

アンカーテキストモデルと検索質問分類による Web 文書検索の高度化

藤 井 敦^{†1}

World Wide 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モデルが異なる。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ものが目的）」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モデル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検索モデルが有効である。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検索モデルが有効である。そこで、本研究はアンカーテキストを用いた検索モデルも提案する。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、個別の提案手法と Web 検索システム全体の有効性を示す。

Enhancing Web Document Retrieval by the Anchor Text Model and Query Classification

ATSUSHI FUJII^{†1}

Several types of queries are widely used on the World Wide Web and the expected retrieval model can vary depending on the query type. We propose a method for classifying queries into informational and navigational types. While content-based retrieval is effective for informational queries, anchor-based retrieval is effective for navigational queries. Our retrieval system combines the results obtained with the content-based and anchor-based retrieval models, in which the weight for each retrieval result is determined automatically depending on the result of the query classification. We also propose a retrieval model using anchor texts. We use the NTCIR test collections and show the effectiveness of individual methods and the entire Web retrieval system experimentally.

1. はじめに

World Wide Web には多種多様な情報が存在し、Web を検索するユーザの情報要求もまた多種多様である。Broder³⁾ 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。

Broder³⁾ 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、一部に推定を含むものの、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。具体的には、調査型、誘導型、取引型の割合は、およそ5:2:3であった。

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モデルが有効であり、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モデルが有効であるという知見が得られている^{5)-7),11),18)}。

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うえで重要である。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、

^{†1} 東京工業大学

Tokyo Institute of Technology

^{*1} <http://trec.nist.gov/> (2010 年 9 月参照)

^{*2} <http://research.nii.ac.jp/ntcir/index-ja.html> (2010 年 9 月参照)

アンカーテキストとは異なる。

Li ら¹⁵⁾ 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、既存のモデルとは異なる検索モデルが有効であった。

しかし、ユーザは検索質問を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、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。

本論文は、検索質問を自動的に分類し、その結果に基づいて検索モデルを動的に変更する手法を提案する。さらに、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モデルは研究の歴史が浅いために改善の余地がある。さらに、提案手法を統合した Web 検索システムを提案する。なお、本論文は調査型と誘導型の検索質問を対象とする。

2 章では、本論文で提案する Web 検索システムについて説明する。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 章で提案手法とシステムの評価実験について説明する。

2. Web 検索システムの概要

本論文で提案する 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、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、検索質問の種類によらずに、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 以外の検索でも用いられている Okapi BM25²⁰⁾ を用いる。ただし、原理的には Okapi BM25 以外の検索モデルを用いてもよい。コンテンツ検索の詳細は本章の最後で説明する。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。

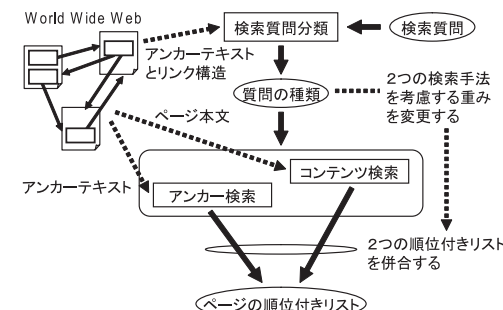


図 1 Web 検索システムの概要

Fig. 1 Overall design of our Web retrieval system.

最後に、2 つの初期リストを併合して出力する。ここで、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 $S(d)$ に基づいて降順に整列される。

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同じであるとは限らないため、単純に加重平均をとることはできない。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。

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

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 以下の値をとり、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、本論文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、図 1 のシステムでは α の値を手で決定する必要がある。

情報検索結果の併合に関する関連研究では、ある検索質問に対して複数の文書コレクションを同時に検索し、複数の文書リストを 1 つに併合する手法²¹⁾ が提案されている。それに対して、本手法は単一の文書コレクションに対して複数の検索モデルを使用して複数の文書

リストを作成し、検索質問の種類に応じて検索モデルの重要度を考慮しながら文書リストを併合する点が異なる。

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。コンテンツ検索のモデルは本論文の提案手法ではないので、本章の残りで説明する。

コンテンツ検索の索引付けでは、ページの本文から HTML タグを削除し、ChaSen^{*1}を用いて形態素解析を行う。形態素解析の結果に基づいて、名詞、動詞、形容詞、辞書未登録語、記号を抽出し、単語と単語バイグラムを索引語として利用する。なお、単語と単語バイグラムは独立の索引語として扱う。単語と単語バイグラムを併用する手法の有効性は過去の情報検索実験¹⁰⁾でも示されている。Okapi BM25 を用いて、検索質問 q に対するページ d のスコアを計算する。具体的には、式 (2) を用いる。

$$\sum_{t \in q} \frac{(k_1 + 1) \cdot f_{t,d}}{k_1 \cdot \left\{ (1 - b) + b \cdot \frac{dl_d}{avgdl} \right\} + f_{t,d}} \cdot \log \frac{N}{n_t} \cdot \frac{(k_3 + 1) \cdot f_{t,q}}{k_3 + f_{t,q}} \quad (2)$$

$f_{t,d}$ と $f_{t,q}$ はそれぞれ d と q における索引語 t の出現頻度である。 dl_d は d の文書長 (バイト) であり、 $avgdl$ は平均文書長である。 N は検索対象文書の総数であり、 n_t は t が出現する文書数である。 k_1 , b , k_3 はパラメータであり、本論文の実験では、過去の文献でよく使われる値に従って、 $k_1 = 2$, $b = 0.75$, $k_3 = 1,000$ とした¹⁶⁾。ただし、Okapi BM25 本来の IDF (逆文書頻度) は $\log \frac{N - n_t + 0.5}{n_t + 0.5}$ で計算され、 n_t が $\frac{N}{2}$ よりも大きい場合はスコアが負の値となる。そこで、式 (2) では通常の IDF を用いている。

さらに、擬似適合性フィードバック (pseudo-relevance feedback)¹⁶⁾ を行う。まず初期検索として式 (2) を用いてページに順位を付ける。次に、上位のページから重みが大きい索引語を検索質問に追加して再検索を行う。再検索においても式 (2) を用いてページのスコアを計算する。しかし、検索質問の索引語が変化しているため、初期検索とはページの順位付けが異なる。具体的には、初期検索の上位 10 ページから重みが大きい索引語の上位 10 件を選択する。上記 2 カ所の「上位 10」という値は、種々のテストコレクションを用いた実験から経験的に決定した。索引語の重みは、式 (2) における TF.IDF に相当する要因であり、式 (3) で計算する。

$$\frac{(k_1 + 1) \cdot f_{t,d}}{k_1 \cdot \left\{ (1 - b) + b \cdot \frac{dl_d}{avgdl} \right\} + f_{t,d}} \cdot \log \frac{N}{n_t} \quad (3)$$

*1 <http://chasen-legacy.sourceforge.jp/> (2010 年 9 月参照)

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。

Yang²⁴⁾ 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、HITS¹³⁾ 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⁵⁾ 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²³⁾ 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 らの手法は本論文の提案手法に最も類似するため、両手法の相違点について 3.3 節で説明する。

3.2 検索モデルの全体像

本論文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。以下、アンカーテキストにおける索引語の頻度分布を用いて $P(d|q)$ を推定する。まず、ベイズの定理を用いて $P(d|q)$ を式 (4) のように変形する。

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \propto P(q|d) \cdot P(d) \quad (4)$$

式 (4) 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。本論文では、 $P(d)$ の計算方法として 2 通りの選択肢を検討する。まず、「 d へのリンク数」と「Web コレクション全体におけるリンク総数」の割合、すなわち最尤推定によって $P(d)$ を計算する。この方法では、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とる。 $P(d)$ を計算する別の方法として、PageRank²⁾ を使うこともできる。PageRank は、ユーザがリンクをたどりながらページ d に到達する確率を計算するからである。5.2 節では、アンカー検索モデルにおける最尤推定と PageRank の有効性を実験によって比較する。

検索質問 q は複数の索引語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 (5) によって $P(q|d)$ を計算する。

$$P(q|d) = \prod_{t \in q} P(t|d) \quad (5)$$

$P(t|d)$ は、ページ d にリンクしている 1 つ以上のアンカーテキストから無作為に 1 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 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。本論文では、 $P(t|d)$ の計算方法として、「文書モデル」と「アンカーテキストモデル」という 2 通りの選択肢を検討する。詳細は 3.3 節で説明する。

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 (5) 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。具体的には式 (6) を用い、索引語 t の同義語 s によって $P(t|d)$ を近似する。

$$\begin{aligned} P(t|d) &= P(t|s, d) \cdot P(s|d) \\ &\approx P(t|s) \cdot P(s|d) \end{aligned} \quad (6)$$

式 (6) において、 $P(t|s)$ は s が t に置換される確率である。式 (6) の 2 行目を導出するために、 s が t に置換される確率が d と独立であると仮定している。 $P(s|d)$ は式 (5) における $P(t|d)$ と本質的に同じであり、その計算方法は 3.3 節で説明する。同義語の抽出と $P(t|s)$ の計算に関する手法は 3.4 節で説明する。

アンカー検索モデルが有効に機能する誘導型の検索質問では再現率よりも精度が重視されるため、過度の平滑化は好ましくない。そこで、 $P(t|d)$ がゼロになる場合のみ式 (6) を使用する。同じ理由により、高い精度で抽出しやすい同義語対として、カタカナによって翻字 (音訳) された用語と元の用語に限定する。

t の同義語がない場合は、 $P(t|d)$ の代わりに $P(t)$ を用いて平滑化する。 $P(t)$ は、検索対象のページ集合に含まれる全アンカーテキストから無作為に 1 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。そこで、 $P(t)$ を用いた場合は、全アンカーテキストにおける t の出現頻度が低いほど $P(q|d)$ が小さくなる。IDF の趣旨からも分かるように、高頻度語よりも低頻度語の方が一般的に検索には有効である。すなわち、検索質問とアンカーテキストの間で索引語の不一致が生じた場合は、それが検索に有効な低頻度語であるほど、 $P(q|d)$ に与える影響が大きくなる。

最後に、スパムへの対策について考察する。ここでいう「スパム」とは、迷惑メールのような事物ではなく、第三者が検索結果を意図的に操作する行為を指す。操作の意図や手法は

多岐にわたるものの、典型的な例は、特定のページが上位で検索されるように当該ページへのリンクを増やす方法である。個人や小規模な団体が行う場合には、自分たちが管理するページ間で相互にリンクするだろう。こうしたスパムへの対策として、本手法では同一の Web サーバ内に閉じたリンクとそれに対応したアンカーテキストは抽出しない。また、あるページから同一のページに対して複数のリンクがあった場合は、最初のリンクとそれに対応したアンカーテキストだけを抽出する。しかし、大規模で組織的なスパムへの対策はそれ自身が 1 つの研究課題であるため⁴⁾、本論文では対象とせずに今後の検討課題とする。

3.3 アンカーテキストモデル

式 (5) より、アンカー検索モデルでは $P(t|d)$ の計算が重要である。Westerveld ら²³⁾ は、ページ d にリンクしている 1 つ以上のアンカーテキストをまとめて 1 つの「代理文書」として扱い、 $P(t|d)$ を計算した。以降、この手法を「文書モデル」と呼ぶ。文書モデルにおける $P(t|d)$ は、ページ d の代理文書から無作為に索引語を 1 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、文書モデルには問題がある。たとえば、Yahoo! Japan のトップページ^{*1}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 Yahoo! Japan のトップページを検索するために等しく重要であると見なされる。しかし、実際には「ヤフー」は単体で Yahoo! Japan 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いはずである。

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、ある特定の個人が $P(t|d)$ の分布を容易に操作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、本論文では、索引語の出現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モデルを提案する。以降、本モデルを「アンカーテキストモデル」と呼ぶ。なお、「アンカー検索モデル」は検索モデルであり、ここで説明する「アンカーテキストモデル」とは、アンカー検索におけるアンカーテキストのモデル化手法である。具体的には、 $P(t|d)$ を式 (7) のように分解する。

*1 <http://www.yahoo.co.jp/> (2010 年 9 月参照)

$$P(t|d) = \sum_{a \in \mathbf{A}_d} P(t|a) \cdot P(a|d) \quad (7)$$

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 (7) の右辺では、索引語の出現確率は $P(t|a)$ で計算される。ここで、式 (7) に対する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。すなわち、式 (7) の $P(t|d)$ 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 Yahoo! Japan 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、現実合っている。

また、文書モデルに比べて、アンカーテキストモデルはスパムに対して頑健である。 $P(t|d)$ の分布を操作するためには、 $P(a|d)$ が大きい a に対して $P(t|a)$ の分布を操作することが効果的である。しかし、 $P(a|d)$ が大きい a とは、多数のユーザが d へのリンクに使用するアンカーテキストである。すなわち、特定の個人が $P(t|a)$ の分布を操作することは困難である。他方において、 d へのリンクにはめったに使用されないアンカーテキスト a を独自に作れば、 $P(t|a)$ の分布を操作することは容易である。しかし、その場合は $P(a|d)$ が小さくなるので、 $P(t|d)$ 全体に与える影響は小さくなってしまう。以上より、アンカーテキストモデルでは、個人や小規模な団体が $P(t|d)$ の分布を操作することが困難である。ただし、組織的なスパムへの対策は今後の検討課題である。

3.4 同義語の抽出

索引語 t の同義語 s を抽出するために、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。

たとえば、Google の日本語トップページ^{*1}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された用語と元の用語を同義語の対として抽出する。具体的には、統計的な翻字手法⁸⁾を利用し、与えられた用語対の一方から

他方に翻字することができるかどうかを検査する。もし翻字できれば、両者を同義語対として検出する。上述の例では、「グーグル」と「google」が同義語対として検出される。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 (8) によって $P(t|s)$ を計算する。

$$P(t|s) = \frac{F(t, s)}{\sum_{r \neq s} F(r, s)} \quad (8)$$

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。

Lu ら¹⁷⁾ は、アンカーテキストとリンク構造を利用して対訳を抽出し、言語横断情報検索に応用した。しかし、本論文の提案手法は、対象を発音が類似した翻字に限定することで抽出精度を高める工夫をしている点と、抽出した翻字をアンカー検索モデルの平滑化に応用している点が異なる。

4. 検索質問分類

4.1 先行研究

Kang ら¹¹⁾ は、種々の特徴量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、評価実験に用いられた TREC のテストコレクションは 10 GB 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¹⁴⁾ 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、Lee らは検索実験を行っていないため、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、本論文は、NTCIR-3 と NTCIR-4 で構築された 100 GB のテストコレクションを対象として、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、本論文では、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。

本論文で提案する手法は、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

*1 <http://www.google.co.jp/> (2010 年 9 月参照)

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論文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。たとえば、情報処理学会のトップページ^{*1}にリンクする場合は、「情報処理学会」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、情報処理学会と無関係なページにリンクする場合は、「情報処理学会」がアンカーテキストとして使われることは稀である。

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。たとえば、「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¹⁴⁾ 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。

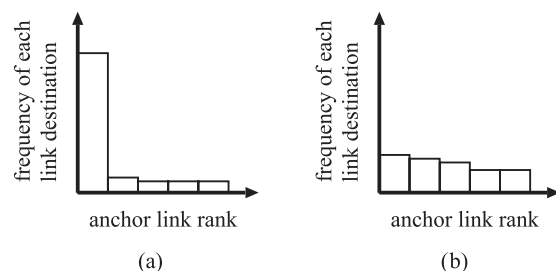


図 2 anchor-link distribution の例
Fig. 2 Examples for anchor-link distribution.

*1 <http://www.ipsj.or.jp/> (2010 年 9 月参照)

本論文では、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、本論文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 はアンカーテキストの一部または全体である。 $\mathbf{D}_t$ 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 (9) によってエントロピーを計算する。

$$H(\mathbf{D}_t|\mathbf{T}_q) = - \sum_{t \in \mathbf{T}_q} P(t) \sum_{d \in \mathbf{D}_t} P(d|t) \log P(d|t) \quad (9)$$

各 t 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。

$P(t)$ は、 q における t の出現確率である。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 の出現確率は一様であると考え、 $P(t) = \frac{1}{|\mathbf{T}_q|}$ とする。 $P(d|t)$ は、 t が出現するアンカーテキストによって d がリンクされている確率であり、図 2 における縦軸の値である。ただし、Lee らの手法が q と t を同一視していることに対して、本手法では q を t に分割し、各 t に対する $P(d|t)$ を統合する点が異なる。

さらに、 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 s 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、3.4 節で提案した手法を使用する。ただし、単に式 (9) 中の t を s に置換するだけであり、式 (8) の $P(t|s)$ を式 (9) に組み込むことはしない。

最後に、式 (10) のように $H(\mathbf{D}_t|\mathbf{T}_q)$ 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。

$$i(q) = \frac{H(\mathbf{D}_t|\mathbf{T}_q)}{\log |\mathbf{D}_t|} \quad (10)$$

$i(q)$ を式 (1) の α とすることで、 $i(q)$ の値が 0.5 より大きい 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。質問の種類を具体的に決めるか否かという点について、5.3 節で比較実験を行う。

以上をまとめると、本論文は、先行研究では未解決であった 3 つの課題を解決した。まず、検索質問と完全一致するアンカーテキストが存在しない場合でも ALD の歪度を計算可

能とした。次に、コンテンツ検索とアンカー検索の結果を統合する際の重みを自動的に決定可能とした。さらに、検索質問の分類において、同義語による平滑化を可能とした。

ただし、本手法には 2 つのパラメータがある。1 つ目は、図 2 のヒストグラムにおける階級幅である。 $P(d|t)$ 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。そこで、現在は 5 としている。2 つ目は、 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かかわらず、すべての $d \in \mathbf{D}_t$ 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 K 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 $K = 10,000$ としている。これら 2 つのパラメータは、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。なお、これらの 2 つのパラメータは、Lee らの手法でも適宜設定する必要があるため、この点において本論文の提案手法が劣るわけではない。

5. 評価実験

5.1 実験方法

本論文で提案した内容の有効性を評価するために 3 通りの実験を行った。具体的には、3 章で提案したアンカー検索モデル、4 章で提案した検索質問分類、2 章で提案した Web 検索システムを評価した。

表 1 に本実験で使用したテストコレクションの概要を示す。まず、NTCIR-3⁷⁾ と NTCIR-4^{6),18)} で構築された Web 検索のテストコレクションを使用した。NTCIR-3 と NTCIR-4 の検索対象は共通であり、JP ドメインから収集された約 100 GB (約 1 千万ページ) のペー

ジ集合である。さらに、NTCIR-5 で構築された Web 検索のテストコレクション¹⁹⁾ も使用した。NTCIR-5 の検索対象は JP ドメインから収集された約 1 TB (約 1 億ページ) のページ集合である。いずれのページ集合もほとんどが日本語のページである。

検索課題も日本語で記述されている。NTCIR-3 では調査型の検索課題だけが作られ、NTCIR-4 では調査型と誘導型の検索課題が作られた。NTCIR-5 では誘導型の検索課題だけが作られた。NTCIR-3 と NTCIR-4 の検索対象が共通であることから、当該テストコレクションを用いて検索質問分類の有効性を評価した。さらに、Web 検索システム全体の有効性も評価した。他方において、NTCIR-4 のうち誘導型の検索課題と NTCIR-5 の検索課題を個別に用いてアンカー検索モデルの有効性を評価した。

各検索課題に対する適合判定のレベルには、「高適合」、「適合」、「部分適合」、「不適合」がある。本実験では、「高適合」もしくは「適合」と判定されたページだけを「適合文書」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、NTCIR-5 の誘導型が 841 件である。表 1 には、「課題あたりの適合文書数」、「検索質問あたりの索引語数」、「本論文中で使用する節」の内訳も示している。「課題あたりの適合文書数」から、誘導型よりも調査型の検索課題に対する適合文書が多いことが分かる。誘導型の検索課題では、対象の事柄に対する代表的なページが多く存在しないためである。「検索質問あたりの索引語数」から、誘導型よりも調査型の検索質問が多く索引語で構成されていることが分かる。

ここで、「検索課題 (search topic)」と「検索質問 (query)」は同義ではない。以下、両者の違いについて具体例を交えながら説明する。なお、「検索課題」を「課題」と略記する場合がある。**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。

図 3 や図 4 に示した課題からどの項目を抽出して、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、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。**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

表 1 本論文の実験で使用したテストコレクション
Table 1 Test collections used for experiments.

	NTCIR-3	NTCIR-4		NTCIR-5
検索課題の種類 (検索課題数)	調査型 (47)	調査型 (80)	誘導型 (168)	誘導型 (841)
文書集合のファイル容量 (文書総数)	約 100 GB (約 1 千万ページ)			約 1 TB (約 1 億ページ)
検索課題あたりの適合文書数	75.7	84.5	1.79	1.94
検索質問あたりの索引語数	2.89	2.39	1.39	1.35
本論文中で使用する節	5.3, 5.4, 5.5			5.2

```

<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

Fig. 3 Example for NTCIR-3 informational search topic.

```

<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

Fig. 4 Example for NTCIR-4 navigational search topic.

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。

本節の最後では, 評価尺度について説明する。Web 検索の有効性を評価する代表的な尺度として, 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。MAP は Web 以外の情報を対象とした情報検索の評価にも用いられる。MRR は, 課題ごとに適合文書が最初に見つかった順位の逆数 (reciprocal rank: RR) を計算し, 全課題の RR を平均した値である。適合文書が 1 位で見つかった課題の RR は 1 になり, 適合文書が 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, 適合文書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 は適合文書の網羅性を考慮していない点異なる。なお, MRR は正解数が少ない質問応答の評価²²⁾にも用いられる。以上の議論から, MAP は調

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

Fig. 5 Examples for informational queries.

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

Fig. 6 Examples for navigational queries.

査型の課題に対する評価に適しており, MRR は誘導型の課題に対する評価に適している。NTCIR の評価においても調査型と誘導型の課題によって 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。調査型の検索では適合文書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, 誘導型の検索では適合文書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 節でアンカー検索の有効性について評価する。5.3 節で検索質問分類の有効性について評価し, 5.4 節で検索質問分類の誤りについて考察する。5.5 節で 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、誘導型の課題 1,009 件 (NTCIR-4: 168 件, NTCIR-5: 841 件) を用いて検索実験を行い、複数の手法を MRR で比較した。混乱を避けるために、まずは本実験において明らかにすべき点を整理する。

- 誘導型の課題に対して、コンテンツ検索モデルの Okapi BM25 (2 章参照) とアンカー検索モデル (3 章参照) のどちらが有効か？ なお、コンテンツ検索モデルは、ページ本文を索引付けする手法とアンカーテキストを代理文書として索引付けする手法⁵⁾に細分することができる。ここで、後者は代理文書を使用するものの、「文書モデル」(3.3 節参照) とは異なる点に注意を要する。
- アンカー検索モデルにおける $P(t|d)$ の計算には、文書モデル²³⁾ と本論文で提案するアンカーテキストモデル (3.3 節参照) のどちらが有効か？
- アンカー検索モデルにおける $P(d)$ の計算には、PageRank と最尤推定 (3.2 節参照) のどちらが有効か？
- 同義語による平滑化 (3.4 節参照) は有効か？
- コンテンツ検索とアンカー検索モデルの統合 (2 章参照) は有効か？

以上の観点を念頭に置いて、本実験で比較した手法 A1~A8 を挙げる。

- A1: ページ本文を索引付けに用いるコンテンツ検索モデル
- A2: アンカーテキストを代理情報として索引付けに用いるコンテンツ検索モデル
- A3: $P(t|d)$ と $P(d)$ にそれぞれ文書モデルと PageRank を用いるアンカー検索モデル
- A4: $P(t|d)$ と $P(d)$ にそれぞれ文書モデルと最尤推定を用いるアンカー検索モデル
- A5: $P(t|d)$ と $P(d)$ にそれぞれアンカーテキストモデルと PageRank を用いるアンカー検索モデル
- A6: $P(t|d)$ と $P(d)$ にそれぞれアンカーテキストモデルと最尤推定を用いるアンカー検索モデル
- A7: A6 に対して同義語による平滑化を適用するアンカー検索モデル
- A8: 式 (1) を用いて A1 と A7 の結果を統合する検索モデル

本論文で提案する手法の変形は、A5~A8 である。A8 では、式 (1) の α を設定する必要がある。 α の値は予備実験によって最適化した。具体的には、NTCIR-4 では $\alpha = 0.3$ とし、NTCIR-5 では $\alpha = 0.1$ とした。

表 2 に各手法の MRR を示す。表 2 の結果について考察する。まず、手法間の優劣は NTCIR-4 と NTCIR-5 で大差なかった。

表 2 誘導型の検索質問に対する MRR の比較

Table 2 Comparison of MRR for navigational queries.

	A1	A2	A3	A4	A5	A6	A7	A8
NTCIR-4	0.063	0.458	0.556	0.590	0.567	0.606	0.612	0.618
NTCIR-5	0.047	0.446	0.556	0.675	0.577	0.691	0.691	0.691

A1 と A2 を比較すると、同じコンテンツ検索モデルであっても誘導型の検索質問にはアンカーテキストを索引付けする手法が有効であることが分かった。

A1 とアンカー検索モデルの変形 (A3~A8) を比較すると、A1 の MRR が極端に低いことが分かった。すなわち、過去の研究と同じように、誘導型の検索質問に対しては、コンテンツ検索モデルよりもアンカー検索モデルの方が有効であるという結果が本実験でも再現された。

A3 と A4 (もしくは A5 と A6) は $P(d)$ の計算方法だけが異なる。表 2 の結果から、 $P(d)$ の計算には PageRank よりも最尤推定の方が有効であることが分かった。

A3 と A5 (もしくは A4 と A6) は $P(t|d)$ の計算方法だけが異なる。表 2 の結果から、 $P(t|d)$ の計算には Westerveld ら²³⁾ の文書モデルよりも、本論文で提案したアンカーテキストモデルの方が有効であることが分かった。

A6 と A7 は同義語による平滑化の有無だけが異なる。表 2 の結果から、NTCIR-4 の検索質問に対して、同義語による平滑化が有効であることが分かった。この差異は、NTCIR-4 の課題番号 0064 「ザ・プリンストン・レビュー・オブ・ジャパン」に起因する。A6 では当該課題の RR が 0 であったことに対して、A7 では「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。

A7 と A8 を比較すると、NTCIR-4 の課題に対しては、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。

NTCIR-4 と NTCIR-5 の結果を比較すると、A4~A8 では NTCIR-5 に対する MRR が高かった。しかし、A5 では NTCIR-4 と NTCIR-5 の MRR に大きな違いはなかった。NTCIR-5 の MRR が顕著に高いグループ (A4, A6, A7, A8) とそれ以外のグループ (A1, A2, A3, A5) に分けると、両グループの違いは、前者が $P(d)$ の計算に最尤推定を使用している点にある。NTCIR-5 の文書集合は NTCIR-4 よりも 10 倍大きいので、リンク情報も密である。その結果、NTCIR-5 では、 $P(d)$ の計算にリンク数を直接的に使用する最尤

推定の効果が顕著であったと考えられる。

以上の結果をまとめると、誘導型の検索質問に対して、a) アンカーテキストモデル、b) 同義語による平滑化、c) アンカー検索モデルとコンテンツ検索モデルの統合が個別に有効であることが分かった。本論文の提案手法は上記の a) と b) であるため、本論文の成果によって、NTCIR-4 の MRR は 0.590 (A4) から 0.606 (A6) もしくは 0.612 (A7) まで向上し、NTCIR-5 の MRR は 0.675 (A4) から 0.691 (A6 または A7) まで向上した。

これらの差が必然か偶然かを検査するために、検索課題を無作為標本と見なして両側 t 検定を行った^{9),12)}。その結果、A4 と A6 の差は NTCIR-4 と NTCIR-5 においてそれぞれ有意水準 0.05 と 0.01 で統計的に有意であった。すなわち、誘導型の検索において、本論文で提案したアンカーテキストモデルは既存の文書モデルよりも有効であった。しかし、A6 と A7 の差は NTCIR-4 と NTCIR-5 のいずれにおいても有意ではなかった。

5.3 検索質問分類の評価

検索質問分類の評価では、NTCIR-3 の調査型 127 件と NTCIR-4 の誘導型 168 件を用いて、「分類」と「検索」の実験を行った。まず、検索質問の分類精度を評価した。本手法では、式 (10) で計算される $i(q)$ の閾値を 0.5 として、 $i(q)$ が 0.5 以上の場合は検索質問 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¹¹⁾ の手法と Lee ら¹⁴⁾ の手法を用いた。Kang らは、検索質問の分類に用いるために、検索質問中の語句に関する 4 つの特徴量を提案した。4 つの特徴量に対する重みは人手で最適化する必要がある。しかし、本実験では Kang らが提案した特徴量のうち「検索質問中の語句がアンカーテキストで使用される頻度の割合 (usage rate as an anchor text)」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最高で 53.9% だった。

比較対象の手法はどれも分類に関するスコアを計算するだけなので、「調査型」か「誘導型」に分類するための閾値が必要である。上述したように、本論文の提案手法では閾値を 0.5 とした。比較対象の手法では複数の閾値を試し、分類精度が最も高くなる閾値を選択した。各手法の分類精度を表 3 に示す。表 3 より、本手法の分類精度は、既存の分類手法よりも高いことが分かった。

表 3 検索質問の分類精度

Table 3 Accuracy for query classification.

Kang	Lee	本手法
75.6%	72.5%	79.3%

アンカー検索と同様に、検索質問の分類においても同義語による平滑化を行う。当該平滑化を行わない場合と比較したところ、平滑化によって正しく分類された課題が 2 件あった。どちらも NTCIR-4 の誘導型課題であり、具体的には、課題番号 0010 の「SHARP、液晶テレビ」と課題番号 0078 の「フランス、観光」であった。平滑化によって分類を誤った課題はなかった。

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。いずれの手法も 5.2 節の A8 を用い、検索質問の分類手法と式 (1) における α の値を決定する方法だけが異なる。

- B1: 検索質問の分類をせずに、つねに $\alpha = 0.5$ 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最適化した結果である。それに対して、B5 は α の値を自動的に決定する。

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。本論文の提案手法は 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 4 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。ただし、B4 と B5 の MAP や MRR には統計的に有意な差がなかった。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 4 MAP と MRR の比較

Table 4 Comparison of MAP and MRR.

	B1	B2	B3	B4	B5	B6
MAP	0.254	0.281	0.265	0.300	0.304	0.312
MRR	0.468	0.504	0.485	0.519	0.517	0.545

適合文書が 1 位で検索された課題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、検索質問の分類がつねに正しい 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。

比較対象の手法 B1～B3 では、B2 の結果が最も良かった。両側 t 検定を用いたところ、B2 と本手法 (B4 と B5 のいずれも) の MAP における差は有意水準 0.05 で有意だった。他方において、MRR では有意な差がなかった。しかし、5.5 節で示すように、本論文の提案手法を統合した検索システム全体の評価では、MAP と MRR の両方で既存の手法に対して有意な改善が確認された。

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 5 に示す。表 5 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 5 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。ただし、(iii) では AP と RR のどちらにおいても「低下」と「向上」がほぼ同数あり、「等価」も多かった。

(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。例として、NTCIR-4 の「0001 オフサイド、サッカー、ルール」がある。検索質問の例において、先頭の数字は課題番号を表す。

(ii) は、分割された索引語すべ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。具体的には、NTCIR-4 の「0030 フォトショップ、tips」がある。しかし、(ii) の分類誤りによって RR が向上した質問もあった。たとえば、NTCIR-3 の「0013 京都、寺、神社」である。この質問に対してアンカー検索の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かかわらず、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。

(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。具体例として、NTCIR-4 の「0159 ビーナスライン、観光」がある。「ビーナスライン」は道路の固有名詞であるため ALD の歪度が大きいものに対して、一般語である「観光」に対する ALD の歪度が小さいために、全体として ALD の歪度が小さくなってしまった。

(iv) は、NTCIR-4 の「0092 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、元の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。具体的には、NTCIR-4 の「0160 巖流島」がある。

(vi) は、分類の閾値に起因する問題である。具体的には、NTCIR-4 の「0192 コカコーラ」であり、他の誤り事例に比べると ALD の歪度は大きいものの、 $i(q)$ の値が 0.549 となり、閾値の 0.5 に比べて若干大きかったことが原因である。

(i)～(iv) は本手法で検索質問を分割したことに起因する。それに対して、(v) と (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

本論文で提案した 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 (ベースラインシステム)」を構成し、比較対象とした。表 4 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 2 より、最も精度が高かった既存のアンカー検索手法は A4 である。そこで、A4 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 4 の B4、B5、B6 と比較した。B4 と B5 は本論文で提案したシステムの変形なので、「本システム」と総称する。B6 は検索質問

表 5 検索質問分類の誤り原因と AP/RR の変化
Table 5 Error types of query classification and changes of AP and RR.

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

表 6 検索システム全体に対する MAP と MRR
Table 6 MAP and MRR for retrieval systems.

	B0	B4	B5	B6
MAP	0.272	0.300	0.304	0.312
MRR	0.491	0.519	0.517	0.545

表 7 検索システムの比較に対する t 検定の結果
Table 7 t-test results for differences between retrieval systems.

	MAP	MRR
B0 vs. B4	≪	<
B0 vs. B5	≪	<
B4 vs. B6	—	≪
B5 vs. B6	—	≪

の分類を手で行っているため、精度に関する上限を与える。比較実験の結果を表 6 に示す。表 6 より、本システムは MAP と MRR の両方においてベースラインを上回った。

さらに、表 6 におけるベースラインと本システムの差が偶然ではないことを確認するために両側 t 検定を行った。検定の結果を表 7 に示す。表 7 において、< と ≪ 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 7 より、ベースラインと本システムの MAP や MRR における差は有意であった。本システムと 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、検索質問の分類がつねに正しいシステムに匹敵することが分かった。

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モデルが異なる。本論文は、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モデル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語句は、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、検索質問中の語句が 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。NTCIR で構築された 100GB と 1TB の Web 検索用テストコレクションを用いた実験の結果、本論文で提案したアンカーテキストモデルと検索質問分類手法の有効性が確認された。さらに、提案した手法を統合した検索システムの有効性も確認

された。今後は、調査型と誘導型以外の検索質問へ対処することや検索ログなどの併用について検討する必要がある。また、組織的なスパムへの対策についても検討する必要がある。

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 18049006, 19024007, 21013003) によって実施された。

参 考 文 献

- 1) Baeza-Yates, R., Calderón-Benavides, L. and González-Caro, C.: The Intention Behind Web Queries, *Proc. 13th International Conference on String Processing and Information Retrieval*, pp.98–109 (2006).
- 2) Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks*, Vol.30, No.1–7, pp.107–117 (1998).
- 3) Broder, A.: A taxonomy of web search, *ACM SIGIR Forum*, Vol.36, No.2, pp.3–10 (2002).
- 4) Chung, Y.-J., Toyoda, M. and Kitsuregawa, M.: A Study of Link Farm Distribution and Evolution using a Time Series of Web Snapshots, *Proc. 5th International Workshop on Adversarial Information Retrieval on the Web*, pp.9–16 (2009).
- 5) Craswell, N., Hawking, D. and Robertson, S.: Effective Site Finding using Link Anchor Information, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.250–257 (2001).
- 6) Eguchi, K., Oyama, K., Aizawa, A. and Ishikawa, H.: Overview of the WEB Task at the Fourth NTCIR Workshop, *Proc. 4th NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization* (2004).
- 7) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K.: Overview of the Web Retrieval Task at the Third NTCIR Workshop, *Proc. 3rd NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering* (2003).
- 8) Fujii, A. and Ishikawa, T.: Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration, *Computers and the Humanities*, Vol.35, No.4, pp.389–420 (2001).
- 9) Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments, *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.329–338 (1993).
- 10) Iwayama, M., Fujii, A., Kando, N. and Marukawa, Y.: Evaluating patent retrieval in the third NTCIR workshop, *Information Processing & Management*, Vol.42, No.1, pp.207–221 (2006).

- 11) Kang, I.-H. and Kim, G.: Query Type Classification for Web Document Retrieval, *Proc. 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.64–71 (2003).
- 12) Keen, E.M.: Presenting Results of Experimental Retrieval Comparisons, *Information Processing & Management*, Vol.28, No.4, pp.491–502 (1992).
- 13) Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment, *Proc. 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp.668–677 (1998).
- 14) Lee, U., Liu, Z. and Cho, J.: Automatic Identification of User Goals in Web Search, *Proc. 14th International World Wide Web Conference*, pp.391–400 (2005).
- 15) Li, Y., Krishnamurthy, R., Vaithyanathan, S. and Jagadish, H.V.: Getting Work Done on the Web: Supporting Transactional Queries, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.557–564 (2006).
- 16) Liu, B.: *Web Data Mining*, Springer (2007).
- 17) Lu, W.-H., Chien, L.-F. and Lee, H.-J.: Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach, *ACM TOIS*, Vol.22, No.2, pp.242–269 (2004).
- 18) Oyama, K., Eguchi, K., Ishikawa, H. and Aizawa, A.: Overview of the NTCIR-4 WEB Navigational Retrieval Task 1, *Proc. 4th NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization* (2004).
- 19) Oyama, K., Takaku, M., Ishikawa, H., Aizawa, A. and Yamana, H.: Overview of the NTCIR-5 WEB Navigational Retrieval Subtask 2 (Navi-2), *Proc. 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp.423–442 (2005).
- 20) Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M. and Gatford, M.: Okapi at TREC-3, *Proc. 3rd Text REtrieval Conference* (1994).
- 21) Tsai, M.-F., Wang, Y.-T. and Chen, H.-H.: A Study of Learning a Merge Model for Multilingual Information Retrieval, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.195–202 (2008).
- 22) Voorhees, E.M. and Tice, D.M.: Building a Question Answering Test Collection, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.200–207 (2000).
- 23) Westerveld, T., Kraaij, W. and Hiemstra, D.: Retrieving Web Pages using Content, Links, URLs and Anchors, *Proc. 10th Text REtrieval Conference*, pp.663–672 (2001).
- 24) Yang, K.: Combining text- and link-based retrieval methods for Web IR, *Proc. 10th Text REtrieval Conference*, pp.609–618 (2001).

(平成 22 年 3 月 15 日受付)

(平成 22 年 9 月 17 日採録)



藤井 敦 (正会員)

1993 年 3 月東京工業大学工学部情報工学科卒業。1998 年 3 月同大学大学院博士課程修了。現在、東京工業大学大学院情報理工学研究科准教授、博士(工学)。自然言語処理、情報検索、音声言語処理、Web マイニングの研究に従事。電子情報通信学会、人工知能学会、言語処理学会、日本データベース学会各会員。