Web検索におけるアンカーテキストのモデル化と質問の自動分類
Modeling Anchor Text and Classifying Queries in Web Retrieval

藤井 敦
Atsushi Fujii
筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba

概要　Webを検索するユーザの情報要求は多種多様であり，検索質問の種類によって必要とされる検索手法が異なる．本研究は，検索質問を「調査型（ある事項に関する一般的な調べ物が目的）」と「誘導型（ある事項に関するトップページを探すことが目的）」に自動分類し，その結果に基づいて検索手法を動的に変更する手法を提案する．調査型の検索質問にはページの本文を用いた一般的な検索手法が有効である．それに対して，誘導型の検索質問にはアンカーテキストを用いたWeb特有の検索手法が有効である．そこで，本研究はアンカーテキストを用いた検索手法も提案する．NTCIRのWeb検索用テストコレクションを用いた実験によって提案手法の有効性を示す．

1　はじめに

World Wide Webには多種多様な情報が存在する．そこで，Webを検索するユーザの情報要求もまた多種多様である．Broder [3]は，Web上の検索質問をユーザの情報要求に基づいて以下の3種類に分類した．

- informational query（調査型の検索質問）

  ある話題についてWeb上の情報を調査するために使用される検索質問である．Web以外の一般的なテキスト検索でも使用される．

- navigational query（誘導型の検索質問）

  ある事柄（人物，組織，商品，イベントなど）に関するトップページや代表的なページを検索するために使用される検索質問である．

- transactional query（取引型の検索質問）

  ソフトウェアのダウンロードやオンラインショッピングなどのように，Webサイトを中継して別の実体に到達するための検索質問である．

Broder [3]は，ユーザに対するアンケート調査と検索エンジンのログ解析によって，各種類の質問が無視できない一定の割合で存在することを示した．

近年の情報検索研究によって，上記3種類の検索質問に対して，必要とされる検索の手法が根本的に異なることが明らかになっている．TREC[1]やNTCIR[2]などのワークショップでは，調査型と誘導型の検索質問を対象としたWeb検索のテストコレクションが構築された．テストコレクションとは，情報検索システムの精度を評価するためのベンチマークである．上記のテストコレクションを用いた種々の実験によって，調査型の検索質問にはページの本文を用いた検索が有効であり，誘導型の検索質問にはアンカーテキストやリンク構造を用いた検索が有効であるという知見が得られている [4, 5, 6, 9, 14] ．

アンカーテキストとは，あるページから別のページにリンクをはるときに使用される文字列である．アンカーテキストは，あるページに対して第三者がどのように評価しているのかを知る上で重要である．なお，リンクとは，ページ間の引用関係に基づく構造であり，アンカーテキストとは異なる．

Liら [13]は，人手で作成した規則によって取引型のページとそれ以外のページを自動的に分類し，取引型の検索質問に対する検索精度を向上させた．すなわち，ページ本文やアンカーテキストとは異なる検索手法が効果的であった．

しかし，ユーザは検索質問は入力しても，検索質問の種類は明示しない，あるいはできない．そこで，

[1] http://trec.nist.gov/
[2] http://research.nii.ac.jp/ntcir/index-ja.html

様々な検索質問に対して高い検索精度を実現するためには，入力された検索質問の種類を自動的に特定し，異なる検索手法を選択的に使用する必要がある．

本研究は，検索質問を自動的に分類し，その結果に基づいて検索手法を動的に変更する手法を提案する．さらに，本研究は，アンカーテキストを用いた検索モデルを提案する．ページ本文を用いた検索モデルは，Web 検索以外の情報検索研究によって一定の成果が得られている．それに比べて，アンカーテキストを用いた検索は研究の歴史が浅いために改善の余地がある．ただし，取引型を対象としたテストコレクションが存在しないため，本研究は調査型と誘導型の検索質問を対象とする．

2 章で筆者が開発した Web 検索システムの全体像について説明する．3 章でアンカーテキストを用いた検索モデルを提案し，4 章で検索質問の分類手法を提案する．5 章で提案手法を評価するための実験と結果について説明する．

2 Web 検索システムの概要

本研究で開発した Web 検索システムの概要を図 1 に示す．図 1 は，大きく分けて,「検索質問分類」,「コンテンツ検索」,「アンカー検索」で構成されている．

システムへの入力は，キーワード集合や自然文による検索質問である．検索質問が自然文の場合は，索引語を抽出して検索に利用する．システムの出力は，検索質問に合致するページの順位付きリストである．ページの順位は，検索質問に対するスコアによって決定される．

まず，検索質問が入力されると，検索質問分類によって「調査型」か「誘導型」に分類される．次に，検索質問の種類に依らずに，コンテンツ検索とアンカー検索を独立に行って 2 つの順位付きリスト（初期リスト）を作成する．コンテンツ検索はページ本文を検索に利用し，アンカー検索はアンカーテキストを検索に利用する．コンテンツ検索には，Web 検索以外の情報検索でも用いられている Okapi BM25 [15] を用いる．アンカー検索には 3 章で提案する検索モデルを用いる．どちらの検索モデルも，検索されたページに対してスコアを計算し，スコアに基づいてページに順位を付ける．

最後に，2 つの初期リストを併合して出力する．こ

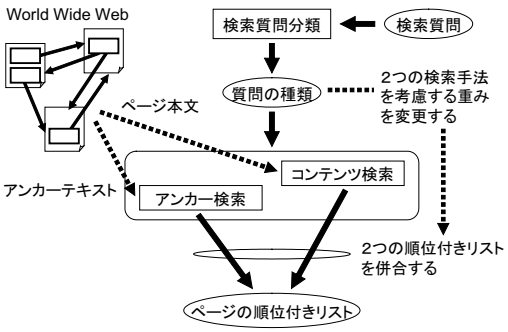

図 1: Web 検索システムの概要

こで，検索質問の種類によって，コンテンツ検索とアンカー検索の結果をそれぞれどの程度重視するかを決定する．具体的には，検索されたページ d に対して，2 つの初期リストにおけるスコアを加重平均し，新しいスコア $S(d)$ を計算する．最終的な順位付きリストにおいて，各ページは $S(d)$ に基づいて降順に配列される．

しかし，コンテンツ検索とアンカー検索で計算されるスコアは互いに意味や範囲が異なるため，単純に加重平均を取ることことはできない．そこで，ページ d の初期スコアにおける「順位の逆数」をとり，これをページ d のスコアとする．さらに，式 (1) によって最終的なスコア $S(d)$ を計算する．

$$S(d) = \frac{\alpha}{R_c(d)} + \frac{1-\alpha}{R_a(d)} \quad (0 \leq \alpha \leq 1) \qquad (1)$$

式 (1) において，$R_c(d)$ と $R_a(d)$ は，それぞれ，コンテンツ検索とアンカー検索で得られた初期リストにおける d の順位である．α は 0 以上 1 以下の値を取り，検索質問の種類によって変更する必要がある．α の値は，調査型の質問に対しては 0.5 よりも大きな値に設定し，誘導型の質問に対しては 0.5 よりも小さな値に設定する．ただし，本研究で提案する検索質問の分類手法は，検索質問の種類を決定する際に α の値も決定する．そこで，α の値を人手で決定する必要はない．

以下，図 1 の「アンカー検索」と「検索質問分類」について，3 章と 4 章で個別に詳説する．

3 アンカー検索モデル

3.1 先行研究

Web 検索に関する先行研究では，リンクやアンカーテキストなどの「ページ本文以外の情報」を用いた検索モデルがいくつか提案されている．

Yang [18] は，コンテンツ検索とリンクによる検索を統合する手法を提案した．リンクによる検索では，HITS [11] を用いてリンク構造に基づいてページの順位付けを行った．しかし，アンカーテキストによる検索は行っていない．Craswell ら [4] は，あるページ d にリンクしているアンカーテキストの集合を d の代理情報と見なしてコンテンツ検索を行った．Westerveld ら [17] はアンカーテキストをモデル化して Web 検索に利用した．Westerveld らの手法は，本研究の提案手法に最も類似する．両手法の相違点については 3.2 節で詳説する．

3.2 手法

本研究で提案するアンカー検索のモデルは，検索質問 q が与えられた条件のもとでページ d が検索される確率 $P(d|q)$ を計算し，この確率でページを順位付ける．ベイズの定理を用いて $P(d|q)$ を式 (2) のように変形する．

$$\begin{aligned} P(d|q) &= \frac{P(q|d)\cdot P(d)}{P(q)} \\ &\propto P(q|d)\cdot P(d) \end{aligned} \tag{2}$$

式 (2) において $P(q)$ は全ページに共通の定数であり，ページ間の相対的な順序には影響しない．そこで，$P(q)$ は無視する．$P(d)$ は検索質問とは無関係にページ d が選択される確率である．$P(d)$ は最尤推定によって求める．すなわち，d にリンクしているアンカーテキストの数と Web コレクション全体におけるリンク数の割合を $P(d)$ とする．そこで，多くのページからリンクされているページほど大きな値を取る．

なお，$P(d)$ として PageRank [2] を使うこともできる．PageRank はユーザがリンクを辿りながらページ d に到達する確率 $P(d)$ を計算するからである．しかし，予備実験の結果，最尤推定の方が検索精度が高かったため，以降は最尤推定だけを考慮する．

検索質問 q は複数の索引語（ターム）で構成されることがある．そこで，索引語の独立性を仮定して，式 (3) によって $P(q|d)$ を計算する．

$$P(q|d) = \prod_{t\in q} P(t|d) \tag{3}$$

$P(t|d)$ は，ページ d にリンクしているアンカーテキストから無作為に一つの索引語を選択したときに，それが t である確率である．ChaSen[3]を用いてアンカーテキストを形態素解析し，名詞，動詞，形容詞，辞書未登録語，記号を索引語として利用する．

以上より，誘導型検索では $P(t|d)$ の計算が重要である．Westerveld ら [17] は，ページ d にリンクしているアンカーテキストをまとめて一つの「代理文書」として扱い，$P(t|d)$ を計算した．すなわち，$P(t|d)$ は，ページ d の代理文書から無作為に索引語を一つ選択したときに，それが t である確率である．具体的には，ページ d の代理文書における索引語 t の相対頻度によって $P(t|d)$ を計算する．

しかし，この手法には問題がある．例えば，日本語 Yahoo!のトップページ[4]が「ヤフー」と「Yahoo Japan」という 2 つのアンカーテキストでリンクされていた場合に,「ヤフー」,「Yahoo」,「Japan」という 3 つの索引語に対して，$P(t|d)$ の値は同じになる．すなわち，これら 3 つの索引語は日本語 Yahoo!のトップページを検索するために等しく重要であることになる．しかし，実際には「ヤフー」は単体で日本語 Yahoo!のトップページを検索するための重要な索引語であるのに対して,「Yahoo」と「Japan」はどちらか片方だけでは重要度が低くなるはずである．

また，ページ d が非常に長いアンカーテキストでリンクされていると，ページ d に対してリンクしている別のアンカーテキストの影響が低くなる．すなわち，第三者が索引語の重要性を意図的に変更することができてしまうため，スパムに弱い．

以上の問題点を解消するために，本手法では，索引語の確率を代理文書の単位で計算するのではなく，アンカーテキストの単位で計算する．具体的には，$P(t|d)$ を式 (4) のように分解する．

$$P(t|d) = \sum_{a\in \mathbf{A}_d} P(t|a)\cdot P(a|d) \tag{4}$$

ここで，$\mathbf{A}_d$はページ d にリンクしているアンカーテキストの集合である．式 (4) の右辺では，索引語の確率は $P(t|a)$ で計算される．ここで，$P(t|d)$ に対し

[3] http://chasen.naist.jp/hiki/ChaSen/
[4] http://www.yahoo.co.jp/

て直感的な解釈を与える．$P(a|d)$ は，ページ d にリンクする際によく使われるアンカーテキストに対して大きな値をとる．$P(t|a)$ は，アンカーテキスト a でよく使われる索引語に対して大きな値をとる．そこで，$P(t|d)$ は，ページ d へのリンクによく使われるアンカーテキストに頻出する索引語 t に対して大きな値をとる．

上述した日本語 Yahoo!の例では，「ヤフー」の $P(t|a)$ は 1 であるのに対して,「Yahoo」と「Japan」の $P(t|a)$ はそれぞれ 1/2 となり,「ヤフー」よりも重要度が低くなるため，現実に合っている．

しかし，ページ d に対するアンカーテキストの中にユーザが入力した質問中の索引語が存在しない場合は，$P(t|d)$ がゼロになる．その結果，式 (3) の計算結果は他の索引語によらずゼロになってしまう．このような場合は平滑化（スムージング）が必要である．本手法は式 (5) を用い，索引語 t の同義語 s によって $P(t|d)$ を近似する．

$$\begin{aligned} P(t|d) &= P(t|s,d) \cdot P(s|d) \\ &\approx P(t|s) \cdot P(s|d) \end{aligned} \tag{5}$$

式 (5) において，$P(s|d)$ は式 (4) を用いて計算する．

本手法は，索引語の同義語を検出するために，同じページにリンクしている複数のアンカーテキストを利用する．あるページが複数のアンカーテキストからリンクされている場合は，各アンカーテキストの文字列が表層的に異なっていたとしても互いに同じような内容を示していることが多い．

例えば，Google の日本語トップページ[5] にリンクしている主なアンカーテキストは,「グーグル」,「google」,「検索エンジン」などである．そこで，同じページにリンクしている複数のアンカーテキストから，カタカナによって翻字（音訳）された言葉と元の言葉を同義語として抽出する．具体的には，筆者が提案した統計的な翻字手法 [7] を利用し，与えられた 2 つの言葉の一方から他方に翻字することができるかどうか検査する．もし翻字できれば，両者を同義語として検出する．上述の例では,「グーグル」と「google」が同義語として検出される．しかし,「検索エンジン」に対する同義語は検出されない．以上の手法で同義語を抽出し，式 (6) によって $P(t|s)$ を計算する．

$$P(t|s) = \frac{F(t,s)}{\sum_{r \neq s} F(r,s)} \tag{6}$$

[5] http://www.google.co.jp/

ここで，$F(t,s)$ は t と s が別のアンカーテキストで使用され，かつ，それらのアンカーテキストが同じページにリンクしている頻度である．

同義語を用いても $P(t|d)$ が計算できない場合は，$P(t)$ を用いて平滑化する．$P(t)$ は全アンカーテキストから無作為に一つの索引語を選択したときに，それが t である確率である．

4 検索質問分類

4.1 先行研究

Kang ら [9] は，TREC の 10 ギガバイト Web テストコレクションを用いて，検索質問を調査型と誘導型に自動分類する手法を提案した．しかし，対象のテストコレクションは 10 ギガバイトという小さな規模である．また，検索質問の分類に焦点が当てられており，分類によって検索精度が向上したかどうかは明らかになっていない．

Lee ら [12] は，被験者の Web 検索行動を観察することによって，検索質問を調査型と誘導型に分類するための特徴量を提案した．しかし，検索実験は行っていない．そこで，Lee らが提案した特徴量を用いて検索質問を分類することによって，Web 検索の精度がどのように変化するかは明らかになっていない．

以上の背景から，本研究は，NTCIR-3 と NTCIR-4 で構築された 100 ギガバイトのテストコレクションを用いて，Web 上の検索質問を調査型と誘導型に自動分類することを目的とする．当該テストコレクションは，Kang らが使用したテストコレクションよりも 10 倍大きく，現実の Web 検索により近い環境での実験を可能とする．さらに，本研究では，検索質問を自動分類することによって，Web 検索の精度がどのように変化するかを明らかにする．

本研究で提案する手法は，Baeza-Yates ら [1] が提案した検索質問分類の手法とは異なり，大量の検索ログを必要としない．大学や研究所で Web 検索の研究を行う場合は，商用の検索サービスとは異なり，大量の検索ログを入手することが困難であることが多い．そのような状況においても本研究で提案する分類手法は適用可能である．

4.2 手法

誘導型の検索質問によって検索されるべきページとは，ある事柄に関するトップページや代表的なページである．そのようなページは，他のページから，特定のアンカーテキストによってリンクされることが多い．例えば，筑波大学のトップページ[6]にリンクする場合は,「筑波大学」がアンカーテキストとして使われることが多いだろう．他方で，筑波大学と無関係なページにリンクする場合は,「筑波大学」がアンカーテキストとして使われることは稀であろう．

しかし，特定のトップページと対応しない一般的な事柄の場合には，上記のような現象は起こりにくい．例えば,「情報検索」というアンカーテキストでリンクされるページは，検索サービスのページや情報検索に関する解説のページなど多岐にわたる．

以上の仮説に基づいて，Lee ら [12] は，検索質問の文字列と同じアンカーテキストからリンクされているページの出現分布（anchor-link distribution: ALD）を分析してヒストグラムを作り，ALD の歪度（skewness）によって検索質問の種類を特定した．歪度が小さい場合は調査型に分類し，大きい場合は誘導型に分類する．図 2 の (a) と (b) は，それぞれ誘導型と調査型の検索質問に対するヒストグラムの例であり，(a) は ALD の歪度が大きい．しかし，Lee らは，検索質問の文字列がそのままの形でアンカーテキストとして使用されている場合だけを対象とした．そこで，検索質問と完全一致するアンカーテキストがなければ ALD を分析することができない．

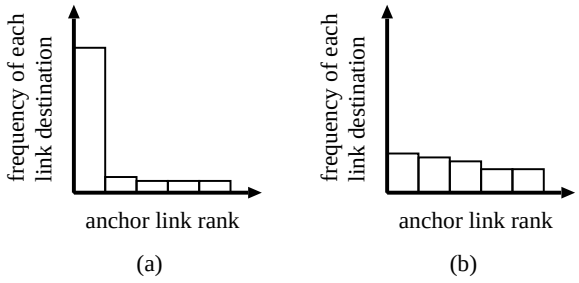

図 2: anchor-link distribution の例

本研究は，この問題を解消するために，検索質問がそのままの形でアンカーテキストとして使われていない場合には，検索質問を索引語に分割して個別に ALD を分析する．以下，本研究で提案する分類手法について説明する．検索質問 q を構成する索引語の集合を $\mathbf{T}_q$ とする．ChaSen を用いて q を形態素解析し，名詞を索引語として使用する．索引語 $t \in \mathbf{T}_q$ に対して，t を含むアンカーテキストによってリンクされているページの集合を $\mathbf{D}_t$ とする．ここで，t とアンカーテキストが同一である必要はない．$\mathbf{D}_t$ から無作為に選んだページが d である確率 $P(d|t)$ の分布を分析し，式 (7) によってエントロピーを計算する．

$$H(\mathbf{D}_t|\mathbf{T}_q) = -\sum_{t\in\mathbf{T}_q} P(t) \sum_{d\in\mathbf{D}_t} P(d|t)\log P(d|t) \tag{7}$$

q に関する ALD の歪度が大きいほど，$H(\mathbf{D}_t|\mathbf{T}_q)$ は小さくなり，q は誘導型に分類されやすくなる．

検索質問は少数のキーワードで構成されることが多いため，検索質問における t の出現頻度は考慮しない．すなわち，$P(t) = \frac{1}{|\mathbf{T}_q|}$ とする．

t を含むアンカーテキストが存在しない場合は，アンカーテキスト集合から t の同義語を自動抽出し，同義語が存在すれば，t の代わりに使用する．アンカーテキスト集合からの同義語抽出には，3.2 節で提案した手法を使用する．

実際には，$P(d|t)$ を計算する際に，複数のページを一定の階級にまとめる方が分類の精度が向上する．

さらに，$H(\mathbf{D}_t|\mathbf{T}_q)$ を $\log|\mathbf{D}_t|$ で割って，0 以上 1 以下の範囲に正規化した値を $i(q)$ とする．$i(q)$ を式 (1) の α とすることで，$i(q)$ の値が大きい q に対してはコンテンツ検索の結果が重視される．そこで，システム内で q の種類を具体的に決定する必要はない．

検索質問の種類を具体的に決定する必要がある場合は，$i(q)$ の値が 0.5 以上の場合は q を調査型に分類し，それ以外の場合は q を誘導型に分類する．

ただし，本手法には 2 つのパラメタがある．一つは，複数のページをまとめるための階級幅であり，現在は 5 としている．また，t がアンカーテキストに出現しない場合は，本来 ALD の歪度が小さいにも拘らず，全ての $d \in \mathbf{D}_t$ に対して $P(d|t)$ が 0 になるため，$H(\mathbf{D}_t|\mathbf{T}_q)$ が小さくなる．このような場合は，t を含むアンカーテキストから N 件のページに対して均一にリンクがはられていると見なす．現在は $N = 10000$ としている．これら 2 つのパラメタは，予備実験を通して経験的に設定した．今後は，対象とするテストコレクションの規模に応じて適切に設定するための基準が必要である．

[6] http://www.tsukuba.ac.jp/

5 評価実験

5.1 実験データと評価尺度

NTCIR-3 と NTCIR-4 で構築された Web 検索のテストコレクションを使用して提案手法の評価実験を行った．NTCIR-3 では調査型の検索課題が作られた [6]．NTCIR-4 では調査型と誘導型の検索課題が作られた [5, 14]．NTCIR-3 と NTCIR-4 の検索対象は共通であり，JP ドメインから収集した 100 ギガバイト（約 1 千万ページ）のページ集合である．

正解のレベルには「適合」と「部分適合」がある．本実験では,「適合」だけを正解として使用した．適合文書が存在する検索課題の件数は，NTCIR-3 の調査型が 47 件，NTCIR-4 の調査型が 80 件，NTCIR-4 の誘導型が 168 件である．課題あたりの適合文書数は，調査型に対して平均 81.3 件であり，誘導型に対して平均 1.79 件である．合計 295 件の検索課題を用いて，100 ギガバイトのページ集合を検索する実験を行った．

図 3 と図 4 に調査型と誘導型の検索課題を 1 件ずつ示す．図 3 と図 4 は NTCIR-3 と NTCIR-4 からの抜粋である．NTCIR-3 と NTCIR-4 の課題に含まれる項目は完全に同一ではない．図 3 と図 4 では，両方の課題に共通の項目だけを示している．1 つの課題は<TOPIC>や<TOPICS>で括られており，<NUM>は課題番号を示す．情報要求に関する記述には<TITLE>, <DESC>，<NARR>の 3 種類がある．<TITLE>は 1 つ以上のキーワードであり，<DESC>はフレーズや文である．<NARR>は複数の文からなる文章であり，背景情報などが<BACK>のようなタグで適宜示されている．<USER>は課題作成者に関する情報である．

```
<TOPIC>
<NUM>0010</NUM>
<TITLE>オーロラ，条件，観測</TITLE>
<DESC>観測のために、オーロラの発生する条件
が知りたい</DESC>
<NARR><BACK>オーロラを観測するために、発生
に必要な条件や、発生のメカニズムが知りたい。
</BACK> ･･･（略）･･･ </NARR>
<USER>大学院修士 1 年，女性，検索歴 2.5 年
</USER>
</TOPIC>
```

図 3: NTCIR-3「調査型」検索課題の例

```
<TOPICS>
<NUM>0015</NUM>
<TITLE>ゼンリン</TITLE>
<DESC>ゼンリンという会社について調べたい
</DESC>
<NARR><BACK>電子地図が有名だから、もっと知
りたい。</BACK></NARR>
<USER>大学院修士 1 年，男性，検索歴 4 年
</USER>
</TOPICS>
```

図 4: NTCIR-4「誘導型」検索課題の例

検索課題のどの項目を使用して，検索質問をどのように構成するのかは実験の目的によって異なる．本実験では，Web 検索エンジンの検索窓に入力されるキーワードを模倣するために，検索課題の<TITLE>に記述されたキーワードを検索質問として使用した．検索質問あたりのキーワード数は，調査型に対して平均 2.57 であり，誘導型に対して平均 1.39 である．図 5 と図 6 に調査型と誘導型の検索質問を課題番号が若い順に 10 件ずつ示す．テストコレクション構築の過程で削除された検索課題があるため，課題番号は連番ではない．また，課題番号はテストコレクションごとに独立して付けられるため，図 5 と図 6 で同じ番号の課題どうしに関係はない．

課題番号	検索質問
0008	サルサ, 学ぶ, 方法
0010	オーロラ, 条件, 観測
0011	遣唐使, 習慣, 文化
0012	正月, 雑煮, 地方
0013	京都, 寺, 神社
0014	夢, 将来, 努力
0015	オゾン層, オゾンホール, 人体
0016	ゲノム, 創薬, 動向
0017	野球, ベースボール, 比較

図 5:「調査型」検索質問の例

Web 検索の精度を評価する典型的な尺度として，MAP（Mean Average Precision）と MRR（Mean Reciprocal Rank）がある．MAP は，課題ごとに精度（precision）と再現率（recall）を考慮して AP（Average Precision）を計算し，全課題の AP を平均した値である．Web 以外の情報を対象とした検索の評価にも用いられる．MRR は，課題ごとに正解が最初に見つかった順位の逆数（Reciprocal Rank: RR）

課題番号	検索質問
0007	マスタードシード, 代理店
0010	SHARP, 液晶テレビ
0012	JPNIC
0014	ギガバイト, マザーボード
0015	ゼンリン
0018	Becky, メールソフト
0027	スカイパーフェク TV, 音楽, 番組
0028	みのもんた, クイズ番組, ファイナルアンサー
0029	吉野ケ里遺跡
0030	登呂遺跡

図 6: 「誘導型」検索質問の例

を計算し，全課題の RR を平均した値である．正解が 1 位で見つかった課題の RR は 1 になり，正解が 2 位で見つかった課題の RR は 0.5 まで落ちる．MAP と MRR は，どちらも大きいほど良い結果を表す．

MAP は精度と再現率の両方を考慮しているため，正解を網羅的に検索した場合に高い値になる．それに対して，MRR は正解の網羅性を考慮していない点が異なる．なお，MRR は正解数が少ない質問応答の評価 [16] にも用いられる．以上の議論から，MAP は調査型の検索質問に対する評価に適しており，MRR は誘導型の検索質問に対する評価に適している．NTCIR でも，調査型と誘導型の検索質問によって，MAP と MRR を使い分けている．

MAP と MRR は，どちらも検索結果の上位から一定数の文書を対象に計算される．調査型の検索では正解の網羅性が重要であり，多くの文書が閲覧されることを考慮して，本実験では検索結果の上位 100 件までを評価対象とした．それに対して，誘導型の検索では，正解の網羅性は重要ではないため，検索結果の上位 10 件までを評価対象とした．

以下，5.2 節では，アンカー検索の有効性について評価する．5.3 節では，検索質問分類の有効性について評価する．5.4 節では，検索質問の自動分類で生じた誤りについて分析する．5.5 節では，Web 検索システムを総合的に評価する．

5.2 アンカー検索の評価

アンカー検索の評価では，誘導型の検索質問 168 件だけを用いて検索実験を行い，以下に示す手法の MRR を比較した．

- A1: ページ本文を用いたコンテンツ検索
- A2: アンカーテキストを代理情報として用いたコンテンツ検索 [4]
- A3: 式 (2) のアンカー検索を用い，$P(t|d)$ の計算に Westerveld ら [17] の文書モデルを用いる
- A4: 式 (2) アンカー検索を用い，$P(t|d)$ の計算に式 (4) のアンカーモデルを用いる
- A5: A4 に対して同義語による平滑化を適用する
- A6: 式 (1) を用いて A1 と A5 の結果を統合する

A1 と A2 のコンテンツ検索では，検索モデルとして Okapi BM25 [15] を用いた．A4 と A5 が本研究の提案手法である．A6 では，式 (1) の α を経験的に 0.3 とした．表 1 に各手法の MRR を示す．

表 1: 誘導型の検索質問に対する MRR の比較

A1	A2	A3	A4	A5	A6
.063	.458	.590	.606	.612	.618

表 1 の結果について考察する．まず，A1 とそれ以外の手法を比較すると，A1 の MRR が極端に低いことが分かる．すなわち，過去の研究で明らかになったように，コンテンツ検索は誘導型の検索質問には効果的でないことが分かった．しかし，A2 のようにアンカーテキストを代理情報として用いると MRR が向上する．

A4 と A5 の MRR は，A1〜A3 の MRR よりも高かった．このことから，本研究で提案したアンカー検索のモデルと同義語による平滑化が効果的であることが分かった．A4 と A5 の差異は，課題番号 0064「ザ・プリンストン・レビュー・オブ・ジャパン」に起因する.「プリンストン」,「レビュー」,「ジャパン」の英語表記が検索キーワードとして使用されたことで，RR が 0 から 1 に向上した．

A5 と A6 を比較すると，誘導型の検索質問に対して，アンカー検索だけでなく，コンテンツ検索の結果を一定の割合で考慮することで MRR がさらに向上した．このことから，アンカー検索とコンテンツ検索は相補的な関係にあることが分かった．ただし，α の値を適切に決める必要がある．

5.3 検索質問分類の評価

まず，検索質問の分類精度を評価した．$i(q)$ の閾値を 0.5 として，$i(q)$ が 0.5 以上の場合は q を調査型に分類し，それ以外の場合は q を誘導型に分類した．

比較対象として Kang ら [9] の手法と Lee ら [12] の手法を用いた．Kang らは，検索質問の分類に用いる 4 つの特徴量を提案した．しかし,「検索質問中の索引語がアンカーテキストで使用される頻度」だけを特徴量として使用した．それ以外の特徴量を用いても検索質問の分類精度は 50%未満だった．人手で決定すべきパラメタは，予備実験によって最適化した．各手法の分類精度を表 2 に示す．表 2 より，本手法の分類精度は，既存の分類手法よりも高かった．

表 2: 検索質問の分類精度

Kang	Lee	本手法
75.6%	72.5%	79.3%

次に，検索質問の分類が検索精度に及ぼす影響について評価した．ここでは，全 295 件の検索質問を用いて，以下に示す B1〜B6 の MAP と MRR を比較した．これらの手法は，式 (1) における α の決定方法だけが異なる．

- B1: 検索質問の分類をせずに，常に $\alpha = 0.5$ とする．
- B2: Kang らの手法で検索質問を分類し，調査型の場合は $\alpha = 0.7$ とし，誘導型の場合には $\alpha = 0.3$ とする．
- B3: Lee らの手法で検索質問を分類し，B2 と同じ方法で α を決定する．
- B4: 本手法で検索質問を分類し，B2 と同じ方法で α を決定する．
- B5: 本手法で検索質問を分類し，$i(q)$ によって α を自動的に決定する．
- B6: 検索質問の正しい分類を使用し，B2 と同じ方法で α を決定する．

B2，B3，B4，B6 における α の値は，予備実験によって検索精度が最も高くなるように設定した結果である．

B1 と B6 の MAP や MRR は，それぞれ期待される検索精度の下限と上限である．本研究の提案手法は B5 である．B4 は B5 の変形であり，α の決定方法を揃えることで B2 や B3 との直接的な比較を行うために使用する．

実験結果を表 3 に示す．MAP と MRR のどちらに対しても，B5 は B2 や B3 の結果を上回った．B5 で α の値を自動的に決定すると，B4 に比べて MAP は若干向上し，逆に MRR は若干低下した．すなわち，検索精度をほとんど保持したまま，α を人手で決定する負担を軽減することができた．

表 3: 検索精度の比較

	B1	B2	B3	B4	B5	B6
MAP	.254	.281	.265	.300	.304	.312
MRR	.468	.504	.485	.519	.517	.545

適合文書が 1 位で検索された質問の件数を調査したところ，B2 は 119 件，B3 は 113 件だったのに対して，B4 と B5 では 127 件に増加した．

本手法は，検索質問の分類が常に正しい B6 に比べると MAP と MRR がともに低下するものの，検索質問の分類を行わない B1 に比べて MRR を向上させた．すなわち，検索質問を自動分類することが Web 検索の精度向上に貢献することが分かった．また，本手法は，既存の分類手法よりも Web 検索の精度を向上させることが分かった．

5.4 検索質問分類の誤り分析

5.3 節の B5 が分類を誤った検索質問を分析し，原因を特定した．また，検索質問の分類誤りによって，B6 と比較して B5 の AP と RR がどのように変化したのかを分析した．分析の結果を表 4 に示す．表 4 において，(i) と (ii) は調査型の検索質問に対する誤りの原因であり，(iii)〜(vi) は誘導型の検索質問に対する誤りの原因である.「↓」,「=」,「↑」は，それぞれ，AP や RR の「低下」,「等価」,「向上」を示す. 表 4 より，分類誤りのために総じて AP と RR が低下したことが分かる．

しかし，分類誤りによって RR が向上した質問もあった．例えば,「京都，寺，神社」という調査型の検索質問である．この質問に対してアンカー検索の

表 4: 検索質問分類の誤り原因と AP/RR の変化

誤りの原因	質問の種類	誤りの件数	AP			RR		
			↓	=	↑	↓	=	↑
(i)	調査型	14	14	0	0	10	3	1
(ii)	調査型	9	9	0	0	5	3	1
(iii)	誘導型	27	8	9	10	6	16	5
(iv)	誘導型	1	1	0	0	1	0	0
(v)	誘導型	4	0	4	0	0	4	0
(vi)	誘導型	1	1	0	0	1	0	0

検索結果が重視され，その結果，京都の観光に関する代表的なページが検索された．これは，ユーザは調査型を意図して質問を入力したにも拘らず，誘導型の質問として処理したことが良い結果につながった例である．

(i) は，検索質問が特徴的な索引語に分割されたために，個々の索引語に関する ALD の歪度が大きくなったことが原因である．(ii) は，分割された索引語全てを含むアンカーテキストが存在したために，ALD の歪度が大きくなったことが原因である．

(iii) は，(i) の逆であり，検索質問が一般的な索引語に分割されたために，個々の索引語に関する ALD の歪度が小さくなったことが原因である．(iv) は,「遺伝子組み換え食品」という検索質問が「遺伝子組み換え食品ホームページ」や「本当はどうなの？ 遺伝子組み換え食品」のように様々な文脈でアンカーテキストに使用されていたために，ALD の歪度が小さくなったことが原因である．(v) は，検索質問も分割後の索引語もアンカーテキストに存在しないことが原因だった．(vi) は，他の誤り事例に比べると ALD の歪度は大きいものの，$i(q)$ が 0.5 という閾値に比べて若干大きかったことが原因である．

(i)〜(iii) は本手法で検索質問を分割したことに起因する．(iv)〜(vi) は検索質問の分類にアンカーテキストを利用することに起因する根本的な問題である．

5.5 Web 検索システム全体の評価

本研究で開発した Web 検索システムを総合的に評価した．図 1 のシステムは，検索質問分類，コンテンツ検索，アンカー検索で構成されている．このうち，コンテンツ検索は既存の手法を用いている．そこで，検索質問分類とアンカー検索を既存の手法で実装することで,「既存の Web 検索システム」を構成し，比較対象とした．表 3 より，最も精度が高かった既存の検索質問分類手法は B2 である．表 1 より，最も精度が高かった既存のアンカー検索手法は A3 である．そこで，A3 と B2 をそれぞれアンカー検索と検索質問分類に使用したシステムを B0 として，表 3 の B5，B6 と比較した．比較実験の結果を表 5 に示す．表 5 より，B5 は MAP と MRR の両方において B0 の結果を上回った．

表 5: 検索システム全体の評価

	B0	B5	B6
MAP	.272	.304	.312
MRR	.491	.517	.545

さらに，表 5 における B0 と B5 の差が偶然ではないことを確認するために検定を行った．具体的には，295 件の検索質問を無作為標本と見なして，両側 t 検定を行った [8, 10]．検定の結果を表 6 に示す．表 6 において，$<$ と $\ll$ は，検定対象の差がそれぞれ有意水準 0.05 と 0.01 で有意であったことを示す．表 6 より，B0 と B5 の MAP や MRR における差は有意であった．B5 と B6 を比較すると，MAP では有意な差がなかった．この結果から，本システムは既存のシステムよりも有用性が高く，また，検索質問の分類が常に正しいシステムに匹敵することが分かった．

表 6: t 検定の結果

	MAP	MRR
B0 vs. B5	$\ll$	$<$
B5 vs. B6	—	$\ll$

6 おわりに

Web 検索では，検索質問の種類によって必要とされる検索手法が異なる．本研究は，検索質問を「調査型」と「誘導型」に自動分類し，その結果に基づいて検索手法を動的に変更する手法を提案した．具体的には，誘導型の検索質問に使われる言葉は，アンカーテキストに出現しやすいという性質を利用する．そこで，検索質問中のキーワードが Web 上のアンカーテキストにどのように分布するかを分析することで，検索質問の種類を特定した．さらに，誘導

型の質問に対する検索精度を向上させるためにアンカーテキストをモデル化し，検索に利用する手法を提案した．ここでは，アンカーテキストから同義語を抽出して検索に利用した．NTCIR で構築された 100GB の Web 検索用テストコレクションを用いた実験の結果，本研究で提案したアンカー検索と検索質問分類の手法が有効であることが確認された．

謝辞

本研究の一部は，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」（課題番号 19024007）によって実施された．

参考文献

[1] Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Cristina González-Caro. The intention behind Web queries. In *Proceedings of the 13th International Conference on String Processing and Information Retrieval*, pp. 98–109, 2006.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, Vol. 30, No. 1–7, pp. 107–117, 1998.

[3] Andrei Broder. A taxonomy of web search. *ACM SIGIR Forum*, Vol. 36, No. 2, pp. 3–10, 2002.

[4] Nick Craswell, David Hawking, and Stephen Robertson. Effective site finding using link anchor information. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 250–257, 2001.

[5] Koji Eguchi, Keizo Oyama, Akiko Aizawa, and Haruko Ishikawa. Overview of the WEB task at the fourth NTCIR workshop. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004.

[6] Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama. Overview of the Web retrieval task at the third NTCIR workshop. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[7] Atsushi Fujii and Tetsuya Ishikawa. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, Vol. 35, No. 4, pp. 389–420, 2001.

[8] David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329–338, 1993.

[9] In-Ho Kang and GilChang Kim. Query type classification for Web document retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 64–71, 2003.

[10] E. Michael Keen. Presenting results of experimental retrieval comparisons. *Information Processing & Management*, Vol. 28, No. 4, pp. 491–502, 1992.

[11] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pp. 668–677, 1998.

[12] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in Web search. In *Proceedings of the 14th International World Wide Web Conference*, pp. 391–400, 2005.

[13] Yunyao Li, Rajasekar Krishnamurthy, Shivakumar Vaithyanathan, and H. V. Jagadish. Getting work done on the Web: Supporting transactional queries. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 557–564, 2006.

[14] Keizo Oyama, Koji Eguchi, Haruko Ishikawa, and Akiko Aizawa. Overview of the NTCIR-4 WEB navigational retrieval task 1. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004.

[15] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, 1994.

[16] Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 200–207, 2000.

[17] Thijs Westerveld, Wessel Kraaij, and Djoerd Hiemstra. Retrieving Web pages using content, links, URLs and anchors. In *Proceedings of the 10th Text REtrieval Conference*, pp. 663–672, 2001.

[18] Kiduk Yang. Combining text- and link-based retrieval methods for Web IR. In *Proceedings of the 10th Text REtrieval Conference*, pp. 609–618, 2001.