DEWS2006 3C-i14

質問修正と再ランキングを用いた文脈依存 Web 検索

河重　貴洋† 　小山　聡† 　大島　裕明† 　田中　克己†

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

あらまし　Web ページやワープロ等の文書を見ている際、その文書中に関連する内容を求めて Web 検索を行うことがある．この場合，文書の閲覧と Web の検索の間には関連がある．そこで，閲覧文書の内容を用い，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，ユーザの意図を反映した検索を行うことができる．よって，本研究ではユーザの入力した検索語での検索結果における各ページでの検索語の周辺テキスト及び，閲覧文書での検索語の周辺テキストの付き合わせを行い，文脈を反映させた質問修正を行う．またユーザが検索語を能動的に入力しない場合にも，閲覧文書の内容を元に自動的に検索語を作成し，検索を行う手法について述べる．

キーワード　情報検索 Web とインターネット データマイニング

Context-aware Web Search with Query Modification and Reranking

Takahiro KAWASHIGE†, Satoshi OYAMA†, Hiroaki OHSHIMA†, and Katsumi TANAKA†

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

Abstract　While reading a web page or a word processor document, we often search the Web to get Web pages related to the document we read. 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, web and internet, data mining

1. はじめに

PC で文書を作成・閲覧している際，その文書の内容に関連した情報や，その文書中の特定のキーワードについて調べるために，検索エンジンを用いて Web 検索を行うことが多い．しかし既存の検索エンジンを用いて検索を行う場合，検索語を自分で考えなければならない．また，自分で作成した検索語が一般的なものであった場合，検索結果には求める情報以外の雑多なページが多く含まれてしまうため，検索語の追加や変更を行わなければならない．このように文書を作成・閲覧していて Web 検索を行う場合，閲覧文書の内容にその検索語を特徴付ける文章が含まれていると考えられる．そこで，閲覧文書の文脈を検索に利用することで，ユーザは検索語を工夫する手間をかけず，より関連したページを得ることができる．ゆえに，本研究では文書中のあるキーワードが初期検索語として選択された際に，検索語の閲覧文書中での周辺テキスト，及び初期検索語で Web 検索を行った検索結果の各ページでの検索語の周辺テキストを用い，質問修正を行う手法について述べる．またユーザが明示的に検索語を指定しない場合においても，閲覧文書の内容を用いることで自動的に検索語の作成を行うクエリフリー検索の手法についても述べる．

2. 章で関連研究について述べ，3. 章で質問修正の手法について述べ，4. 章でクエリフリー検索の手法について述べ，5. 章でまとめと今後の課題について述べる．

2. 関連研究

2.1 閲覧文書の文脈を用いた手法

Lev Finkelstein ら [1] はユーザが文書中から選択した検索語に，選択箇所の周辺テキストからキーワードを抽出し，そのキーワードを検索語に追加することで検索質問の修正を行って

	2.1	2.2	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

いる．この研究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，あらかじめ作成した意味ネットワークを用いて単語の距離を測り，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．この点で本研究とは異なっている．

Watson Project [2] [3] [4] 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，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，また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．

2.2　検索結果の文脈を用いた手法

Jinxi Xu ら [5] や Shiping Yu ら [6] 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．

2.3　閲覧文書と検索結果の文脈の使用

本章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質問修正の際に閲覧文書の文脈，検索結果の文脈を使用の有無を表 1 にまとめる．

3.　閲覧文書と初期検索語での検索結果を利用した質問修正

3.1　提案手法の概要

文書閲覧時にその文書中に出現する興味を持った語や，ユーザの知らない語が現れた際に，Web 検索を行いその語に関する情報を得ようとすることがよくある．例として，京都に関する文書を閲覧していて，三条通りについての記述が出てきたとき，三条がどのようなところかを知りたくなり “三条” を検索語として検索し，関連する情報を得ようとする場合を考える．しかし，三条は京都だけではなく新潟にも同じ名前の地名が存在するので，このように “三条” の 1 語だけを検索語として検索を行った場合，検索結果には求める京都の三条のページだけではなく新潟の三条のページも含まれてしまう．この問題を解決するために，ユーザは検索語をさらに追加するか，検索語の変更を行い，京都の三条のページを多く取得できるように工夫しなければならない．このような状況の検索においては，ユーザは京都に関する文書を閲覧している際に，その文書中に出現する “三条” について検索を行ったのだから，文書中には京都に関するキーワードが含まれていると考えられる．ゆえに,、閲覧中の文書を解析し，その中から三条の京都を特徴付けるようなキーワードを抽出することができれば，そのキーワードを検索語に追加することで，ユーザが求める京都の三条に関するページを得る支援ができる．ここでユーザが閲覧文書中から選択したキーワードを初期検索語と呼ぶことにする．本章では初期検索語を閲覧文書と検索結果の両方の文脈を用いて質問修正する手法ついて述べる．本提案手法の流れを以下に示す．

（1）　初期検索語の選択

最初の段階として，ユーザは閲覧中の文書の中から知りたい，または興味があるキーワードを初期検索語として能動的に選択する．

（2）　閲覧文書からの追加候補のキーワードの取得

初期検索語を特徴づけるキーワードは，閲覧文書中でも特にその初期検索語の周辺のテキストに含まれていると考えられる．そこで閲覧文書において初期検索語の周辺テキストからキーワードを抽出し，それらのキーワードの中から初期検索語を最も特徴付けると考えられる語の選択を行う．このキーワードの抽出については 3.2 章で詳しく述べる．

（3）　候補のキーワードの重み付け

次に抽出キーワードの中から初期検索語を最も特徴付けると考えられる語の選択を行う．閲覧文書と同じ文脈で初期検索語が用いられている文章において，初期検索語との共起度の高いキーワードは初期検索語を特徴づける語であると考えられる．よって初期検索語で Web 検索を行い，その検索結果での検索語の周辺テキストでの共起数を計算し，キーワード候補の重み付けを行う．重み付けの手法の詳細については 3.3 章で述べる．

（4）　検索語の追加・検索

(3) で求めた重みの値により，初期検索語を特徴付けるキーワードを選択し，そのキーワードを初期検索語に追加する．こうして修正された検索語で Web 検索を行い，ユーザに得られた結果を表示する．

このようにユーザが選んだ初期検索語に閲覧中の文書の文脈を反映させたキーワードを追加することで，より閲覧文書に関連した Web ページをユーザに提示することができる．以下で質問修正を行う手法の詳細ついて具体的に説明する．また質問修正の流れを図 1 に示す．

3.2　閲覧文書からの名詞の抽出

閲覧文書の解析を行う．まず閲覧文書中からユーザが選択した初期検索語を含む文，及びその前後の数文を抜き出す．この初期検索語の周辺テキストを初期検索語の閲覧文書における文脈ととらえる．この中には初期検索語の閲覧文書での文脈を特徴付けるキーワードが含まれていると考えられる．そこで，この周辺テキストの形態素解析を行い，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．図 1 においてはステップ 1. でユーザが “三条” を検

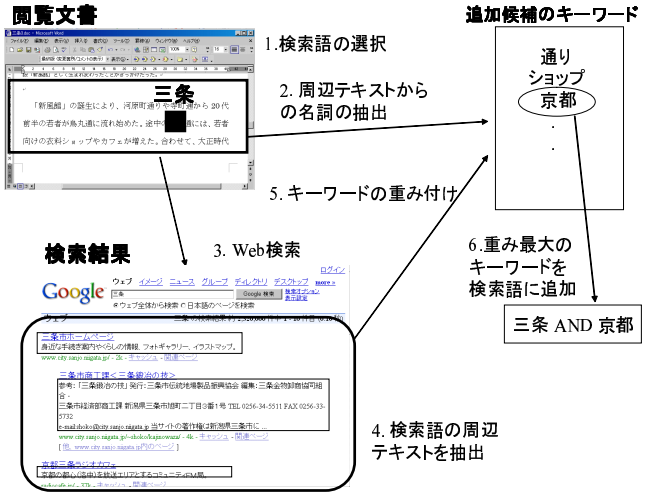

図 1　質問修正の概要

索語として選択し，ステップ 2. において “三条” の周辺テキストを解析し “通り”，“ショップ”，“京都”，“情報” 等の名詞を追加候補のキーワードとして取得する部分に対応する．

3.3　重 み 付 け

次に抽出されたキーワード候補の中から最も閲覧文書の文脈を反映させるキーワードの選択を行う．選択するために選択基準となるキーワード候補の重み付けを行う．重み付けには初期検索語での Web 検索結果を用いる．まず初期検索語で Web 検索を行い，検索結果のページ群を取得する．これは図 1 において，ステップ 3. の部分で初期検索語 “三条” での Web 検索を行い，結果を取得している部分に対応する．取得したページの中から，閲覧文書と同じ文脈で検索語が使われているページをできるだけ多く重み付けに使うことが理想的である．しかし検索結果の全てを利用することは現実的でない．また計算に用いるページ数が多くなれば，重み付けの処理にかかる実行時間が増加してしまう．よって取得した検索結果のうちの N ページを選択し，重み付けに利用する．

3.3.1　候補の名詞の検索結果における共起数の計算

本提案手法の基本的な考え方としては，初期検索語の周辺にあるキーワードの中で，初期検索語を特徴付ける語は検索結果の中の同じ文脈のページにおいても初期検索語の周辺に出現するという仮定がある．そこで取得した N ページの検索結果において，検索語の周辺テキストにおける共起数を計算し，この共起数をキーワードが閲覧文書の文脈を代表する度合いを表しているとみなす．そこで重み付けの最初の段階として，取得した検索結果の各ページ $P_i, i = 1, ..., N$ それぞれにおいて，検索語の周辺テキスト $T_i, i = 1, ..., N$ を取得する．この T_j を検索結果の文脈として考え，重み付けの計算に用いる．システムの実装においては，簡単のため Google [7] の検索結果の各ページの要約部分 (Snippet) に検索語とその周辺テキストが出現することが多いため，これを検索結果の文脈の代用として用いる．次に，それぞれのキーワード候補 K が検索結果の文脈 $T_i, i = 1, ..., N$ に出現する回数を求める．

$$f_j(K) = \begin{cases} 1 & K \text{ が } T_j \text{ に出現する} \\ 0 & K \text{ が } T_j \text{ に出現しない} \end{cases}$$

上式のように $f_j(K)$ を定め，キーワード K の検索結果における出現数 $o(K)$ を 1 式のように定める．

$$o(K) = \sum_j f_j(K) \tag{1}$$

3.3.2　重みの計算

名詞の重みとして (1) 式の検索結果の文脈での出現数のみを考慮する場合，あるキーワードが初期検索語に関連がない場合でも，その語がどのような内容にも現れる一般的な語であった場合，出現数が多くなることがある．例えば「関連」や「ページ」といった一般的な名詞は閲覧文書の文脈に関係なく多くの文書に出現するため，共起度が高くなり，追加語として選ばれてしまうことがある．このようなキーワードはその文書の文脈を特徴付けているとはいえないので，初期検索語に追加した場合，質問修正として適切であるとはいえない．そこで一般的な名詞の重みを低くするために，キーワード K で Web 検索を行った際の検索結果の検索件数を $R(K)$ として，一般的なキーワードの重みを低くするようにキーワード K の重みを (2) 式のように定義する．

$$w = \frac{o(K)}{R(K)} \tag{2}$$

(2) 式の重みをすべてのキーワード候補について計算し，重みが最も高いキーワードを検索語に追加する．この式では一般的な名詞では検索結果の総数 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 IDF 法のようなものである．これは図 1 において，ステップ 5. の検索結果の文脈を用いてキーワード候補の重み付けを行うという部分に対応する．一般的な名詞の重みを低くする処理を行うことにより，候補中から “情報” のような一般的なキーワードではなく，“京都” のように三条の京都を特徴づけるようなキーワードが初期検索語に追加される．こうして修正された検索語で Web 検索を行い，ユーザに結果の Web ページを提示する．

3.4　重み付けに使用する検索結果の選択

この章では，閲覧文書における初期検索語の周辺テキストを追加キーワードの抽出に用い，初期検索語での検索結果中の N ページを選び，そのページ中での検索語の周辺テキストをそのキーワードの重み付けに使用することで，初期検索語に関連のある検索語を追加する手法について述べた．ここで，検索結果の中から重み付けに利用する N ページを選ぶ方法について述べる．本提案手法では，重み付けの際，閲覧文書と同じ文脈で使われているページにおいて共起数の多い語を検索語に追加する語としてふさわしいと考える．そこで検索結果の中から同じ文脈のページをを多く取得し，重み付けに用いることが重要である．

3.4.1　検索結果の順位を利用

検索結果から N ページを選択する一番単純な方法として，初期検索語での検索結果での上位文書に閲覧文書と同じ文脈の文

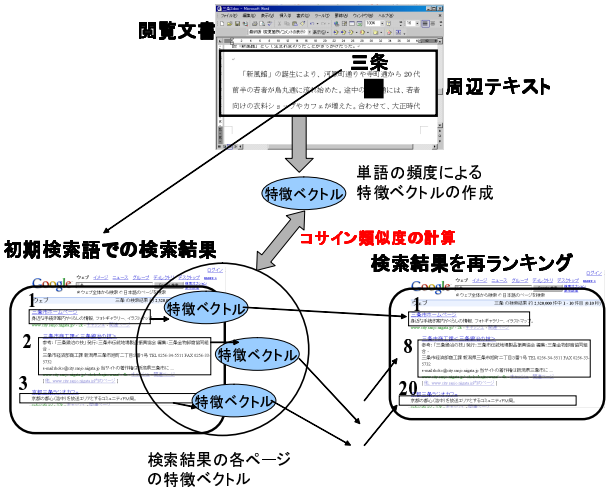

図 2　類似度による再ランキング

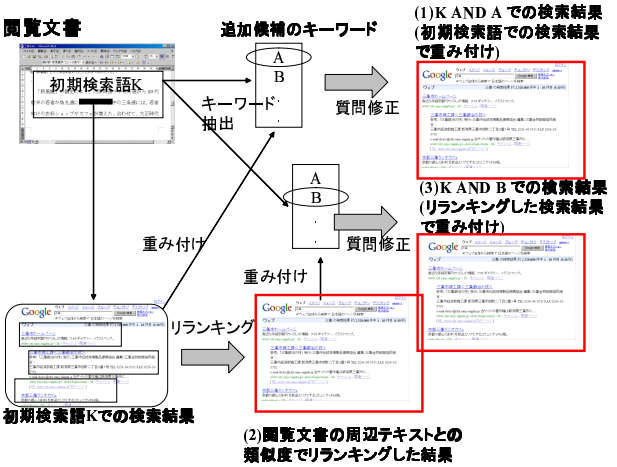

図 3　比較する手法の流れ

書が含まれると仮定する方法がある．この方法では初期検索語で Web 検索を行い，得られた検索エンジンでの順位をそのまま利用し，上位 N ページを取得するという方法である．

3.4.2　周辺テキストの類似度

3.4.1 章の手法で N ページを取得する方法では，閲覧文書で初期検索語が余り一般的でない使われ方をしていた場合，検索結果の上位には閲覧文書と同じ文脈のページが少なくなり，キーワードの適切な重み付けができない可能性がある．そこで，同じ文脈のページをより多く選択するために，周辺テキストの類似度を用いる手法について述べる．この手法では検索結果の上位 R ページ $(R > N)$ の中から閲覧文書との類似度が高い N 件を重み付けに利用することになる．まず閲覧文書の検索語の周辺テキスト，及び検索結果の各ページにおける検索語の周辺テキストのそれぞれについて名詞を抽出し，その頻度を計算し，特徴ベクトルを作成する．キーワード k_i の周辺テキスト ST における値 $v_{ST}(k_i)$ を (3) 式で求める．

$$v_{ST}(k_i) = TF_{ST}(k_i) \tag{3}$$

ここで $TF_{ST}(k_i)$ はキーワード k_i が周辺テキスト ST に出現する回数を表す．このようにして閲覧文書の周辺テキスト ST_s，及び検索結果の各ページごとの周辺テキスト $ST_{r_j}, j = 1, ..., R$ について名詞の頻度による特徴ベクトル v_{ST_s}，$v_{ST_{r_j}}$ を計算する．そして求めた閲覧文書の特徴ベクトル v_{ST_s} と検索結果の各ページの特徴ベクトル $v_{ST_{r_j}}$ 間でのコサイン類似度を (4) 式により求める．

$$cos(v_{ST_s}, v_{ST_{r_j}}) = \frac{\sum_i (v_{ST_s}(k_i) * v_{ST_{r_j}}(k_i))}{\sqrt{\sum_i v_{ST_s}(k_i)^2 * \sum_i v_{ST_{r_j}}(k_i)^2}} \tag{4}$$

こうして求めた検索結果の各ページのコサイン類似度の値の高い上位 N 件を重み付けに使用するページとして選択する．

3.5　実　　験

ここまでに述べてきた手法について実際に処理を行い，得られる結果の評価を行う．実験は以下の 3 手法について行うするものとする．

（1）　検索結果の上位 N 件を重み付けに用いた質問修正

本提案手法で質問修正を行う．キーワード候補の重み付けの際に，3.4.1 章で述べた検索結果の上位の 20 ページを用いる．

（2）　初期検索語での検索結果の特徴ベクトルの類似度による再ランキング

3.4.2 章では，検索結果の中から重み付けに使用する N ページの選択をする際に，閲覧文書の周辺テキストの特徴ベクトル，及び検索結果の各ページでの検索語の周辺テキストの特徴ベクトル間での類似度を計算し，その値の高いものを重み付けに用いる手法を説明した．しかし，ここで重み付けによる質問修正を行わなくとも類似度の高いページそのものがユーザが求める閲覧文書に関連するページである可能性もある．そこで比較のために，質問修正をおこなわず，類似度の高い順に検索結果を並べ替えただけの手法についても実験を行う．

（3）　重み付けに特徴ベクトルの類似度で再ランキングした結果を用いた質問修正

本提案手法で質問修正を行う．キーワード候補の重み付けの際に，検索結果の中から 3.4.2 章の手法を用い類似度を求め，その値の高い上位 N ページを用いる手法である．

実験を行う 3 つの各手法の流れを図 3.5 に示す．実験の条件として，ユーザは初期検索語として多義性のある語を選択したという状況を想定する．具体的には，京都に関する文書の中に出現する “三条”，新潟に関する文書の中に出現する “三条”，車に関する文書の中に出現する “ジャガー”，動物に関する文書の中に出現する “ジャガー” のそれぞれの語を初期検索語としてユーザが選択したと仮定し，3 つの手法それぞれを用い処理を行い，最終的に得られる結果を 20 件取得する．この 20 件の文書の中で，閲覧文書と同じ意味の文書の割合 (例えば，京都の文書で “三条” を初期検索語として処理を行い，得られた結果の 20 件中の京都のページの割合) を適合率として計算し，この適合率の値により手法の比較を行う．閲覧文書の周辺テキストとしては，初期検索語を含む文，及びその前後の 2 文づつの計 5 文を用いる．検索結果のページにおける周辺テキストは Google [7] の要約部分を用いる．また (2)，(3) の手法でコサイ

		(1) 質問修正	(2) 再ランキング	(3) 質問修正 +再ランキング
三条	京都	77.0	62.5	75.5
	新潟	70.0	76.5	77.0
ジャガー	車	79.0	75.0	77.5
	動物	14.3	2.1	8.6

表 2　各手法での平均適合率 (%)

ン類似度を初期検索語での検索結果の全てについて計算することはできないため，検索エンジンの検索結果の上位 100 件について計算を行うことにする．4 種類の初期検索語それぞれについて (1)，(2)，(3) の手法それぞれで 10 回程度実験を行い，適合率の平均を計算した．その結果を表 2 に示す．

表 2 を見ると，動物のジャガーの場合はどの手法でも適合率が低い結果となった．これは “ジャガー” での検索結果の上位はほとんどが車メーカーのジャガーに関するページばかりで，動物に関するページがほとんど含まれなかったためであると考えられる．他の例を見てみると，再ランキングのみを用いた (2) の手法よりも重み付けに再ランキングを用いて質問修正を行う (3) の手法の方が良い結果となっている．しかし実験例が少ないため，他の例においても今後比較を行わなければならないと考えている．また現在は質問修正や再ランキングを行い，閲覧文書に関連したページ群を集めるという段階までの処理を行っているが，今後は集めてきたページの中からユーザが必要としているページを提示する手法についても考えていきたい．

4.　検索語の自動検出によるクエリフリー検索

4.1　概　　要

ユーザが閲覧文書中のキーワードを初期検索語として能動的に選択した上で，質問修正を行う手法について述べてきた．この手法により，ユーザが文書閲覧中に何か知りたいことある場合に Web 検索を行う際の支援ができる．しかし，検索を行いたいと能動的に示さない場合にも，関連する Web ページを自動的に提示することで，ユーザの文書閲覧支援が行えると考えられる．本章では自動的にキーワードを抽出し，検索語を作成するクエリフリー検索を行う手法について述べる．ユーザは文書を読み進めいるだけで，現在の文書位置に関連した Web ページが自動的に次々と提示され，もし関心がある情報や必要なページであると判断すると即座にそのページを閲覧することができる．文書閲覧中のユーザは文書全体の中でも，現在閲覧している一部分について興味を持っていると考えられる．そこでまず閲覧している一部分を特徴付けるキーワードを発見する．システムの実装の際には，現在の閲覧箇所を正確に特定することはできないため，MS Word においてカーソルの位置を閲覧箇所と近似することにする．しかし，閲覧箇所を特徴付ける語の決定ができたとして，そのキーワードのみを用いて検索を行う場合，その語の対象範囲が広すぎて意味のないものとなってしまう可能性がある．例えば京都の文書を閲覧しているとする．その文書には京都の観光や食べ物，文化などいくつかのトピックについて書かれている．現在はその文書の中でも食べ物に関する部分を読んでいるとすると，その部分の特徴として “食べ物” や “飲食” といった語が選択された場合，一部分を特徴づける語としては正しいがそのまま検索語として利用したのでは飲食に関する余りに広範囲のページが得られることとなり，ユーザの閲覧支援ができているとは言いがたい．そこで一部分の特徴語だけでなく，文書全体での特徴語というものを決定し，現在閲覧している箇所の特徴語と全体の特徴語を組み合わせて検索に利用することで，より関連したページの検索を行うことができると考える．

4.2　検索語の決定

4.2.1　現在の文書位置を特徴づけるキーワードの検出

最初の段階として，閲覧中の文書の中で，現在注目している一部分を特徴付けるキーワードの抽出を行う．現在の文書位置での特徴語を決定するために，文書中の各文において，その文中に出現するキーワードがその文を特徴づける度合いを表す重みの計算を行う．

（1） 文単位での特徴ベクトルの作成

文書全体を文単位に区切り，その各 1 文での名詞の頻度による特徴ベクトルを作成する．i 番目に出現する文の特徴ベクトルを $TF_{s_i}(i = 1, ..., N)$ と書く．

（2） 特徴ベクトルの変形

通常文書を読み進めていく際に，前にある文の内容を覚えたままで次の文へ読み進んでいく．このように文書において文が並んでいる時，ある文と次の文の間には関連がある．そこで，ある文でキーワードが出現すると，次の数文にもそのキーワードの影響が残るように特徴ベクトルを変形する．変形後の特徴ベクトルを $TF'_{s_i}(i = 1, .., N)$ を (5) 式のように定める．

$$TF'_{s_i} = \sum_{l=max(0,i-range)}^{i-1} TF_{s_l} * (1 - \frac{i-l-1}{range}) \qquad (5)$$

$range$ はあるキーワードが出現して何文後まで影響を及ぼすかの範囲を表す．$(1 - \frac{i-l-1}{range})$ の部分は，例えば影響範囲を 5 文とすると，あるキーワードが登場して次の文には大きさ 1 で影響が残り，2 文目では $\frac{1}{2}$，3 文目では $\frac{1}{3}$ と後になればなるほど影響が弱くなるように影響が残る関数となっている．この影響の及ぼす様子を図 4 に示す．この変形によりある文にキーワードが現れると，後続の数文にも現れたとみなしている．

（3） 重みベクトルの計算

その文における頻度だけでなく，その文章の中で局所的に頻出しているキーワードがその部分を特徴づける語であると考えられるため，局所性の度合いも考慮する．あるキーワード K に対して，そのキーワードが出現する文の数を文頻度 SF_K として求める．次に求めた SF_K を文書に含まれる文の総数 N で正規化し，(6) 式の ISF_K を得る．

$$ISF_K = \log \frac{N}{SF_K} \qquad (6)$$

文書に出てくる全ての名詞について ISF を求め ISF ベクトルを作成する．出現する文の数が少なければ SF の値は低くなるため ISF の値は高くなる．以上より i 番目の文に出現する名詞 n_j の重み $w_{i,j}$ を (7) 式のように定義する．

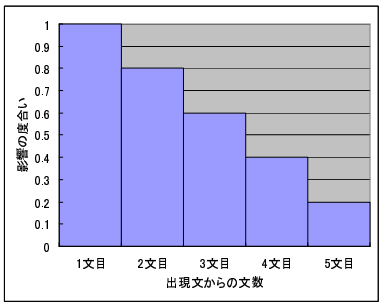

図 4　単語出現による影響の伝播

$$w_{i,j} = TF'_{s_i} * ISF_j \tag{7}$$

この重みの計算方法は 1 文を 1 つの文章としてみなし，閲覧文書を文書集合とみなした場合の $TF - IDF$ 法による重みの計算方法と同じである．

以上によりキーワードの重み付けをすることで，文書中でのある 1 文を特徴づけるキーワードの選択を行う．

4.2.2　文書全体を特徴づけるキーワードの検出

ある 1 文を特徴付ける語を選択する際には，局所的に頻出するキーワードが部分を特徴付ける語としてふさわしいと考えた．文書全体を特徴づける語を選択する際には，局所的に出現する語よりも全体的に出現する語の方がふさわしいと考えられる。そこで SF_K の値をキーワードを文書全体の特徴語の選択基準とし，この値の高いキーワードが文書全体を特徴付けている語と判断する．

4.2.3　検　　索

こうして全体の特徴語と部分の特徴語が抽出される．これらのキーワードを “AND” で組み合わせて検索語を作成し，Web 検索を行い，ユーザに結果のページを提示することになる．

4.3　実　行　例

ここで本提案手法により，部分の特徴語，及び全体の特徴語としてどのような語が抽出されるかの実験を行う．閲覧文書としては京都の烏丸通り付近の町並みについて書かれた文書を用いる．その文書の中では，飲み屋や飲食店，自転車問題など複数の話題が含まれている．その文書の中で文書全体，及び一部分の特徴語としてどのような語が選ばれるかを見る．

最初に一部分での特徴語を見る．この文書中のある 5 文について，特徴語を各文につき 5 語抽出し，抽出される各文の特徴語とその重みの値を表 3 に示す．この 5 文では話題として「河原町周辺に若者向けの衣料店が開店」,「昭和初期の建築物のブーム」,「百貨店と自転車の駐輪問題」等が述べられている．次に文書全体における特徴語について見る．文書全体に含まれるキーワードのうち SF の値の高い上位 5 語とその値を表 4 に示す．このように抽出された語を元に，例えば 2 文目を閲覧している際は部分から “ショップ” や “カフェ”，全体から “京都” や “烏丸” を選んで検索語を作ることで自動的に検索が行うことができる．検索語としては “ショップ AND 京都” や “(ショップ OR カフェ) AND (烏丸 OR 烏丸)” 等の検索語が作成できる．こうして自動的に作成された検索語で検索を行うことで，ユーザは京都のショップや烏丸のカフェについての情報を自動的に得ることができる．

文 1	キーワード	若者	河原町	寺町	前半	誕生
	w	3.84	2.70	2.70	2.70	2.70
文 2	キーワード	ショップ	カフェ	衣料	途中	向け
	w	2.70	2.70	2.70	2.70	2.70
文 3	キーワード	昭和	初期	ブーム	建築	建物
	w	2.70	2.70	2.70	2.70	2.50
文 4	キーワード	流れ	自然	散策	藤田	浮上
	w	3.84	2.70	2.70	2.21	1.35
文 5	キーワード	駐輪場	百貨店	立地	移動	不法
	w	2.70	2.70	2.56	2.56	1.93

表 3　文書の一部分から抽出される特徴語

キーワード	烏丸	京都	オープン	条	四条
SF	32	14	12	11	9

表 4　文書全体から抽出される特徴語

5.　まとめ及び今後の課題

本論文では，文書を閲覧している際の Web 検索支援を行う手法について述べた．まずユーザが文中のある語を初期検索語として選択した際に，その周辺からキーワードを抽出し，そのキーワードの中から初期検索語での検索結果を用いて，初期検索語を最も特徴付けると思われる語を選択し，検索語に追加する手法について述べた．次にユーザが初期検索語を入力しない場合において，閲覧部分の特徴語と文書全体での特徴語を検出することで自動的に検索語を作成する手法について述べた．

今後の課題としては，3.5 章の実験の際の動物のジャガーの例のように，現在の手法では検索結果の中から閲覧文書と同じ文脈の文書が集められず，その結果検索語を特徴付けるキーワードの重みが適切に計算できない場合の問題の解決を図りたいと考えている．また 4. 章の手法で一部分を特徴付ける語及び全体を特徴づける語の抽出を行った．現在は全体の特徴語と部分の特徴語での “AND” 検索を行うことを考えているが，この 2 つを用いて検索語を作成する手法は他にも考えられる．文書の一部分に関する文書が欲しいのか，文書全体に関する文書が欲しいのかに応じて，特徴語の組み合わせ方法を変えて検索語を作成する等，より状況を反映させた検索語を作成する手法についても考えていきたい．また，現在まだクエリフリー検索において検索語を作成し検索を行った結果の評価が行えていないため，今後実験を行い，検索語の評価を行いたいと考えている．

謝　　辞

本研究の一部は，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」、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 17 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，平成 17 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] 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).

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

[3] 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).

[4] 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).

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

[6] 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).

[7] Google
`http://www.google.com/`.