいる割合が大きくなるので，対象文書 2 により大きく周辺文書のクエリ尤度を付与する必要がある．このように隣接文書のクエリ尤度を検索対象文書に反映する方法を手法 ave1 とし，以下の式 (6) で表す．

$$P'_{ave1}(Q|M_d) = P(Q|M_d)(\sum_{u_j \in U_d}^{n} \frac{P(Q|M_{u_j})}{N_U}) \tag{6}$$

式 (6) 右辺第二項では，隣接文書のクエリ尤度の平均値を算出している．この値を対象文書のクエリ尤度 $P(Q|M_d)$ に乗じて $P'_{ave1}(Q|M_d)$ を算出する．図 6 においては対象文書 1 の場合，文書 1 のクエリ尤度が 0.1, 隣接文書のクエリ尤度がそれぞれ 0.2, 0.3, 0.4, 0.5 であったとすると，実際に対象文書 1 の 再計算されたクエリ尤度 $P'_{ave1}(Q|M_1)$ は $0.1 * (\frac{0.2+0.3+0.4+0.5}{4}) = 0.35$ となる．対象文書 2 の場合，文書 2 のクエリ尤度が 0.1, 隣接文書のクエリ尤度が共に 0.5 であったとすると，実際に対象文書 2 の 再計算されたクエリ尤度 $P'_{ave1}(Q|M_2)$ は $0.1 * (\frac{0.5+0.5}{2}) = 0.05$ となる．

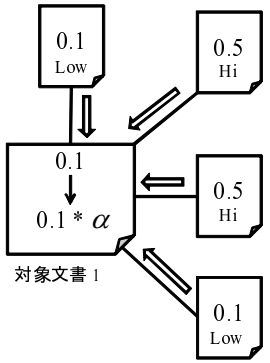

図 5 クエリ尤度の低い隣接文書を含む場合

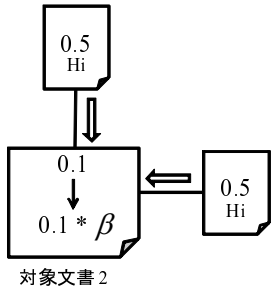

図 6 クエリ尤度の高い隣接文書のみを含む場合

3.2.2 隣接文書のクエリ尤度の付加手法 2

前述した提案手法 1 では，隣接文書のクエリ尤度の総和及び平均値を検索対象文書のクエリ尤度に乗じていたため，隣接文書の出現の仕方によってあらかじめ算出されている検索対象文書のクエリ尤度から大きく変動する可能性がある．このような隣接文書のクエリ尤度に依存した提案手法 1 に対して本提案手法は，あらかじめ算出された検索対象文書のクエリ尤度を基準値として再計算を行う手法を提案する．提案手法 1 で説明した隣接文書のクエリ尤度の総和による反映方法 (式 (5))，平均値による反映方法 (式 (6)) の考え方を元に，検索対象文書のクエリ尤度に依存した隣接文書のクエリ尤度の反映方法をそれぞれ手法 sum2, ave2 とし，以下の式 (7)，(8) で表す．

$$P'_{sum2}(Q|M_d) = P(Q|M_d)((\sum_{u_j \in U_d}^{n} P(Q|M_{u_j})) + 1) \tag{7}$$

$$P'_{ave2}(Q|M_d) = P(Q|M_d)((\sum_{u_j \in U_d}^{n} \frac{P(Q|M_{u_j})}{N_U}) + 1) \tag{8}$$

ここでは，式 (7)，(8) の右辺第二項でそれぞれ隣接文書のクエリ尤度の総和，平均値を算

出した値に対して 1 を加えている．この値を検索対象文書のクエリ尤度に乗じることによって，そのクエリ尤度の値に依存した再計算が可能となる．

4. 評価実験

本実験は，隣接文書のクエリ尤度を考慮した文書特徴付け手法の妥当性を評価することを目的とする．この妥当性と評価を行うために，後述するテストコレクションを用いて評価実験を行った．なお，比較対象として用いる従来手法は，隣接文書の内容を考慮しないクエリ尤度モデルである．また，従来手法及び提案手法におけるスムージングの際のパラメータには過去の研究[17] において良い検索精度を実現している 40% を採用しており，この値が一般的なパラメータの値であると言及されている[7]．

4.1 評価尺度

本実験に用いる評価尺度には 101 点 平均精度 (101-point average precision) 及び，再現率 - 精度グラフ (recall-precision graph) を用いている．再現率 (recall) とは情報検索システムの完全性を評価するための尺度，精度 (precision) とは情報検索システムの正確性を評価するための尺度であり，以下のように定義される．

$$\text{再現率 } R_k = \frac{\text{上位 } k \text{ 番目までに検索された正解文書数}}{\text{正解文書数}} \tag{9}$$

$$\text{精度 } P_k = \frac{\text{上位 } k \text{ 番目までに検索された正解文書数}}{\text{上位 } k \text{ 番目までに検索された文書数}} \tag{10}$$

ここで，本評価尺度を用いる上で必要である 101 点の再現率レベルと検索結果から得られる再現率は実際に異なるため，補完精度 (interpolated precision) を用いて 101 点の再現率レベルでの精度を求める．この再現率レベル x 補完精度 $P(x)$ は以下の式を用いて算出する．

$$P(x) = max_{x \leq R_k} P_k \tag{11}$$

式 (11) の右辺は再現率レベル x 以上の検索結果から得られる精度のうち最大のものを求めている．このように算出された再現率を x 軸に，補完精度を y 軸に描き，再現率が 0% から 100% までの 101 点の各点においての精度を推移を描画するグラフが再現率 - 精度グラフである．このグラフでは，精度の推移が描画された曲線がグラフ上部にあれば，描かれた曲線の検索システムは精度が高いと言える．またその 101 点の精度それぞれの平均を算出したものが 101 点平均精度であり以下の式を用いて算出する．

$$\tilde{P} = \frac{1}{101} \sum_{l=0}^{100} P(\frac{l}{100}) \tag{12}$$

4.2 テストコレクション

本実験では 2008 年の INEX (INitiative for the Evaluation of XML Retrieval) テストコレクションを用いて各実験及び評価を行った．INEX テストコレクションは，XML 部分文書検索のための国際プロジェクトである INEX Project ⋆1 によって 2002 年より構築作業が行われている XML 部分文書検索システム用の性能評価データセットである．これは，検索対象となる約 66 万個の XML 文書で構成される 2008 Wikipedia document collection[2]，285 個のクエリの集合である INEX topics，それらのクエリに対する正解部分文書集合及びその評価が記述されている INEX relevance assessment の三つで構成されている．本実験ではこの 2008 Wikipedia document collection に対して，不要語リスト⋆2 に基づいて不要語を取り除き，Porter Stemmer ⋆3 を用いてステミング処理を行ったものを使用している．また，本実験のタスクは部分文書検索ではなく全文検索であるため，正解部分文書集合が存在している文書を解答文書と見なして評価を行った．

4.3 実験結果及び考察

本節では，INEX topics から正解部分文書が用意されていた 70 個のクエリを用いて従来手法と提案手法の検索精度比較を行った．実験の結果を表 1，図 7，8 に示す．

手法 sum1, ave1 ではそれぞれの手法で 101 点平均精度において従来手法を下回る結果となった．各再現率での精度の推移は，手法 sum1 が再現率 0.0 から 0.02 において従来手法より精度が上回る結果となった．

手法 sum2, ave2 ではそれぞれの手法ともに従来手法の検索精度を上回る結果となった．各再現率での精度の推移は，手法 sum2 においては各再現率の点での精度で従来手法を上回る結果となった．一方，ave2 においては従来手法とほぼ同じ値という結果となった．以上の各結果よりまず，手法 sum1, ave1 の実験結果について考察する．

手法 sum1, ave1 は，3.2.2 節でも述べたように，隣接文書の数，クエリ尤度に大きく依存する手法である．よって，ある検索対象文書のクエリ尤度が高く正解文書であったとしても，その文書の隣接文書が少なかったり，存在してもクエリ尤度が少ない場合，再計算され

⋆1 http://www.inex.otago.ac.nz/
⋆2 ftp://ftp.cs.cornell.edu/pub/smart/english.stop
⋆3 http://www.tartrus.org/%7Emartin/PorterStemmer/

Vol.2009-DBS-149 No.3
2009/11/20

る検索対象文書のクエリ尤度は大幅に減少する．その結果，各文書を順位付けした場合に正解文書が低く順位付けされ，検索精度の低下を招いているのではないかと考えられる．

次に sum2, ave2 の実験結果について考察する．これらの手法は，検索対象文書のクエリ尤度を基準に考える手法であるため，手法 sum1, ave1 と比較すると，クエリ尤度の再計算を行った際の検索対象文書のクエリ尤度の変化は小さくなる．よって，たとえ正解文書の隣接文書が少ない，また存在する隣接文書のクエリ尤度が低くとも，その文書のクエリ尤度は保たれる．そして，隣接文書のクエリ尤度が高い場合は，検索対象文書のクエリ尤度も高くなるよう再計算されるため，検索精度の向上を導いたと考えられる．

さらに四つの提案手法全ての検索精度を比較すると，手法 sum1, sum2 に比べて手法 ave1, ave2 のほうが検索精度の低下を招いていることが分かる．このような結果を導いた原因を分析するために，テストコレクションに対して索引語の尤度を算出するための統計量を表 2 のように算出した．

表 1 101 点平均精度

	101 点平均精度	上昇率 (%)
従来手法	0.0710	-
手法 sum1	0.0634	- 10.67%
手法 ave1	0.0319	- 55.02%
手法 sum2	0.0850	+ 19.75%
手法 ave2	0.0715	+ 0.6791%

この統計量から，各文書に対してスコアとして算出されるクエリの尤度は非常に小さな値であることが分かる．この値に対して手法 sum1, sum2 では隣接ページのクエリ尤度の総和を求め，その値を元に対象文書のクエリ尤度を再計算している．隣接文書のクエリ尤度の総和をとった値を用いることによって，対象文書のクエリ尤度に対して十分に変化を与えられる影響力を持った値となり，正解である可能性がある文書に対してスコアを上昇させることができたと考えられる．

一方，手法 ave1, ave2 では隣接文書のクエリ尤度の総和に対して，隣接文書の数で割った平均値を元に対象文書のクエリ尤度を再計算している．その結果，非常に小さい値で算出される隣接文書のクエリ尤度に対して，さらに値そのものを小さくする処理を加えていることとなる．特に表 2 の隣接ページに関する統計量から分かる通り，隣接ページが多い文書では 70,000 文書を超える文書が存在している．このような文書数で除する処理を行った結果，対象文書に対して影響力の弱い値となってしまい，手法 ave1 では，大きく正解文書

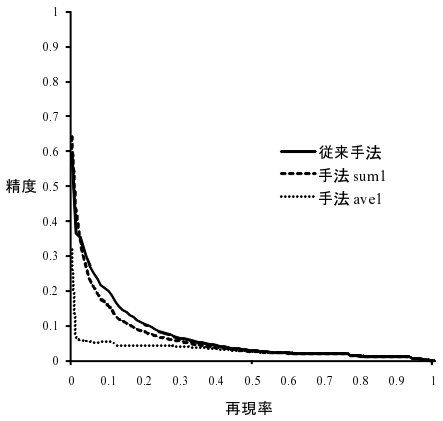

図 7 実験結果： 提案手法 1

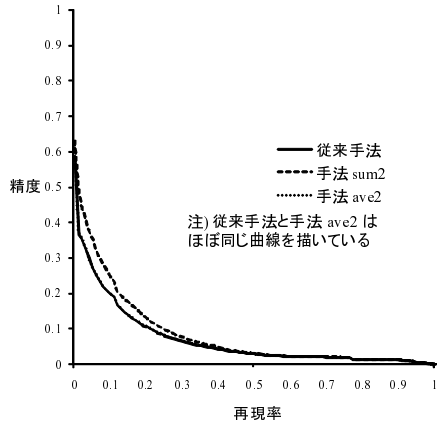

図 8 実験結果： 提案手法 2

表 2 テストコレクションに関する統計量

	統計量
全索引語数	148,634,720
全文書数	659,388
異なり語数	5,083,553
各文書における索引語の最大出現回数	27,221
各文書における索引語の最小出現回数	1
各文書における平均索引語数	225.41
クエリの数	70
クエリの最大単語数	6
クエリの最小単語数	1
クエリの平均単語数	3.1
各文書における最大隣接文書数	75,085
各文書における最小隣接文書数	1
平均隣接文書数	37.67

のクエリ尤度を下げる結果に，また手法 ave2 では正解文書に対しても大きな影響を与えることができなかったと考えられる．このことから手法 ave1, ave2 では，隣接文書にクエリ尤度が高い文書が存在したとしても，他のクエリ尤度の低い文書によって打ち消しあうことになり，その結果，隣接文書のクエリ尤度を十分に考慮することができなくなってしまったと考えられる．しかし，非常に小さな値ではあるが従来手法を上回っているため，この考

え方をもとに，隣接ページのクエリ尤度が低い文書の扱い方などを考慮に入れた改良手法の提案が今後の課題の一つとして挙げられる．

5. おわりに

本稿では，Web 文書に対して正確な特徴付けを行うために，隣接文書のクエリ尤度を考慮した特徴付け手法を実装し，その評価を行った．我々が提案した手法の中で，特に検索対象文書のクエリ尤度に重きをおいた検索手法において従来手法の検索精度を上回る結果となった．また隣接文書におけるクエリ尤度の扱い方やクエリ尤度の低い隣接文書の扱い方に改良の余地があるという知見が得られた． 今後の課題として，このような問題を解決できるような手法の提案が挙げられる．

謝辞 本研究の一部は，独立行政法人日本学術振興会 科学研究費補助金 基盤研究 (C) (課題番号：21500284) によるものである．ここに記して謝意を表す．

参 考 文 献

1) IBM AlmadenResearch Center. Clever searching. `http://www.almaden.ibm.com/cs/k53/clever.html`.
2) Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML corpus. *SIGIR Forum*, Vol.40, No.1, pp. 64–69, 2006.
3) Jon Michael Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pp. 668–677, Philadelphia, PA, USA, 1998. Society for Industrial and Applied Mathematics.
4) Amy N. Langville and Carl D. Meyer. Google 's Pagerank and Beyond: The Science of Search Engine Rankings. Princeton University Press, 2006.
5) Page Lawrence, Brin Sergey, Motwani Rajeev, and Winograd Terry. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, 1999.
6) David J. C. MacKay. A hierarchical dirichlet language model. *Natural Language Engineering*, Vol.1, No.3, pp. 1–19, 1995.
7) Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.
8) Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.
9) Qiaozhu Mei, Hui Fang, and Chengxiang Zhai. A study of poisson query generation model for information retrieval. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 319–326. ACM, 2007.
10) Satoshi Nakamura, Shinji Konishi, Adam Jatowt, Hiroaki Ohshima, Hiroyuki Kondo, Taro Tezuka, Satoshi Oyama, and Katsumi Tanaka. Trustworthiness Analysis of Web Search Results. *Research and Advanced Technology for Digital Libraries*, pp. 38–49, 2007.
11) Lan Nei, Brian D. Davison, and Xiaoguang Qi. Topical link analysis for web search. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 91–98. ACM, 2006.
12) Jay M. Ponte and William B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275–281. ACM, 1998.
13) Xiaoguang Qi and Brian D. Davison. Classifiers without borders: incorporating fielded text from neighboring web pages. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 643–650. ACM, 2008.
14) Gerard M. Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, Vol.24, No.5, pp. 513–523, 1988.
15) Gerard M. Salton and Michael J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, 1983.
16) Gerard M. Salton, A. Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Commun. ACM*, Vol.18, No.11, pp. 613–620, 1975.
17) Fei Song and William Bruce Croft. A general language model for information retrieval. In *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management*, pp. 316–321. ACM, 1999.
18) Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa, and Shunsuke Uemura. Improvement in tf-idf scheme for web pages based on the contents of theirhyperlinked neighboring pages. *The transactions of the Institute of Electronics, Information and Communication Engineers. D-I*, Vol.87, No.2, pp. 113–125, 20040201.