検索語の閲覧文書と検索結果における文脈を利用した質問修正

河重　貴洋† 　大島　裕明†† 　小山　聡†† 　田中　克己††

† 京都大学工学部情報学科 〒 606-8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科社会情報学専攻 〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{takahiro,ohshima,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　現在 Web ページやワープロ等の文書を見ている際にその文書中に出現するテキストをクエリとして Web 検索を行うことが多くある。この場合、文書の閲覧と Web の検索の間には関連があり、難しい問題ではあるがユーザが検索に至る経緯を考慮した質問修正を行うことでよりユーザの意図を反映した検索結果を得ることができる。そこでユーザの入力した検索語の検索結果と閲覧文書中の検索語の周辺テキストを元にキーワードの追加を行い質問修正する手法について述べる。閲覧文書の文脈から追加キーワードを選択する際に検索結果の文脈を考慮することで適切なキーワードを追加することができる。

キーワード　情報検索 知識発見 Web マイニング

Query Modification Using Query Contexts both in the Reading Document and in the Search Results

Kawashige TAKAHIRO†, Hiroaki OHSHIMA††, Satoshi OYAMA††, and Katsumi TANAKA††

† School of Informatics, Kyoto University Yoshidahonmati, Sakyou-ku, Kyoto,606-8501 Japan
†† Department of Social Informatics, Graduate School of Informatics, Kyoto University Yoshidahonmati, Sakyou-ku, Kyoto 606-8501 Japan
E-mail: †{takahiro,ohshima,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　While reading a web page or a word processor document, we often search the Web by using a word in the reading document as a query. In this case there is a relation between the intention of the search and the reading document. Modifying the query to reflect the user's intention hidden in the reading document can improve the relevance of search results. However, finding an appropriate query modification is a difficult problem. We propose the query modification using surrounding text of the query terms both in the reading document and in the search results. Considering the context of the query word in the initial search results, we can select appriciate additional keywords for query modification from the reading document.

Key words　information retrieval, knowledge discovery, web mining

1. はじめに

現在，インターネットおよびＰＣが普及し個人が簡単にあらゆる情報を Web 上で閲覧できるようになった．また PC 上でさまざまな作業を行っている際に何らかの情報が知りたいと思ったときには検索エンジンを用いて Web 検索を行うことが多い．しかし Web ページは近年増加の一途をたどっており，Web 検索エンジンの Google [1] に至っては検索対象ページ数が 80 億ページにも上っている．このような莫大な量の Web ページの中から自分の必要とするページを見つけ出すことは困難なことである．

たとえばユーザがＰＣ上で何かの文書を閲覧していて，その文書中の興味を引かれる事柄やわからない単語について調べようとして Google 等の検索エンジンを用いて検索を行うことがある．この場合，検索語をその単語単体とした場合にはその検索語が最も一般的な意味で使われるページが上位にランク付けされて表示されるため，閲覧文書での使われ方がその一般的な用法と違っていた場合には自分の必要とするページを探すのに手間がかかり，場合によっては検索結果の絞込みが行えるように検索語にさらにキーワードを何語か追加して再検索を行うということを繰り返さなければいけない．このような検索質問に適切なキーワードを追加することは検索に慣れていない初心者にとっては難しい作業であるし，検索に慣れたユーザにとっても手間のかかる作業である．

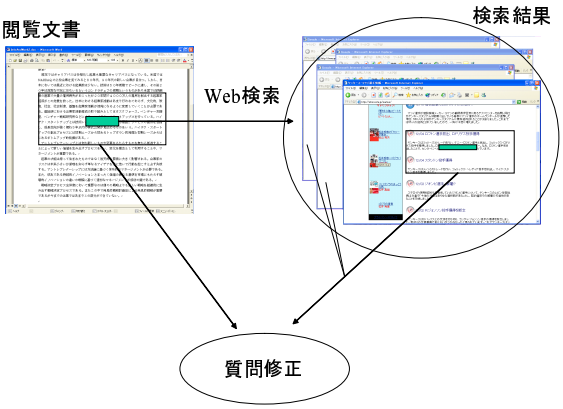

図 1　研究の概要

このような作業を支援するために，本研究ではユーザの入力した検索質問にキーワードを自動的に追加するシステムを提案する．このシステムを用いることによってユーザはただ単に既存の検索エンジンを用いて検索を行った場合に比べ自分の必要とする情報を少ない手間で入手することができるようになると考える．ここで，ユーザはワードプロセッサーの文書や Web ページなどを閲覧していて，その中に現れるテキストで Web 検索を行おうとしているという状況を想定する．そのような場合において，ユーザが入力した検索語と現在閲覧中の文書の内容の間には関連があると考えられる．そこで検索語に現在閲覧中の文書の内容を元にキーワードを追加し質問修正を行うことで検索を行おうとした意図に合った検索結果を得ることができると考える．そこで本研究では検索語の閲覧文書中での文脈と検索結果における文脈を付き合わせることによって質問修正を行う手法について述べる．つまりユーザの入力した検索語を修正する際に検索語の閲覧文書中での周辺テキストと，検索語そのままで Web 検索を行った結果の Web ページの中での検索語の周辺テキストを用いる手法について提案する．この概要を図で示すと図 1 のようになる．この論文では 2. で基本的事項，関連研究について述べ，3. で提案する質問修正の手法について，4. で実装について，5. で実際に実験を行った結果について述べ，6. で課題点について述べ，7. で結論について述べる．

2.　基本的事項および関連研究

2.1　関 連 研 究

2.1.1　Placing Search in Context

Lev Finkelstein ら [2] はユーザが閲覧している文書中で選択されたテキストに選択箇所の周辺テキストからキーワードを追加して検索質問をつくり検索を行うという手法を提案している．この研究では周辺テキストから検索語に追加する選択箇所に関連したキーワードを決定する際に，あらかじめ意味ネットワークを作成しておきそれを用いてキーワードを決定している．これはあらかじめ 27 種類の知識領域，たとえばコンピューターやビジネス，娯楽というような事柄についてそれぞれ文書を集めておく．そして単語を 27 次元のベクトルで表現し，ベクトルの各次元には 27 種類の知識領域が対応する．ベクトルの各値はその知識領域の文書におけるその単語の出現頻度となっている．この意味ネットワークを用いて単語を多次元ベクトルで表現しそのベクトル間の距離を測り，距離の近い単語を検索語に追加するということを行っている．本研究では関連のあるキーワードを選ぶ際に初期検索語での検索結果における検索語の周辺テキストを用いている．よってこの点が本研究とは異なっている．

2.1.2　Watson

Watson Project [3] [4] [5] では閲覧文書を元にユーザの質問修正や検索質問を自動的に生成し検索してユーザに提示するということを行っている．検索質問の修正・作成には文書中の単語を頻度や出現位置等で重み付けを行うことによって追加キーワードの決定や，検索質問の形成をしている．本研究では検索語に追加するキーワードの決定に本研究では閲覧文書だけではなく初期検索語の検索結果を用いている．このように検索結果の文脈を考慮しているという点がこの研究と異なっている．また Watson では閲覧中の文書と反対の意見の Web ページの検索を行ったり，企業情報や地図情報の提示などユーザが必要としている情報を提示するということも行っている．

2.1.3　Automatic Query Expansion (Pseudo Relevance Feedback)

Jinxi Xu ら [6]，Shiping Yu ら [7] はユーザの検索質問の検索結果の上位のページの検索語の周辺テキストからキーワードを抽出し質問修正を行っている．本研究では検索語に追加するキーワードは現在閲覧中の文書中から抽出してきている．この研究では検索結果の文脈は利用しているが閲覧文書の文脈は考慮していないという点が本研究とは異なっている．

2.2　基本的事項

2.2.1　茶　　筌

茶筌 [8] は奈良先端科学技術大学院大学自然言語処理学講座にて開発された日本語形態素解析器であり使用者によって品詞体系，単語認定基準などを容易に変更できるようになっている．本研究では閲覧文書の検索語の周辺テキストを解析して名詞の抽出を行う際に用いている．

2.2.2　適　合　率

適合率とは結果中の正解の文書の割合のことである．

結果の文書総数を M，そのうちの正解の文書数を N とすると適合率 P は以下の式で表される．

$$P = \frac{N}{M} \tag{1}$$

本研究では質問修正の評価において，検索結果を閲覧文書の内容と関係があるかどうかで適合，不適合の判定を行い，適合率を計算することで手法の優劣の基準としている．

3.　質問修正の手法

3.1　提案手法の概要

提案手法の一連の流れは以下のようになる．

（1）　ユーザが閲覧文書に出現するテキストを検索語として選択

（2）　閲覧文書の検索語の周辺テキストの抽出，解析

（3）　そのままの検索語で Web 検索を行い結果の取得

	2.1 の手法	2.2 の手法	2.3 の手法	提案手法
閲覧文書の文脈	○	○	×	○
検索結果の文脈	×	×	○	○

表 1　文脈の使用状況

（4） 取得した検索結果の各文書中から検索語の周辺テキストの抽出

（5） 2 で抽出した名詞を 3 を元に重み付け

（6） 検索語に 5 を元にキーワードを追加

（7） 修正後の検索語で Web 検索

（8） ユーザに検索結果の提示

（9） 必要に応じてキーワードの再追加

3.2 閲覧文書と検索結果の文脈の使用

本研究では閲覧文書の文脈と検索結果の文脈を用いて質問修正を行っている．ここで 2.1 で紹介した関連研究と本研究での提案手法のそれぞれについて閲覧文書の文脈，検索結果の文脈の使用の有無を表 1 にまとめる．

3.3 質問修正の手法

ユーザが閲覧文書中のテキストを検索語として選択した場合に検索語にキーワードを追加して質問修正を行う手法について説明する．

3.4 閲覧文書からの名詞の抽出

まず閲覧文書中からユーザが入力した初期検索語を含む文およびその前後の数文を抜き出し，形態素解析を行う．そしてその中に含まれる名詞を全て抽出する．これらの名詞が検索語に追加されるキーワードの候補となる．これを図で示すと以下の図 2 のようになる．

3.5 初期検索語での検索結果の取得と解析

次にユーザが選択した初期検索語をそのまま検索語として Web 検索を行い検索結果を数十件取ってくる．その検索結果のそれぞれのページ P_i について検索語の出現箇所，およびその周辺テキスト T_i を抽出する．この検索語の周辺テキストを用いて追加するキーワード候補の重み付けを行い，重みの最も高い名詞を初期検索語に追加する．これを図 3 に示す．

3.6 追加キーワードの決定

3.6.1 重みの計算

ここで閲覧文書から抽出した名詞 m_i について

$$x(m_i) = \begin{cases} 1 & m_i \text{ が } T_i \text{ に出現する} \\ 0 & m_i \text{ が } T_i \text{ に出現しない} \end{cases}$$

上式のように $x(m_i)$ を定めるとして名詞 m_i の重み w_i を次

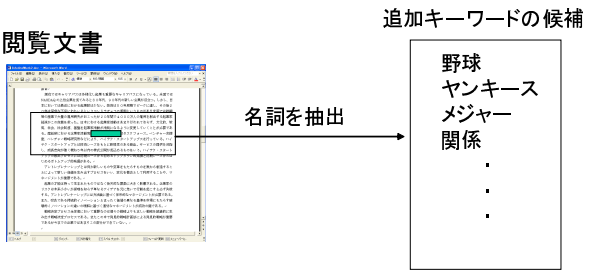

図 2　名詞の抽出

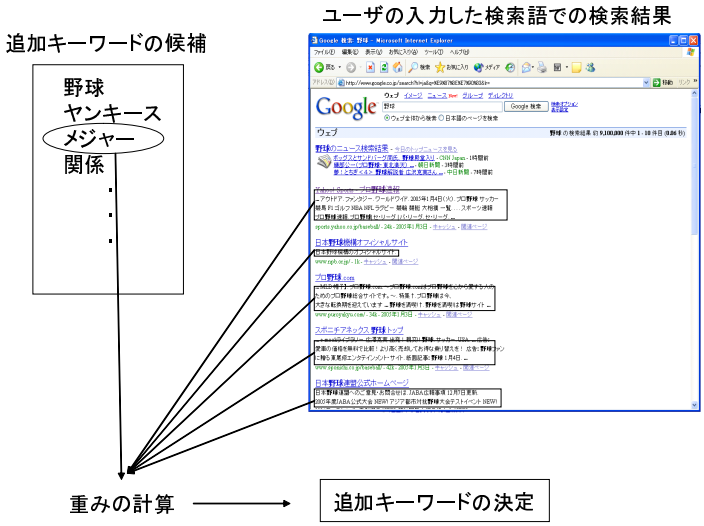

図 3　重みの計算

のように定める．

$$w_i = \sum_{P_i} x(m_i) \tag{2}$$

そして最も重み w_j の大きな名詞 m_j をユーザが最初に選択した検索語に追加して新しい検索語とする．つまり閲覧文書中の検索語の周辺の名詞を抽出し，各名詞に対して初期検索語での検索結果の各ページの検索語の周辺テキストに出現するかどうかを判定し，その出現するページ数をカウントする．そしてそのページ数が最も多い名詞を最初の検索語に追加して新しい検索語とするということである．

3.6.2 フィルタリング

名詞の重みを (2) 式のように定めた場合，検索結果での検索語の周辺箇所の出現頻度だけで追加する名詞を決定しているため多くのページに含まれるような一般的な名詞，例えば「関連」や「関係」といった名詞が追加されてしまうことがある．こういった名詞を追加した場合，質問修正を行わない場合の検索結果の上位ページが多く現れることになり質問修正を行った意味があまりなくなってしまうということがある．このような場合，ユーザの閲覧文書の文脈を反映した質問修正を行っているとはいえない．そこで一般的な名詞の重みを低くするために名詞 m_i で Web 検索を行ったときの検索結果の文書数を $R(m_i)$ として (2) 式を以下のように修正することでこのような問題を解決することを考える．

$$w_i' = \frac{\sum_{P_i} x(m_i)}{R(m_i)} \tag{3}$$

(3) 式の m_i' を新たに名詞 m_i の重みとしてこの重みが最も高い名詞を検索語に追加することにする．この式では，一般的な名詞では検索結果の文書総数 R_i は高くなるためその名詞の重みは低くなる．これは一種の TF-IDF 法のようなものである．T F-IDF 法ではある単語についてその単語の文書中の頻度 t f (term frequency)，およびその単語が出現する文書数 df (document frequency) の逆数 idf (inverse document

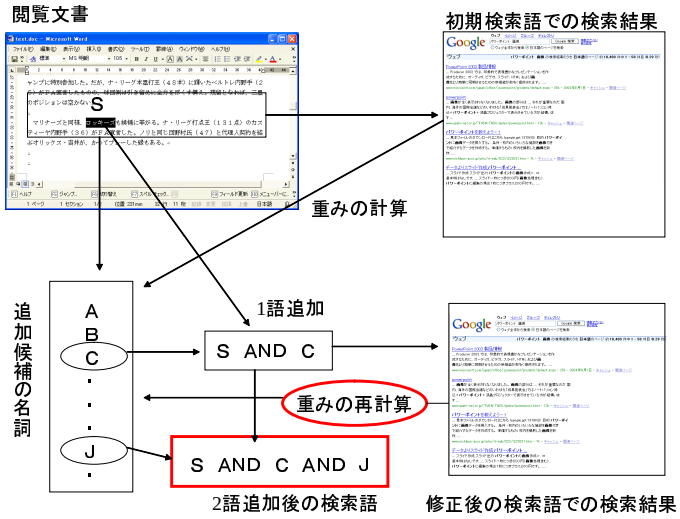

図 4　質問の再修正

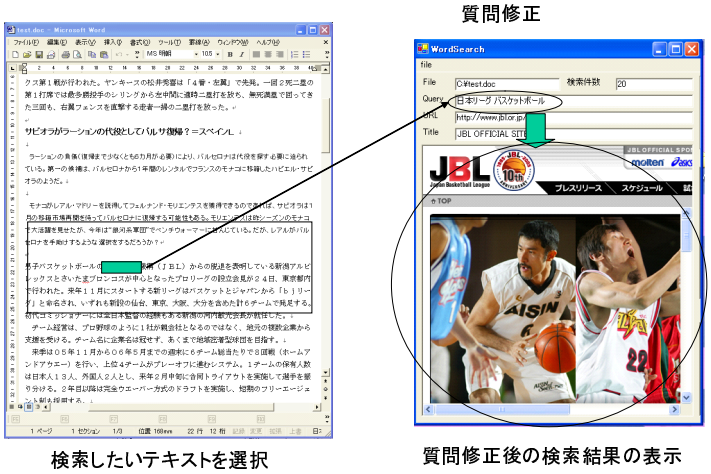

図 5　実 装 画 面

frequency) を求め，tf と idf を掛け合わせた値が単語がその文書を特徴づける度合いとなる．式 3 において R_i で割っている部分が idf に相当する．これによってユーザが最初に選択したテキストに関連した名詞を追加キーワードとして選択することができる．

3.7　キーワードの再追加

この章のこれまでに説明してきた手法で検索語に 1 語キーワードを追加することができる．しかしこの修正後の検索語で Web 検索を行ってもまだ検索結果に満足がいかない場合がある．この場合には検索語のキーワード数を増やして検索結果を絞り込むことが必要である．

そこでさらに検索語にキーワードを追加していくことを考える．

3.7.1　質問修正を繰り返す

検索語にキーワードを単純に追加する方法としては重みの高い順に次々と追加していくという方法が考えられる．しかし 1 語目に追加したキーワードと関連のあるキーワードをさらに追加することで適切な検索結果の絞込みが行えると考えられる．よって，初期検索語に 1 語追加した検索語を元に追加候補の再重み付けを行うことを考える．

3.6 の方法で初期検索語に 1 語を追加してできた検索語で Web 検索を行いその結果を取得し，その検索結果で追加キーワード候補の名詞を再度重み付けを行い，その新たな重みを元に検索語にキーワードをさらに追加する．この流れを図で表すと図 4 のようになる．

初期検索語による検索結果でキーワードの重み付けを行う場合，その検索結果には現在閲覧中の文書に関連する文書だけでなく関連しない文書も多く含まれる．追加候補のキーワードのうち検索結果の周辺テキストに含まれないものの重みは 0 となるため追加候補全体で重みが 0 のものが多く存在する．しかし，初期検索語に 1 語追加した後の検索語の検索結果は初期検索語での検索結果よりも閲覧文書の内容に関連したものが多くなっていると考えられるため重みが 0 のキーワードは最初の場合よりも少なくなる．よって初期検索語の検索結果での重み付けにより重みの上位のものを追加する場合に比べ，重みを再計算した上で重みの高いものをさらにキーワードに追加するほうが追加候補の数が増えることになる．よって複数キーワードを追加していく際に適切なキーワードを追加できる可能性があがると考えられる．よって本研究では複数キーワードを追加していく際には重みの再計算を行う方法を使用する．

4.　実　　装

4.1　実 装 環 境

実装に使用した環境は以下の通りである．

- Microsoft Visual Studio C# .NET
- Microsoft Word

4.2　実 装 方 法

4.2.1　形態素解析

閲覧文書の検索語の周辺テキストの形態素解析を行い名詞を抽出するために茶筌 [8] を用いた．

4.2.2　検索および結果の取得

検索結果や検索結果の文書数など Web 検索を行って取得する情報は Google [1] を用いて取得した．また検索結果における検索語の周辺テキストとして Google の検索結果のページの検索語の周辺テキストを使用した．

4.3　システムの概要

このシステムを実際に実行した場合の画面例を図 5 に示す．

このシステムではユーザがまず MS Word で何か文書を閲覧しているという状況を想定している．その状況で検索を行いたいと思うテキストを Word 上で選択する．するとシステムがその選択箇所の周辺テキストの抽出や，選択箇所での検索などを行って質問修正を行い，その結果生成された質問語で検索を行いその結果をユーザに提示する．

4.4　実　行　例

ここで実際にこのシステムを実行してみる．実際の質問修正の様子を図 6 に示す．

質問修正に用いる検索結果の取得件数は 20 ページ，閲覧文書の周辺テキストの長さは検索語が含まれる文および前後のそ

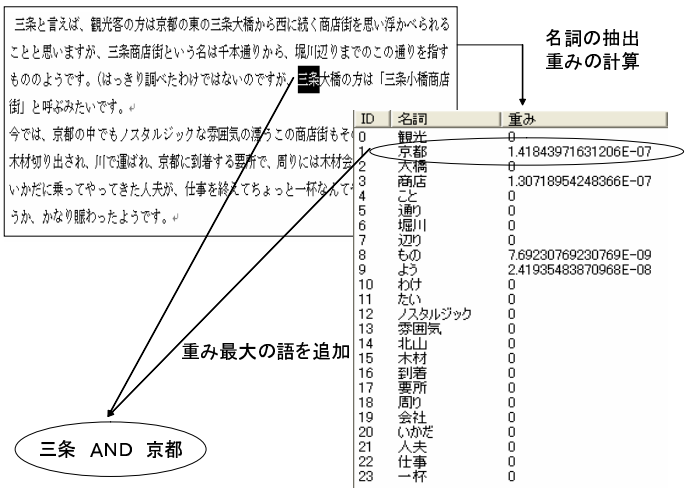

図 6　質問修正例

れぞれ 2 文とする．ユーザは現在 Word で京都の三条商店街ついて書かれている文書を閲覧しているとする．ここで「三条」について検索したいと選択したとする．すると 6 のように周辺テキストを抽出し 3 章で述べた手法で重み付けを行い重みの最も高い「京都」という名詞が追加され修正後の検索語は「三条 AND 京都」というようになる．こうして「三条」でも新潟の三条市の情報などではなく京都の三条通りの情報を検索することができた．

5. 実験結果

この章では本提案手法において実際に質問修正を行った場合の結果について述べる．まずは本提案手法においてパラメータを変化させた場合の修正後質問での検索結果の評価を行いパラメータの値を決定する．そして次に本提案手法での質問修正を行った場合の検索結果と他の手法によって質問修正を行った場合の検索結果について比較を行う．

5.1 パラメータの決定

5.1.1 閲覧文書の周辺テキストの長さ

ここで閲覧文書から抽出する周辺テキストの長さについて考察する．提案手法においては周辺テキストから検索語に追加するキーワードを選択するため，この長さを適切な長さにしておくことが重要である．周辺テキストの長さを長く取ればその分検索語に追加するキーワードの候補数が増えるため，より適切であると思われるキーワードを追加できる可能性が上がる．しかし長く取り過ぎると選択語からの距離が遠い文では選択された検索語とはあまり関係のない内容になっている可能性があり，またその分の計算時間が余分にかかってしまい無駄である．また周辺テキストの長さを短くすると計算時間が少なくてすむが，追加キーワードの候補数が少なくなるため追加するのに適切な単語が抽出できないことが考えられる．そこで周辺テキストの長さを変化させてそこから抽出される追加キーワード候補の名詞の数を比較することで妥当であると考えられる周辺テキストの長さを決定する．周辺テキストを選択語を含む文およびその文の前後それぞれ 0 文，1 文，2 文，3 文と変化させた場合の追加候補の名詞の平均数を表 2 に示す．表 2 前後の 0 文，1 文の場合では名詞数が 20 個以下となった．この場合関連のない名詞が含まれていることを考えると追加するのに適切な名詞の数は少なすぎると考えられる．また 3 文以上では，文の数を長く取ればとるほど名詞数は増えるが計算の実行時間が長くなってしまう．そこで今後は周辺テキストの長さを選択語を含む文および前後の 2 文とする．

5.1.2 検索結果の文脈として使用するページ数

次に閲覧文書の周辺テキストから抽出した追加キーワード候補の名詞の重み付けに使用する検索結果のページ数について考える．このページ数によって名詞の重み付けが変わるため，この数しだいでは追加するキーワードが変わってしまうことがある．使用するページ数が少なすぎる場合には閲覧文書の内容に関連のあるページが検索結果に現れないことが得る．この場合関連のある名詞の重みが高くならないため選択語とはあまり関連のないキーワードが検索語に追加されてしまうことになる．また使用ページ数が多すぎる場合においては計算時間が長くなりすぎてしまうという問題がある．そこで使用する検索結果のページ数を変化させたときに重みが 0 より大きい名詞数を比較することで妥当であると考えられるページ数を決定する．表 3 に使用する検索結果のページ数を 5 件，10 件，20 件，30 件とした場合の重みが 0 より大きい名詞の平均個数を示す．

表 3 より 5 件，10 件の場合には重みが 0 より大きくなる名詞数が 2 個程度となっている．これでは追加候補数が少なすぎるため適切な質問修正が行えないと考えられる．20 件以上を用いた場合には多くすればするほど候補数が増えるがその分計算時間も増えるので以後は重みの計算に使用する検索結果のページ数を 20 件とする．

5.2 他手法との比較実験

ここでは本提案手法と他の手法との質問修正の比較を行う．比較を行う他の手法としては，次の閲覧文書の文脈を用いて質問修正を行う手法，そして検索結果の文脈を用いて質問修正を行う手法の 2 通りを用いる．

（1） 閲覧文書の検索語の周辺テキストにおける名詞の出現頻度が高いものを検索語に追加する

（2） 初期検索語の検索結果の上位文書の検索語の周辺テキストの中の名詞の出現頻度が高いものを検索語に追加する

質問修正を行う初期検索語としては同音で複数の使われ方をする以下の単語を用いる．

- 府中市
  - 広島県の府中市
  - 東京都の府中市
- ピッチャー

表 2　候補数の変化

	0 文	1 文	2 文	3 文
平均候補数	6.6	18	33.4	44.2

表 3　重みが 0 より大きい平均候補数

	5 件	10 件	20 件	30 件
平均数	1.8	2.4	5.0	6.2

– 投手
– 水差し
● マフラー
– 車・バイクのマフラー
– 襟巻き
● キーボード
– 楽器
– PC 等の入力機器
● 三条
– 京都の三条通り
– 新潟の三条市
● ジャガー
– 車のジャガー
– 動物のジャガー

各単語の 2 通りの意味で使われている文書をそれぞれ 5 文書程度用いて，その単語を初期検索語としたときの質問修正後の検索結果 20 件中における元の文書の使われ方との適合文書の適合率を見ることで比較を行う．

また元の検索語に 1 語追加する場合と 2 語追加する場合について比較する．

5. 2. 1 検索語に 1 語追加する場合

初期検索語に 1 語追加した場合の各手法における適合率および初期検索語のみで検索を行った場合の適合率の平均値を表 4 に示す．

検索結果のみで質問修正を行った場合の適合率を見てみると，どの単語の場合においてもどちらか一方の適合率が極めて高く，もう一方の適合率は 0 に近い値となっている．府中市の結果を見てみると東京都の府中市に関するページを見ているときの検索結果の中の東京都のページの割合は 100%だが，広島県の府中市の場合は 5%となっている．これは閲覧文書の文脈を考慮していないため，どのような文書を閲覧していても常に検索結果に占める割合の多いページに関連するキーワードを追加しているからである．本提案手法においては閲覧文書の文脈を考慮しているため常にどちらか一方の意味に質問修正をしてしまうということは起こっていない．

表 4　1 語追加時の適合率の平均値

		提案手法	閲覧文書	検索結果	修正なし
府中市	広島県	85.7	51.4	5	25
	東京都	69.4	70	100	70
ピッチャー	投手	94	90	100	30
	水差し	100	51.3	0	50
マフラー	車・バイク	90	65	100	65
	襟巻き	74	89	0	20
キーボード	楽器	10.6	41.4	0	5
	入力機器	95.4	89.2	100	95
三条	京都	100	58.8	5	30
	新潟	93.3	82.5	70	50
ジャガー	車	65	84	70	40
	動物	7	40	0	5

表 5　京都の三条の追加例と適合率

	提案手法		閲覧文書	
	キーワード	適合率 (%)	キーワード	適合率 (%)
京都の三条	カフェ	100	烏丸	100
	中京	100	10	30
	京都	100	商店	35
	京都	100	プラン	70
水差しのピッチャー	水差し	100	ウツボカズラ	10
	水差し	100	取っ手	90
	水差し	100	もの	40
	水差し	100	作品	65

次に本提案手法と閲覧文書の頻度による手法の適合率について比較を行う．表 4 を見てみると「府中市」,「三条」,「ピッチャー」の例では本提案手法における適合率が閲覧文書のみによる手法と比べて同じくらい，またはそれ以上の値となっている．適合率の差が大きい「京都の三条」,「水差しのピッチャー」の場合について実際に追加されたキーワードとその時の適合率を表 5 に示す．表 5 を見ると閲覧文書の頻度で質問修正を行った場合にはその文書に関連したキーワード，たとえば「京都の三条」の文書では「烏丸」であり「水差しのピッチャー」では「取っ手」といった単語の頻度が高い場合は高い適合率となっているがそれ以外の場合にはあまり関連のないキーワードが追加されて適合率が低くなっている．本提提案手法においては閲覧文書の頻度に関係なく関連のある「京都」や「水差し」といったキーワードが追加できている．

逆に「音楽のキーボード」,「動物のジャガー」の適合率は本提案手法における適合率が極めて低い値となっている．「音楽のキーボード」の例においてそれぞれの手法での実際に追加されたキーワードとその時の適合率を表 6 に示す．初期検索語のみで検索を行った場合の適合率を見ると，本提案手法での適合率が極めて低い箇所と初期検索語のみでの適合率が極めて低い箇所が一致することがわかる．これは初期検索語においての検索結果に元の閲覧文書と関連するページがあまり含まれない場合には，関連するキーワードの重みが低くなってしまいその結果あまり関連しない「コントロール」や「現実」といったキーワードを追加してしまうためであると考えられる．したがってこのような場合においては本提案手法では適切な質問修正ができない．

また広島県の府中市の平均適合率は 85.7%と閲覧文書の頻度

表 6　音楽のキーボードの追加例と適合率

	追加されたキーワード	適合率 (%)
提案手法	コントロール	20
	現実	5
	コンパクト	0
	マルチ	0
閲覧文書	キーボーディスト	75
	バンド	60
	鍵盤	70
	ライブ	60

表 7　2 語追加時の適合率の平均値

		提案手法	閲覧文書
府中市	広島県	85	62.6
	東京都	95.6	89.4
ピッチャー	投手	96	87
	水差し	100	64
マフラー	車・バイク	96.2	76.5
	襟巻き	82	96
キーボード	楽器	25.6	75.6
	入力機器	90.4	88.5
三条	京都	100	100
	新潟	99.2	88
ジャガー	車	66	91
	動物	31.3	35

による手法よりも高い値となっている．そのうち大半の質問修正例では「府中市　広島」といった広島県の適合率が高くなるような検索語に修正することができた．しかし東京にある広島県の府中市のアンテナショップについての文書で「府中市」を質問修正した場合には「府中市　東京」と質問修正をしてしまい適合率が 0%となってしまった．このように 2 通りの使われ方をする名詞の 2 通りのそれぞれに関連する名詞が周辺テキストに出現する場合には本提案手法では質問修正がうまくいかないことがあった．

5.2.2　検索語に 2 語追加する場合

次に検索語に 2 語キーワードを追加する場合についての比較を行う．なお検索結果の頻度による手法は 5.2.1 での検索語に 1 語追加した場合の結果からどのような文書を閲覧していても検索結果の上位文書中の占める割合が多いページに関連する質問修正となってしまうことが分かったのでここでは提案手法と閲覧文書の頻度による手法についての比較を行う．また 2 語追加する手法としては提案手法においては 3.7.1 で述べた重みの再計算を行う方法を用い，閲覧文書の頻度による手法では単純に頻度の上位 2 個を追加するものとする．

２語追加した場合の本提案手法と閲覧文書の頻度による質問修正での検索結果の適合率の平均を表 7 に示す．

5.2.1 での 1 語追加した場合と同様に「楽器のキーボード」，「動物のジャガー」を本提案手法で質問修正した場合の適合率は低くなっている.「ジャガー」と「キーボード」以外の例を見てみると「襟巻きのマフラー」の例を除いては 5.2.1 の 1 語追加した場合と同様に，本提案手法のほうが閲覧文書の頻度で質問修正を行った場合よりも適合率が高い結果となっている．

また 1 語追加した場合の適合率と比較して最も適合率の上昇が大きい「東京の府中市」の場合について実際に検索語に追加されたキーワードをいくつか表 8 に示す．文書１，２は東京の府中市で起きた三億円事件に関する文書であり，文書３，４は東京都府中市の郷土の森博物館に関する文書である．いずれの場合も閲覧文書の頻度で質問修正を行った場合よりも本提案手法で質問修正を行ったほうが適合率が高くなっている．

表 8　東京の府中市の質問修正例

	提案手法		閲覧文書	
	追加キーワード	適合率 (%)	追加キーワード	適合率 (%)
文書 1	東京　事件	95	支店　ボーナス	85
文書 2	12 月　事件	85	捜査　本部	80
文書 3	郷土　博物館	100	００　郷土	95
文書 4	プラネタリウム 郷土	95	博物館　敷地	85

6. 課　題　点

本提案手法の課題点としては以下があげられる．

- 重み付けに使用する検索結果に関連文書が少なかった場合

5.2.1 でも述べたように，重み付けを行う際に取得した検索結果中に現在閲覧中の文書と関連したページが少なかった場合は適切なキーワード追加ができないという問題がある．この問題を解決するには検索結果の使用件数を増やして関連ページ数を多くすることが考えられるが，計算時間が増えることや，検索件数を増やしても関連ページがあまり含まれていないこともあるため今後検討が必要である．

- 周辺テキストに別の意味に特定されるようなキーワードがある場合

5.2.1 の最後で述べたように「広島県の府中市」についての文書を閲覧している場合の周辺テキストに「東京」が出現していて「府中市 AND 東京」と質問修正してしまうことがある．このような問題も今後検討する必要がある．

- 計算時間の縮小

本提案手法での検索においてはユーザが必要としたときにすぐに実行，表示を行う必要があり，あまり時間がかかってしまっては実用性がない．そこで計算時間をできる限り短縮することが必要である．そこで現在質問修正を行う際に必要な情報はそのつど取得してきているがデータをローカルにキャッシュすること等で計算時間の短縮を検討したいと考えている．

7. 結　　論

本研究では閲覧文書の文脈と検索結果の文脈を用いることでユーザが検索に至る経緯を考慮した質問修正を行う手法について提案した．ユーザの現在閲覧している文書から初期検索語に追加するキーワードを抽出し，そのキーワードを初期検索語での検索結果によって重み付けを行うことで，閲覧文書の単語の頻度だけで質問修正を行う手法や検索結果の頻度だけで質問修正を行う手法よりも閲覧中の文書の内容に応じた検索結果を多くの場合に得ることができた．

しかし本提案手法で質問修正を行う場合計算時間が長くなってしまう．特に今後検索語に追加するキーワード数を増やしていくことを考えると計算時間を短縮することを考えなければいけない．

また，現在は閲覧中の文書に関連する Web ページを検索することを目的としているが，今後は閲覧中だけでなく，文書作

成中，編集中にも，また関連する文書だけでなく，反対の内容の文書や情報を補完する文書などユーザが必要とする情報を提示するような質問修正ができるようにしていきたいと考えている．

謝　　辞

本研究の一部は，文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検索・統合ソフトウェア開発」（代表：田中克己）および，平成 16 年度科研費特定領域研究 (2) 「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247，代表：田中克己）および，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」および，平成 16 年度科研費若手研究（B）「参照の同一性判定に基づく複数 Web ページの検索閲覧方式の研究」（課題番号：16700097，代表：小山 聡）によるものです．ここに記して謝意を表すものとします．

文　　献

[1] Google
http://www.google.com/.

[2] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman and E. Ruppin.: “Placing search in context: The concept revisited.”, In Proceedings of the Tenth International World Wide Web Conference (WWW10) (May 2001).

[3] J. Budzik and K. Hammond: “Watson: Anticipating and contextualizing information needs”, 62nd Annual Meeting of the American Society for Information Science, Medford, NJ (1999).

[4] J. Budzik and K. Hammond: “User interactions with everyday applications as context for just-in-time information access”, Proceedings of the 2000 International Conference on Intelligent User Interfaces, New Orleans, Louisiana, ACM Press (2000).

[5] J. Budzik, K. J. Hammond, L. Birnbaum and M. Krema: “Beyond similarity”, Proceedings of the 2000 Workshop on Artificial Intelligence and Web Search, AAAI Press (2000).

[6] J. Xu and W. B. Croft: “Query expansion using local and global document analysis”, SIGIR (1996).

[7] S. Yu, D. Cai, J.-R. Wen and W.-Y. Ma: “Improving pseudo-relevance feedback in web information retrieval using web page segmentation”, Technical report, Microsoft Corporation (2002).

[8] 形態素解析システム茶筌
http://chasen.naist.jp/hiki/ChaSen/.