リンク情報に基づく周辺文書の索引語尤度を考慮した文書検索手法の提案と評価

田村　航弥† 　波多野賢治†† 　宿久　洋††

† 同志社大学大学院 文化情報学研究科　〒 610-0394 京都府京田辺市多々羅都谷 1-3
†† 同志社大学 文化情報学部　〒 610-0394 京都府京田辺市多々羅都谷 1-3
E-mail: †tamura@ilab.doshisha.ac.jp, ††{khatano,hyadohis}@mail.doshisha.ac.jp

あらまし　近年，確率的言語モデルを用いた情報検索に関する研究が盛んに行われている．この検索モデルは数理的な枠組みから説明されている点が特徴であり，過去に提案された検索手法に適用可能である点から非常に柔軟な検索モデルであるといえる．また，この検索モデルは，文書内の索引語尤度のみを用いて文書の順位付けを行っているが，Web 文書検索を想定した際には，各 Web 文書にはテキスト情報の他にリンク情報も有しており，この情報は Web 文書検索を行う際に非常に有用な情報であるといえる．そこで本稿では，Web 文書のリンク情報を利用し，隣接文書の索引語尤度まで考慮した文書検索手法を提案し，その有用性を確認する．

キーワード　Web 文書検索，確率的言語モデル，リンク構造解析

Link-Based Retrieval Considering Term Likelihood of Neighboring Pages

Koya TAMURA†, Kenji HATANO††, and Hiroshi YADOHISA††

† Graduate School of Culture and Information Science, Doshisha University
1-3 Tatara-Miyakodani, Kyotanabe, Kyoto 610-0394, Japan
†† Faculty of Culture and Information Science, Doshisha University
1-3 Tatara-Miyakodani, Kyotanabe, Kyoto 610-0394, Japan
E-mail: †tamura@ilab.doshisha.ac.jp, ††{khatano,hyadohis}@mail.doshisha.ac.jp

1. はじめに

今日のインターネットの普及に伴い，World Wide Web (WWW) に存在する Web 文書や電子テキストなどのデータは氾濫している．このような状況の中，大量のデータから必要な情報を検索する Web 検索エンジンは，我々の情報収集するための手段として欠かせないものとなってきている．しかし，このデータの増加に伴い，ユーザは自身にとって有益な情報を入手することが困難になっている．このような問題があるなかで，如何にしてユーザに対して有用な情報を提示するかという情報検索システムの検索精度に関する研究は，この研究分野において非常に重要な課題である．

この研究分野において，近年は確率的言語モデルを用いた情報検索モデル（以下，クエリ尤度モデル）に関する研究が盛んに行われている．確率的言語モデルとは，ある言語内での語の並びを確率的に表現した統計モデルであり，主に音声認識や機械翻訳の分野において利用されてきた．この確率的言語モデルが Ponte らによって 1998 年に情報検索の分野にも応用され，クエリ尤度モデルが提案された [7]．このクエリ尤度モデルは数理的な枠組みから説明されている点が特徴であり，また過去の研究で経験的に提案されてきた検索手法に対して適用可能である点から，非常に柔軟な検索モデルであるといえる [11]．

このような特徴のあるクエリ尤度モデルであるが，これは基本的に文書内のテキスト情報のみを用いて各文書の順位付けを行っている．しかし我々はこの点に関して，特に Web 文書検索を考慮した際に問題であると考える．Web 文書はテキスト情報のほかにもリンク情報を有しており，これは Web 検索を行う際には非常に有用な情報である．過去の研究において，リンク情報のみを用いて文書を順位付けする PageRank [5] や HITS [4]，ベクトル空間モデルに対してリンク情報を用いて検索を行う研究 [9] など，リンク情報を用いる事によって検索精度の向上が確認されている．

そこで本稿では，クエリ尤度モデルに対してリンク情報を用い，周辺文書の索引語単位でその尤度を考慮した文書検索手法を提案する．以下 2. ではクエリ尤度モデルの基本的事項と関連研究について述べる．3. では提案手法について詳述し，4.

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. では，まとめと今後の課題について述べる．

2. 基本的事項と関連研究

2.1 クエリ尤度モデルに関する基本的事項

クエリ尤度モデルとは，ユーザによって与えられたクエリが各検索対象文書に適合している確率を算出する検索モデルである [7] [8]．クエリ尤度モデルの基本的な概念として，各検索対象文書の文章は言語現象をモデル化した言語モデルに基づいて生成されたサンプル文書としてみなされる．よって，クエリと検索対象文書の適合確率を算出するためには，各文書内での言語モデル，すなわち文書モデルを推測する必要がある．ここで，情報検索には一般的にユニグラムモデルが用いられる．ユニグラムモデルは，各文書の索引語の生起に対して二項分布の一般化である多項分布を仮定，すなわち，各索引語が独立して生起すると仮定するモデルである．このような言語モデルを用いた場合，以下の式によってクエリと各検索対象文書の適合確率が算出される．

$$\hat{P}(Q|M_{d_i}) = \prod_{t_{ij} \in Q} \hat{P}(t_{ij}|M_{d_i}) \tag{1}$$

ここで，$d_i (i = 1, 2, \ldots, l)$ は各文書，Q はユーザが与えたクエリキーワード集合である．また $t_{ij} (j = 1, 2, \ldots, m)$ は各文書に含まれている索引語である．ここで算出された $\hat{P}(Q|M_{d_i})$ がクエリと文書の適合度であり，これをクエリ尤度と呼ぶ．このクエリ尤度の大きさによって，文書の順位付けを行う．

上記のクエリ尤度モデルを用いるにあたり，各文書モデル各クエリキーワードの出現確率 $P(t_{ij}|M_{d_i})$ を算出する必要がある．ここでは，ユニグラムモデルにおいて一般的に用いられる最尤推定を用いて算出する．

$$P_{mle}(t_{ij}|M_{d_i}) = \frac{tf_{t_{ij}}^{d_i}}{N_{d_i}} \tag{2}$$

ここで $tf_{t_{ij}}^{d_i}$ は索引語 t_{ij} が文書 d_i に出現している頻度を表し，N_{d_i} は文書 d_i の文書長である．

このようにクエリキーワードの出現確率を算出し最終的にクエリ尤度を求めるが，クエリキーワードが検索対象文書に出現しない場合は，出現確率 $P(t_{ij}|M_{d_i}) = 0$ と算出される．その結果，他のクエリキーワードの出現確率が存在していたとしても，式 (1) によってクエリ全体の尤度が 0 になる．これを一般的に零確率問題と呼ばれる．この問題を回避するためにスムージングという手法が用いられる．スムージング技術には複数の手法が提案されているが，ここでは過去の比較研究より [1] 往々にして良い検索精度を実現している線形補完法を用いる [3]．線形補完法とは文書におけるクエリキーワードの出現確率と文書集合全体での出現確率をパラメータによって線形結合することで実現され，以下の式によって表される．

$$P(t_{ij}|M_{d_i}) = \omega P_{mle}(t_{ij}|M_{d_i}) + (1 - \omega) P_{mle}(t_{ij}|M_C) \tag{3}$$

ここで，ω はウェイティングパラメタであり，$0 < \omega < 1$ の値をとる． また，M_c はコーパスモデルと呼ばれ，検索対象文書全体で出現している索引語の確率を算出している．このコーパスモデルを用いることによって，クエリキーワードの尤度が 0 すなわち $\hat{P}_{mle}(t_{ij}|M_{d_i}) = 0$ となっても，クエリ全体の尤度が 0 になることを回避している．また，ここで用いているコーパスモデルを以下のように定義する．

$$\hat{P}_{mle}(t_{ij}|M_C) = \frac{\sum_{d_i \in C} tf_{t_{ij}}^{d_i}}{\sum_{d_i \in C} N_{d_i}} \tag{4}$$

2.2 関 連 研 究

周辺文書の特徴を考慮した確率的言語モデルによる文書検索手法は過去の研究において複数提案されている．

我々は，隣接文書のクエリ尤度を考慮した文書特徴づけ手法を提案している [10]．この手法では，検索対象文書とリンクによって接続された隣接文書は，検索対象文書に記述されている内容と関連があると考え，2.1 で述べたクエリ尤度に加え，隣接文書のクエリ尤度を検索対象文書のクエリ尤度に加味して最計算するというアプローチをとっている．この手法によって，隣接文書を考慮しない検索手法よりも良い検索精度を実現した．しかしこの手法では，最終的に算出された文書のクエリ全体の尤度を用いて検索対象文書のクエリ尤度を再計算しているため，個々のクエリキーワードの尤度まで考慮できていないという問題点がある．

Liu らは，文書クラスタを用いた検索手法に言語モデルを適用している [6]．過去の研究において，検索対象文書単体で検索を行うよりも，その文書と内容が類似した文書の特徴も考慮した検索手法のほうが良い検索精度を実現している．この知見を元に Liu らは，まず検索対象文書全体に対してハードクラスタリングを行い，そのクラスタ内でのクエリキーワードの出現確率を算出している．そして，文書内でのクエリキーワードの尤度とクラスタ内での尤度，文書全体での尤度を線形結合する事によって新たなクエリ尤度算出法を提案している．この手法は，各文書に類似している文書を見つけ出すためにクラスタリングを行っているが，これを行うための処理コストが非常に大きいことが問題としてあげられる．また，索引語の情報のみを用いて行われている点においても我々の手法と異なる．

3. 提 案 手 法

我々は，リンクで接続している文書の内容は何らかの関連性があると考え，その隣接文書の特徴を考慮して各 Web 文書のクエリ尤度を算出する手法を提案する．2.2 で述べたように，過去に提案した我々の手法は，隣接文書のクエリ全体の尤度を加味するアプローチをとっていた．しかし，クエリは複数のクエリキーワードから構成されている場合があり，隣接文書のクエリ全体の尤度を用いて検索対象文書のクエリ尤度を再計算すると，個々のクエリキーワードの尤度が保持している特徴を正確に捉えることができない．

具体的に図 1 を用いて説明する．ここではクエリ Q は obama，family，tree の三つのクエリキーワードから形成されているとする．よってこの場合ユーザの情報要求として，オバ

マ米大統領の家系についての情報であると考えられる．図 1 の状況においては，検索対象文書 d_1 のクエリ尤度 $P(Q|M_{d_1})$ に対して，隣接文書である文書 d_2，d_3 のクエリ尤度 $P(Q|M_{d_2})$，$P(Q|M_{d_3})$ を反映する必要がある．ここで，文書 d_2，d_3 の各々に算出されているクエリ全体の尤度及びクエリキーワードの尤度に注目する．文書 d_2 のクエリキーワードの尤度はそれぞれ，$P(\text{obama}|M_{d_2}) = 0.01$，$P(\text{family}|M_{d_2}) = 0.1$，$P(\text{tree}|M_{d_2}) = 0.1$ と算出されている．そしてその全ての積を算出する事で $P(Q|M_{d_2}) = 0.0001$ という値が算出されている．このクエリキーワードの尤度から文書 d_2 には，family と tree のクエリキーワードの出現確率は高く，obama のクエリキーワードの出現確率は低くなっている．よって文書 d_2 は family と tree との適合度が高い文書であると考えられる．一方，文書 d_3 のクエリキーワードの尤度はそれぞれ $P(\text{obama}|M_{d_3}) = 0.1$，$P(\text{family}|M_{d_3}) = 0.01$，$P(\text{tree}|M_{d_3}) = 0.1$ と算出されている．このクエリキーワードの尤度から文書 d_3 には，obama と tree のクエリキーワードの出現確率は高く，family のクエリキーワードの出現確率は低くなっている．よって文書 d_3 は obama と tree との適合度が高い文書であると考えられる．

ここで問題なのは，文書 d_3 のクエリ全体の尤度は文書 d_2 と同じ 0.0001 と算出されて点である．このような状況で過去の我々の手法を適用すると，クエリキーワードの出現確率が異なった文書であっても同じ特徴を持つ文書としてみなされる．図 1 の例では，オバマ米大統領についての記述が少なく，family と tree についての内容が含まれている文書 d_2 と，family に関する記述が少なく，オバマ米大統領と tree に関する記述がされている文書 d_3 を同じ特徴値で扱うことになる．よって隣接文書の特徴を正確に捉えるためには図 2 のように隣接文書の個々のクエリキーワードの尤度を検索対象文書のクエリキーワードの尤度に反映させる必要があると考える．

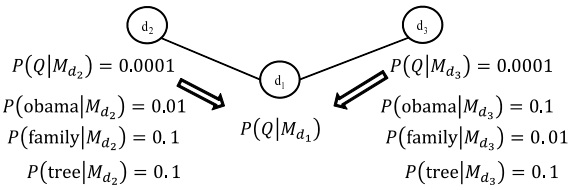

図 1　隣接文書のクエリ尤度を考慮

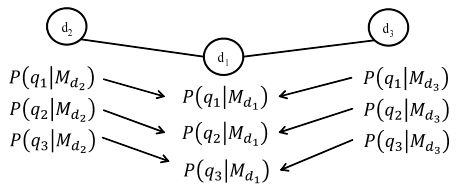

図 2　隣接文書のクエリキーワードの尤度を考慮

上記のような考えを基に我々は，リンクで接続された文書のクエリキーワードの尤度を考慮した検索対象文書のクエリ尤度算出法；**L**ink-**B**ased **L**anguage **M**odel (LBLM) を提案する．具体的に説明すると，我々の提案手法は従来のクエリ尤度モデルで用いられていた文書モデルとコーパスモデルから算出されたクエリキーワードの尤度に加え，検索対象文書の隣接文書集合内でのクエリキーワードの出現確率をリンクモデルとして算出する (図 3)．そしてこの三つの尤度をパラメタによって線形結合する手法であり，以下のような式によって表される．

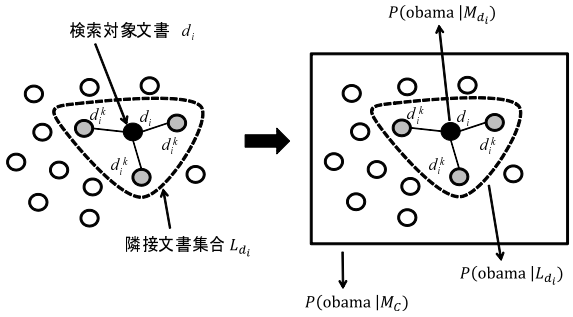

図 3　リンクで接続されている文書集合をリンクモデルとして定義

$$P(t_{ij}|M_{d_i}) = \lambda_1 P(t_{ij}|M_{d_i}) + \lambda_2 P(t_{ij}|L_{d_i}) + \lambda_3 P(t_{ij}|M_C) \quad (5)$$

ここで λ はパラメタであり，$\lambda_1 + \lambda_2 + \lambda_3 = 1$，$0.1 \leqq \lambda_1, \lambda_2, \lambda_3 \leqq 0.8$ の条件を満たす．よって，この条件を満たすパラメタの組合せは 36 通りとなる．

また，L はリンクによって接続された文書集合内での索引語の出現をモデル化したリンクモデルであり，$P(t_{ij}|L_{d_i})$ はリンクモデルによって算出される索引語 t_{ij} の尤度である．この尤度は以下のように算出される．

$$P(t_{ij}|L_{d_i}) = \frac{\sum_{d_i^k \in L_{d_i}} tf_{t_{ij}}^{d_i^k}}{\sum_{d_i^k \in L_{d_i}} N_{d_i^k}} \quad (6)$$

ここで d_i^k は，文書 d_i にリンクによって隣接している文書であり，その集合を L_{d_i} とする．また $N_{d_i^k}$ は 隣接文書 d_i^k の文書長である．

式 (5) によって算出された尤度を各クエリキーワードの尤度として扱い，2.1 で述べた式 (1) を用いて最終的なクエリの尤度を算出する．

4. 評価実験

本実験は，隣接文書のクエリキーワードの尤度を考慮したクエリ尤度算出手法の妥当性を評価する事を目的とする．この妥当性の評価を行うために，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．また，この比較実験の予備実験として，用いるリンクの違いによる検索精度比較及び，パラメータ設定による検索精度比較を行った．なお，比較対象として用いる手法は，隣接文書の内容を考慮しないクエリ尤度モデル (BaseLine) と我々が過去に提案した隣接文書のクエリ尤度を考慮した検索手法 (ST) [10] である．

4.1 テストコレクション

本実験では 2009 年の TREC (注1)(Text REtrival Conference) が提供している ClueWeb09 Dataset Category B [2] を

(注1)：http://trec.nist.gov/

用いた．TREC とは情報検索の研究分野に特化したワークショップであり，アメリカ国立標準技術研究所 (NIST) とアメリカ国防総省の共催で行われている．このワークショップには現在七つのタスクが存在しており，我々が使用するテストコレクションはその中の一つである Web Track のデータである．この Web Track の中でも二種類のタスクが存在する．一つは Adhoc タスク，もう一つは Diversity タスクである．Adhoc タスクは各クエリに対して一種類の解答文書集合が用意されており，それを用いて評価を行う．これに対して Diversity タスクはクエリに対して複数の subtopic を持ち，同じクエリでも異なった情報要求を満たせているかを判定するためのタスクである．本実験では特に Adhoc タスクに焦点あてて実験及び評価を行っている．

このテストコレクションには，2009 年時点でクロールされた約 5 千万の Web サイト (Unique URLs: 428,136,613，Total Outlinks: 454,075,638) とそれに対する 50 個の検索課題，そしてその検索課題に対する解答文書集合が用意されている．このクロールされた Web 文書にはリンク情報も記載されており，クロールされた 5,000 万文書内にリンク先が存在しているリンク全てを考慮することができる．そして最終的この検索課題に対して，各手法ごとに上位 1,000 件の文書リストを作成し，評価を行う．この検索対象文書に対して我々は，不要語リスト(注2)に基づいて不要語を取り除き，Porter Stemmer (注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．検索課題に関しては図 4 のような XML 文書で用意されている．これには検索要求である検索要求である query フィールド，その情報要求についての記述である description フィールド，そして diversity タスクで用いる subtopic フィールドによって構成されている．このような記述から我々は，クエリキーワードが記述されている query フィールドのみを抽出し検索を行う．

```
<topic number="1" type="faceted">
  <query>obama family tree</query>
  <description>
    Find information on President Barack Obama's family
    history, including genealogy, national origins, places and
    dates of birth, etc.
  </description>
  <subtopic number="1" type="nav">
    Find the TIME magazine photo essay "Barack Obama's
    Family Tree".
  </subtopic>
  <subtopic number="2" type="inf">
    Where did Barack Obama's parents and grandparents
    come from?
  </subtopic>
  <subtopic number="3" type="inf">
    Find biographical information on Barack Obama's mother.
  </subtopic>
</topic>
```

図 4　Web Track における検索課題の記述例

また，評価尺度として本実験では精度 (Prec.) と解答文書取得数 (Rel.Retr.) を用いる．精度は以下の式によって算出される．

$$\text{精度} = \frac{\text{検索された解答文書数}}{\text{検索された文書数}} \tag{7}$$

特に精度に関しては，検索結果の上位 10 件の精度 (P@10)，上位から 100 件ごとの精度 (0.0 - 1.0)，平均精度 (MAP) を算出している．また解答文書取得数は検索することができた解答文書数を意味する．

4.2　異なる二種類のリンクを用いた検索結果比較

各 Web 文書には，その文書から他の文書へリンクするアウトリンクと，他の文書からリンクされているインリンクの二種類が存在する．アウトリンクはある文書の作成者自身から意図的にリンクを作成し他の文書へつなげるが，インリンクは他の文書作成者からリンクを生成されるのでリンクされた文書の作成者の意思とは無関係に生成されるものである．このような異なった特徴をもつリンクが存在している中で，どのようなリンクを用いて隣接文書として定義するれば最適であるかが問題となる．よってここでは用いるリンクを変化させて隣接文書を定義する事によって，提案手法の検索結果にどのような影響を与えるのかを確かめるための実験を行う．具体的には，4.1 で述べた Web Track のデータを用い，提案手法である LBLM 法に対して隣接文書を定義する際に，インリンクのみを用いて定義する手法，アウトリンクのみを用いて定義する手法，双方のリンクを用いて定義する手法における検索結果の比較を行う．比較するデータは，前述した三つのリンクの用い方について，36 種類のパラメタの組合せ全てに対して検索結果を算出し，それぞれの結果に対して平均精度と解答文書取得数の基本統計量を用いる．各手法の実験結果に対する統計量を表 1〜3 に示す．

表 1　インリンクを用いた LBLM 法による検索結果の基本統計量

MAP. Inlink	
平均	0.2174
標準偏差	0.03404
最小	0.1540
最大	0.2593

Rel.Retr. Inlink	
平均	4,727
標準偏差	754.8
最小	3,213
最大	5,557

表 2　アウトリンクを用いた LBLM 法による検索結果の基本統計量

MAP. Outlink	
平均	0.2518
標準偏差	0.02054
最小	0.1825
最大	0.2638

Rel.Retr. Outlink	
平均	5,469
標準偏差	152.6
最小	4,980
最大	5,580

表 3　インリンク，アウトリンクを用いた LBLM 法による検索結果の基本統計量

MAP. In, Outlink	
平均	0.2176
標準偏差	0.03399
最小	0.1541
最大	0.2594

Rel.Retr. In, Outlink	
平均	4,735
標準偏差	751.9
最小	3,228
最大	5,562

(注2)：ftp://ftp.cs.cornell.edu/pub/smart/english.stop
(注3)：http://tartarus.org/~martin/PorterStemmer/

比較実験の結果，平均精度及び解答文書取得数の平均値においてもっとも良い結果であったのはアウトリンクのみを用いる手法であった．このアウトリンクのみを用いる手法は，このほかの最小値や最大値においても，他の二手法と比較して良い結果となった．また，各手法に対して標準偏差においてはインリンクのみを用いた手法，及びインリンクとアウトリンク双方を用いる手法が大きな値をとっている．これは，LBLM を用いる際のパラメタの設定方法によって検索結果に大きな差が生まれるということである．この点に関してアウトリンクのみを用いる手法は，他の手法に比べて標準偏差が低い．よって，アウトリンクのみを用いる手法は，どのようなパラメタを設定しても比較的情報要求に沿った質の良い検索結果を提示していると考えられる．以降 4.3，4.4 節の実験においては，隣接文書をアウトリンクのみで接続されている文書として定義する．

4.3 パラメータ設定

3. で述べた LBLM 法では，パラメタの設定が必要となる．よってここでは LBLM 法に用いているパラメタ λ_1，λ_2，λ_3 の設定について述べる．

LBLM 法のパラメタは λ_1，λ_2，λ_3 であり，それぞれ文書内でのクエリキーワードの尤度，リンクモデル内でのクエリキーワードの尤度，文書集合全体でのクエリキーワードの尤度の考慮する度合いを調節する役割を持っている．ここでは 3. で述べたように，$\lambda = 0.1$ から $\lambda = 0.8$ までを考慮し，$\lambda_1 + \lambda_2 + \lambda_3 = 1$ の条件を満たすパラメタの組合せである 36 通りの実験結果を提示する．この実験結果について，各パラメタにおける平均検索精度と解答文書取得数を表 4 に示す．なおここでは用いる LBLM 法には，隣接文書集合は各文書からのアウトリンクのみ接続している文書集合として扱っている．

実験の結果，平均検索精度及び解答文書取得数においてもっとも良い結果であったパラメタの組合せは $\lambda_1 = 0.4$，$\lambda_2 = 0.1$，$\lambda_3 = 0.5$ であった．また全ての組合せについて比較してみると，精度及び解答文書取得数の良い組合せについては，リンクモデルのパラメタが 0.5 以下のものに多く見られた．よって，リンクモデルを用いる際は文書内でのクエリキーワードの尤度や文書集合全体でのクエリキーワードの尤度よりも小さな影響度を与えて反映させることによって，検索システムの性能の向上につながるということが分かった．以降，4.4 の実験においては，LBLM 法には $\lambda_1 = 0.4$，$\lambda_2 = 0.1$，$\lambda_3 = 0.5$ を採用する．

4.4 実 験 結 果

実験結果を表 5, 6 に示す．この表には，上述した評価尺度での各手法の評価，及びその結果に対してウィルコクソンの符号順位和検定を行い有意差があるものに対して * で示している．

まず LBLM と BaseLine の比較では，3.72 % の平均精度の向上，5.78% の再現率の向上が確認する事ができた．この結果に対して，有意差検定を信頼区間 95 % で行った結果，平均精度に対しては有意差を確認することができたが，再現率に対しては有意差を確認する事ができなかった．

一方 LBLM と ST の比較では，1.10 % の平均精度の向上，1.51 % の解答文書取得数の向上を確認する事ができた．この結果に対しても同様に検定を行った結果，こちらでは解答文書取得数に対しては有意差を確認する事ができたが，平均精度に対しては有意差を確認する事はできなかった．

このような結果から，BaseLine に対しては検索結果上位及び，平均精度に対して有用性を，また ST 法に対しては解答文書取得数に対して有用性を示すことができたと考える．この結果は，アウトリンクによって接続された文書の内容まで考慮することによって得られた結果であると考える．よって，アウトリンクによる隣接文書の内容を考慮して Web 文書を検索する事によって，従来手法より多くの解答文書を取得できることが明らかになった．

表 5　BaseLine と LBLM の検索結果比較

	BaseLine	LBLM	Improvement(%)
Rel.	12,544	12,544	
Rel.Retr	5,275	5,580	5.78%
Prec.			
P@10	0.7896	0.7979	1.06% *
0.0	0.9153	0.9143	-0.11%
0.1	0.3846	0.3990	3.74% *
0.2	0.2703	0.2838	4.97% *
0.3	0.2244	0.2383	6.22% *
0.4	0.1917	0.2056	7.26%
0.5	0.1691	0.1809	6.97%
0.6	0.1513	0.1616	6.77%
0.7	0.1376	0.1455	5.75% *
0.8	0.1265	0.1331	5.21% *
0.9	0.1163	0.1228	5.57%
1.0	0.1082	0.1146	5.87%
MAP.	0.2541	0.2636	3.72% *

表 6　ST と LBLM の検索結果比較

	ST	LBLM	Improvement(%)
Rel.	12,544	12,544	
Rel.Retr	5,497	5,580	1.51% *
Prec.			
P@10	0.7646	0.7979	4.36%
0.0	0.8932	0.9143	2.36%
0.1	0.4046	0.3990	-1.39%
0.2	0.2833	0.2838	0.15% *
0.3	0.2357	0.2383	1.12%
0.4	0.2034	0.2056	1.08% *
0.5	0.1786	0.1809	1.28% *
0.6	0.1598	0.1616	1.09% *
0.7	0.1442	0.1455	0.95% *
0.8	0.1313	0.1331	1.35%
0.9	0.1209	0.1228	1.55%
1.0	0.1128	0.1146	1.53% *
MAP.	0.2607	0.2636	1.10%

4.5 計算量についての考察

ここでは，本提案手法の計算量について，2.2. で述べた Liu らのクラスタモデルを用いた研究 [6] と比較し，考察する．ここで比較する処理は，各文書の周辺文書を取得する際に行う処

表 4　パラメタの変化による LBLM の検索精度比較

λ_1, λ_2, λ_3	MAP	Rel.Retr.
0.1, 0.1. 0.8	0.2587	5,393
0.1, 0.2, 0.7	0.2502	5,376
0.1, 0.3, 0.6	0.2435	5,346
0.1, 0.4, 0.5	0.2341	5,292
0.1, 0.5, 0.4	0.2182	5,212
0.1, 0.6, 0.3	0.2041	5,142
0.1, 0.7, 0.2	0.1931	5,067
0.1, 0.8, 0.1	0.1825	4,980
0.2, 0.1, 0.7	0.2627	5,526
0.2, 0.2, 0.6	0.2622	5,525
0.2, 0.3, 0.5	0.2594	5,514
0.2, 0.4, 0.4	0.2534	5,512

λ_1, λ_2, λ_3	MAP	Rel.Retr.
0.2, 0.5, 0.3	0.2506	5,494
0.2, 0.6, 0.2	0.2465	5,446
0.2, 0.7, 0.1	0.2423	5,400
0.3, 0.1, 0.6	0.2633	5,561
0.3, 0.2, 0.5	0.2635	5,563
0.3, 0.3, 0.4	0.2630	5,559
0.3, 0.4, 0.3	0.2615	5,537
0.3, 0.5, 0.2	0.2574	5,533
0.3, 0.6, 0.1	0.2543	5,508
0.4, 0.1, 0.5	0.2636	5,580
0.4, 0.2, 0.4	0.2634	5,573
0.4, 0.3, 0.3	0.2636	5,565

λ_1, λ_2, λ_3	MAP	Rel.Retr.
0.4, 0.4, 0.2	0.2634	5,547
0.4, 0.5, 0.1	0.2629	5,540
0.5, 0.1, 0.4	0.2628	5,571
0.5, 0.2, 0.3	0.2631	5,572
0.5, 0.3, 0.2	0.2630	5,563
0.5, 0.4, 0.1	0.2622	5,544
0.6, 0.1, 0.3	0.2627	5,565
0.6, 0.2, 0.2	0.2630	5,569
0.6, 0.3, 0.1	0.2622	5,548
0.7, 0.1, 0.2	0.2618	5,571
0.7, 0.2, 0.1	0.2610	5,544
0.8, 0.1, 0.1	0.2600	5,559

理である．

Liu らの研究では，周辺文書を取得するために文書集合に対して k-means クラスタリングを行っている．k-means クラスタリングでは，文書全体をクラスタリングするために各文書間距離を全ての組合せにおいて計算する必要がある．よって，検索対象文書全体の数が n とするとクラスタリングによって周辺文書を取得するためには $O(n^2)$ の計算量が必要となる．これに対して我々の手法は各文書に存在しているリンク情報を収集する事によって周辺文書を取得することが可能である．4.1. で述べたテストコレクションでは，各文書には平均 1 つのアウトリンクが存在していることになるので，ここで k-means クラスタリングの文書間の距離を算出する計算量と，Web 文書からリンクを抽出する計算量が同じと仮定すると，我々の手法での周辺文書取得の際に必要な計算量は $O(n)$ となる．よって，Liu らの手法に対して計算量の観点から優位であるといえる．

5. ま と め

本研究では，隣接文書のクエリキーワードの尤度を考慮した言語モデルによる検索手法の提案を行った．複数の予備実験の結果，リンクモデルの影響度を低い値に設定する事によって精度の向上したこと，そして考慮すべきリンクはアウトリンクのみを用いた場合がもっとも良い検索結果を取得することができることを確認した．評価実験の結果，BaseLine と我々が過去に提案した手法との比較において有効性を示すことができた．

今後は，リンクによって接続された文書のみを考慮するのではなく，クエリに対して最適な隣接文書を得るために，文書間距離の概念の適用や情報粒度の概念を適用する必要があると考える．

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 若手研究 (B) (課題番号：22700248) によるものである．ここに記して謝意を表す．

文　献

[1] Graham Bennett, Falk Scholer, and Alexandra Uitdenbogerd. A comparative study of probabilistic and language models for information retrieval. In *Proceedings of the nineteenth conference on Australasian database - Volume 75*, ADC '08, pp. 65–74, Darlinghurst, Australia, Australia, 2007. Australian Computer Society, Inc.

[2] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. Overview of the trec 2009 web track. In *Text Retrieval Conference(TREC)*, 2009.

[3] Frederick Jelinek and Robert L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceeding of the Workshop on Pattern Recognition in Practice*, pp. 381–397, 1980.

[4] Jon M. 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.

[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] Xiaoyong Liu and William B. Croft. Cluster-based retrieval using language models. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 186–193, New York, NY, USA, 2004. ACM.

[7] 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.

[8] Fei Song and William B. 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.

[9] 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, 2004.

[10] Koya Tamura, Kenji Hatano, and Hiroshi Yadohisa. Characterizing web pages based on the query likelihoods of neighboring pages. In *Proceedings of the 5th International Conference on Digital Information Management (ICDIM 2010)*, pp. 392–397, 2010.

[11] 江口浩二．情報検索のための確率的言語モデルに関する動向と課題．電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 3, pp. 157–169, 2010.