質問キーワードの意味的関連と近接性に着目したウェブ検索の精度改善

田　馳† 手塚　太郎‡ 小山　聡‡ 田島　敬史‡ 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: †tianchi@dl.kuis.kyoto-u.ac.jp, ‡{tezuka,oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし ウェブ検索エンジンに入力される質問キーワードの数は，一つあるいは二つである場合が高い割合を占める．既存のウェブ検索エンジンは質問キーワードの数が一つの場合に高い適合率を達成しているが，二つ以上の場合，適合率は低下していく場合がある．本稿では質問キーワードの数が二つの場合を対象に，質問キーワード間の意味的関連を分類した上で，文書内における質問キーワードの近接性を用いて，ウェブ検索結果を改善する手法を提案する．また，近接を表す三つの指標として，初出単語距離・最小単語距離・局所的出現密度を定義した．さらに，各指標に対する重み付けをクライアント側で調整し，必要に応じてリランキングの結果を動的に変更できるユーザインタフェースを実装した．このシステムによって，ウェブ検索結果の適合率が改善されることを評価実験によって示した．

キーワード Web とインターネット，情報検索，ユーザインタフェース

Improving Web Retrieval Precision based on Semantic Relationships and Proximity of Query Keywords

Chi TIAN† Taro TEZUKA‡ Satoshi OYAMA‡ Keishi TAJIMA‡ Katsumi TANAKA‡

†Department of Informatics, Faculty of Engineering, Kyoto University,

Yoshida-honmachi, Sakyo-ku, Kyoto,606-8501 Japan
‡Department of Social Informatics, Graduate School of Informatics, Kyoto University,

Yoshida-honmachi, Sakyo-ku, Kyoto,606-8501 Japan
E-mail: †tianchi@dl.kuis.kyoto-u.ac.jp, ‡{tezuka, oyama, tajima, tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract According to recent studies, one or two keywords are the most common queries in Web search. While most Web search engines perform very well for a single-keyword query, their precisions are not as good in case of two or more keywords. In this paper, we propose a method to improve Web retrieval precision based on semantic relationships and proximity of keywords for two-keyword queries. We define *the first appearance term distance, the minimum term distance* and *the local appearance density* as three measures for proximity. In addition, the user interface of the system enables the user to dynamically change the weight which is put on the original rank and our proposed measures. The experiment showed that our proposed method improves the precision of the Web search results.

Key words Web and Internet, Information retrieval, User interface.

1. はじめに

近年の調査によれば，ウェブ検索エンジンに入力される質問キーワードの数は，一つあるいは二つである場合が高い割合を占める[1][2]．一方，既存のウェブ検索エンジンの多くは質問キーワードの数が一つの場合に高い適合率を持つが，二つ以上の場合，関連の薄い

ページが多数取得され,適合率が低下する場合がある.その理由として，以下の要因が挙げられる.

1.一般に，ウェブページは異なる主題を述べる複数の部分から構成されることが多く，質問キーワードが異なった部分に含まれる場合は検索結果として適切でないことが多く，適合率が低下する.

2.検索エンジンのランキングアルゴリズムにおいて，ウェブページ内の主題構造が考慮されていることは少ない．例として，代表的な検索エンジンであるGoogle[3]で用いられているPageRankは被リンク数の多寡に基づいており，ページ内の主題構造は踏まえられていない.

例として，二つの質問キーワード「四条」「中華料理」を用いて検索を行った場合，四条という場所に存在する中華料理店のページだけでなく，四条に存在する中華料理以外の店の情報と，四条以外の場所に存在する中華料理店の情報が共に載ったページが検索結果上位に含まれてしまう．このようなノイズを除去するための一つの方法は，質問キーワード間の近接性に着目することである．また，質問キーワードの意味的関連を利用してフィルタリングを行う手法が考えられる.

本研究では，質問キーワードの数が二つの場合を対象に,質問キーワード間の意味的関連を分類した上で,文書内における質問キーワードの近接性を用いて，検索結果を改善する手法を提案する．また，近接性を表す指標として，1) 質問キーワード間の初出単語距離，2)質問キーワード間の最小単語距離，3) 質問キーワードの局所的出現密度を定義した．開発されたシステムはこれらの指標を用いて，検索エンジンの結果をリランキングする．さらに，各指標に対する重み付けをクライアント側で動的に調整し，ウェブ検索エンジンによる本来のランキングと結合することで，ユーザの要求に応じてランキングの変更を可能にさせる．システムの概要を，図 1 に示した.

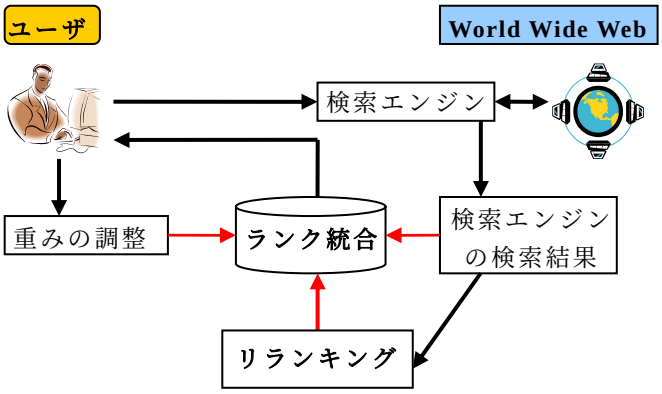


図 1　システムの概要

以降，本論文の第 2 章では，関連研究を述べる．第 3 章では質問キーワードの意味的関連を分類する.第 4 章では提案手法の概要とリランキングメカニズムについて説明する．第 5 章では実装について述べる．第 6 章では評価実験の結果を示し，第 7 章で全体のまとめを行う.

2. 関連研究

複数の質問キーワードによるウェブ検索を改善させる手法に関しては，いくつかの関連研究が存在する．本章ではそれらの研究のうち，本研究で提案される手法に関連するものについて述べる.

2.1 質問キーワードの単語距離の利用

複数の質問キーワードを用いて検索を行う際，文書内における質問キーワード間の単語距離を指定させる手法は,情報検索の分野では広く用いられている[4][5].

Callan は，文書全体よりもその限定的な領域に対して情報検索を行うことで，適合率を上げられることを示した[6]．複数の質問キーワードの近接の度合いに基づく文書検索を実現させる手法として,Sadakane らは,検索クエリとして与えられた多数の質問キーワードが一定の範囲内でまとまって現れる文書を抽出する高速アルゴリズムを実現している[7]．ウェブ検索の精度改善に関しては，単語距離だけではなく，局所的出現密度を用いる点が，本研究が異なっている点である．また，単語距離を用いた近接演算は多くの全文検索システムにおいて実現されているが，ウェブ検索エンジンでそれを明示的に指定できるものは少ない.

2.2 質問キーワードの密度分布の利用

文書内の質問キーワードの密度を検索結果のランキングに反映させる手法は，ウェブ検索エンジンを含む情報検索一般において広く用いられている[1].

黒橋らは，語の出現密度分布に基づき，その語に対する重要説明箇所を取得する手法を提案した[8]．佐野らはウェブ文書の内部構造をキーワードの出現密度分布を用いて抽出し，スコアリングを行う手法を示した[9]．中谷らは，頻出単語の出現密度分布を用いてウェブ文書を意味単位に分割する手法を提案している[10].本研究では,質問キーワードの出現密度分布ではなく,質問キーワードの局所的出現密度を用いる点が異なっている.

2.3 質問キーワードの役割の利用

通常，検索クエリとして用いられた複数の質問キーワードの間には，主題およびそれを修飾する語といったように，非対称な関係が成り立つ場合が多い．そこで，小山らは質問の階層的構造化を用いたウェブ検索手法を提案している[11]．この研究ではユーザの質問

中の主題的なキーワードと付加的なキーワードを区別し，ウェブページのタイトルと本文のそれぞれにマッチさせることで，検索精度を向上させている．本研究では，質問キーワードの役割に関して，質問キーワードの構造ではなく，質問キーワードの意味的関連に着目する点が異なっている．

3. 複数キーワードの意味的関連

本章では，質問キーワード間の意味的関連について述べる．まずは，質問キーワードの数が二つの場合について述べ，次に質問キーワードの数が三つ以上の場合について述べる．

3.1 質問キーワードの数が二つの場合

質問キーワードの数が二つの場合，両者の意味的関連のうち，もっとも代表的なものは以下の二種である．ここで，A，B はユーザによって入力された質問キーワードを表す．

3.1.1 主題修飾型

主題修飾型の場合，一方の質問キーワードは一つの主題を表し，もう一方は主題を表す質問キーワードを修飾している．二つの質問キーワードは従属関係にある．例として，質問キーワード「四条　中華料理」,「アイルランド　歴史」などがこの意味的関連に属している．この場合，二つの質問キーワードを「A の B」という形で繋げられることを意味する．

ユーザが特定の主題を絞り込む形で検索を行いたい場合，このタイプの質問キーワードが用いられる．

3.1.2 主題並置型

主題並置型の場合，二つの質問キーワードはそれぞれ異なる主題を表し，両者は並列関係にある．例えば，質問キーワード「結婚年齢　子供の数」,「就職率　景気」などがこの意味的関連に属している．この場合，二つの質問キーワードを「A と B」という形で繋げられることを意味する．

ユーザが二つの主題の関連について調べたい場合，このタイプの質問キーワードが用いられる．

3.2 質問キーワードが三つ以上の場合

質問キーワードが三つ以上の場合，質問キーワード間の意味的関連は主題修飾型と主題並置型が再帰的に混在した形になっていることが多いと考えられる．例として，質問キーワードが「福井　原子力発電　岩手　風力発電」の場合は，図 2 で示すような意味的関連に分解できる．すなわち，この質問例は「“福井の原子力発電”と“岩手の風力発電”」という形に解釈できる．ただし，本研究では，質問キーワードの数が三つ以上の場合が扱われていない．

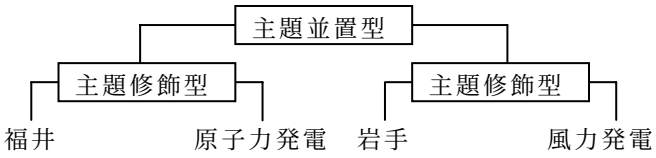


図 2 キーワードの意味的関連

4. 提案手法

本章では，提案手法の概要，ならびに使用される指標について述べる．図 3 に提案手法の概要を示す．

4.1 提案手法の概要

提案手法の一連の流れは，以下に示すものである．

(1) 質問キーワードをウェブ検索エンジンに送り，検索結果を取得してくる．
(2) 検索結果に対して形態素解析を行い，質問キーワード間の初出単語距離・最小単語距離・局所的出現密度を求める．
(3) 各検索結果に対して，(2)で求めた初出単語距離・最小単語距離・局所的出現密度に基づき，新しいランクを算出する．
(4) 検索エンジンのランクと(3)で求めたランクに重みを付け，統合する．
(5) (4)で求めた統合ランクに基づき，検索結果をリランキングする．
(6) 統合結果を表示する．

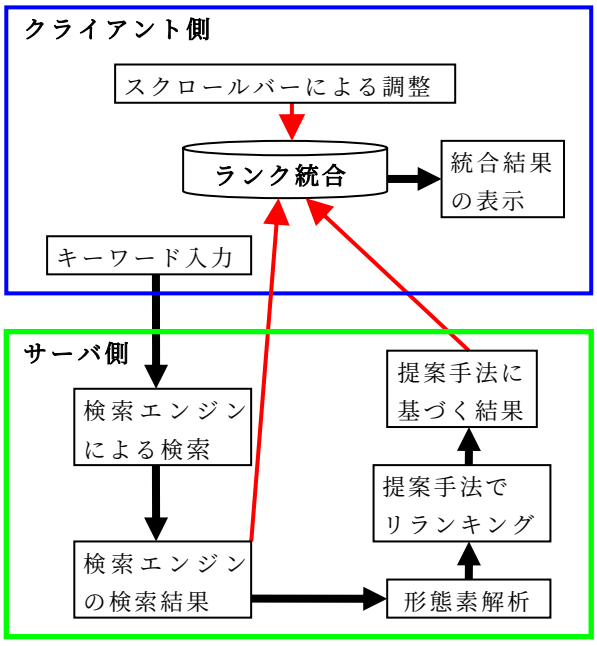


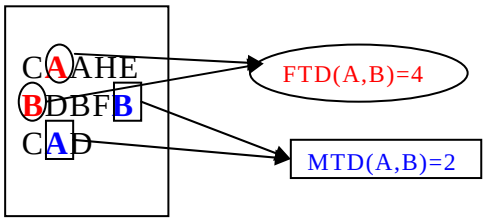
図 3 提案手法の概要

4.2 単語距離(TD)

単語距離(Term Distance)は，文書内部において，一つの単語からもう一つの単語までの単語数を意味する．

本研究では，質問キーワード間の単語距離を解析する際には，以下に述べる二種の単語距離を使用する．なお，本章で使われる関数は以下のように定義される．ここで，質問キーワードを A，B と置いた．

- $TD(A,B)$: AとBの単語距離
- $first(A)$: 文書の中に最初に現れたA
- $last(A)$: 文書の中に最後に現れたA
- $first(A,B)$: AとBのうち、先に現れたもの
- $last(A,B)$: AとBのうち、後に現れたもの
- $f_{\{M,N\}}(A,B)$: 範囲{M,N}のうち、AとBが現れた総数



文書 I

図 4 初出単語距離と最小単語距離

4.2.1 初出単語距離(FTD)

$$FTD(A,B) = TD(first(A), first(B)) \qquad (1)$$

初出単語距離(First-appearance Term Distance)は，文書の中で最初に現れた A と B の間の単語距離を表す．例えば，図 4 の文書 I における A と B の初出単語距離は 4 となる．

初出単語距離の使用は，重要な単語は文書の初期に現れる傾向が高いという仮説に基づくことである．すなわち，二つの質問キーワードが共に主題である場合，いずれも文書の先頭部分に現れると推測する．

4.2.2 最小単語距離(MTD)

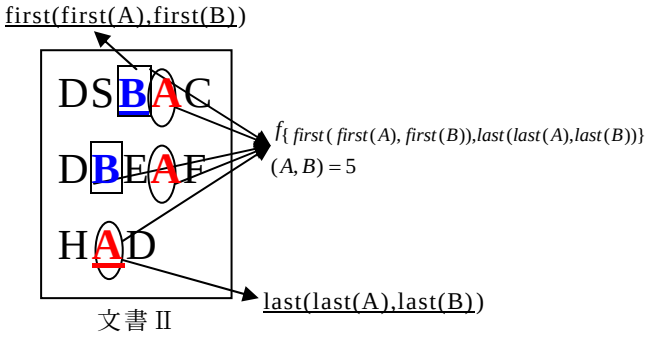
$$MTD(A,B) = \min(\{TD(A,B)\}) \qquad (2)$$

最小単語距離(Minimum Term Distance)は，文書の中で現れたあらゆる A と B の単語距離の中で最小のものである．例えば，図 4 の文書 I における A と B の最小単語距離は 2 となる．

最小単語距離の使用は，関連する単語同士は近接して現れるという仮説に基づくことである．

4.3 局所的出現密度(LAD)

質問キーワードの局所的出現密度(Local Appearance Density)は，式(3)のように定義する．ここで，質問キーワードを A，B と置いた．



文書 II

$$LAD(A,B) = \frac{f_{\{first(first(A),first(B)),last(last(A),last(B))\}}(A,B)}{TD(first(first(A),first(B)),last(last(A),last(B)))+1} = 5/(9+1) = 0.5$$

図 5 局所的出現密度

$$LAD(A,B) = \frac{f_{\{first(first(A),first(B)),last(last(A),last(B))\}}(A,B)}{TD(first(first(A),first(B)),last(last(A),last(B)))+1} \qquad (3)$$

質問キーワードの局所的出現密度は，最初に現れた質問キーワード（A 或いは B）から最後に現れた質問キーワード（A 或いは B）までの範囲内の二つの質問キーワード A と B の総数の割合と定義した．例として，図 5 の文書 II における A と B の局所的出現密度は 0.5 となる．

局所的出現密度の使用は，重要な単語は文書内で繰り返し現れるという仮説に基づく．

4.4 リランキング手法

4.4.1 初出単語距離法

質問キーワード間の初出単語距離を解析した結果を用いてリランキングを行う手法である．初出単語距離の大きさによってソーティングを行い，小さいものから順にランク順位を与えていく．初出単語距離が小さいものほど高いランクを得る．

4.4.2 最小単語距離法

質問キーワード間の最小単語距離を解析した結果を用いてリランキングを行う手法である．最小単語距離の大きさによってソーティングを行い，小さいものから順にランク順位を与えていく．最小単語距離が小さいものほど高いランクを得る．

4.4.3 局所的出現密度法

質問キーワードの局所的出現密度を解析した結果を用いてリランキングを行う手法である．局所的出現密度の高さによってソーティングを行い，大きいものから順にランク順位を与えていく．局所的出現密度が高いものほど高いランクを得る．

4.5 ランキング結果の統合

ユーザインタフェース上で実行可能な操作として，検索エンジンのランキングと提案手法によるランキングの動的な統合を提案する．統合値 Z は式(4)で定義され，Z の昇順に従って統合ランクを付ける．ここで，X は検索エンジンの検索結果におけるランクであり，Y は提案手法によるランクである．$S(S \in [0,1])$ は Y の重みであり，(1-S)は X の重みである．

$$Z = (1-S)X + SY \tag{4}$$

$$\Rightarrow \begin{cases} S=0 \Rightarrow Z = X \\ S=1 \Rightarrow Z = Y \end{cases} \tag{5}$$

式(5)において示したように，S=0 の時，統合値 Z は検索エンジンにおけるランクと等しいことで，統合ランクは検索エンジンにおけるランクと等しい．同様に，S=1 の時，統合ランクは提案手法におけるランクと等しい．ユーザはスクロールバー等で S の値を変更することにより，ランキング結果を変更することができる．検索エンジンによるランクを重視する場合は S の値を 0 に近付け，提案手法によるランクを重視する場合は S の値を 1 に近付ければよい．

4.6 提案手法のまとめ

本研究では，質問キーワードの数が二つの場合を対象に，テキスト内における異なる質問キーワード間の初出単語距離，最小単語距離，ならびにその局所的出現密度を用いて，検索エンジンでの検索結果を改善する手法を述べた．さらに，ユーザがクライアント側で提案手法によるランキングの重みを変更することにより，提案手法と検索エンジンを統合する．

5. 実装

5.1 実装環境

提案手法の有効性を示すため，システムのプロトタイプを実装した．以下に環境を示す．

- OS：Microsoft Windows XP Professional Version
- CPU: Intel Pentium M 1.20GHz, RAM: 1.00GB
- 開発環境：Microsoft Visual Studio.NET(C#)

5.2 モジュール

5.2.1 形態素解析

茶筌[12]は奈良先端科学技術大学院大学自然言語処理講座にて開発された日本語形態素解析器である．設定ファイルの編集によって，品詞体系，単語認定基準等を容易に変更できるようになっている．本研究では，質問キーワード間の初出単語距離，最小単語距離，ならびに局所的出現密度を解析する際に用いている．

5.2.2 ウェブ検索

ウェブコンテンツの URL，スニペット（質問キーワードを含む文書中の一部分），およびテキスト部分の取得には，SlothLib[13]を使用した．SlothLib は Google を用いたウェブ検索，ならびにその検索結果に対する自然言語処理を行う C#のライブラリである．

5.3 システムの概観

本研究で提案された手法を用いて，ウェブ検索結果のリランキングを行う検索システム SSRP(Search by Semantic Relationship and Proximity)を実装した．ユーザが質問キーワードを入力すると，それをクエリとして検索エンジンに送り，その上位 20 件分のウェブページを取得し，それぞれに対して形態素解析を行う．さらに，質問キーワードの初出単語距離，最小単語距離，及び局所的出現密度を用いて，リランキングを行う．

図 6 SSRP の実行例

図 6 に示したように，システムは以下の四つの領域から構成されている．

Ⅰ.入力エリア　　Ⅱ.解析エリア

Ⅲ.要約エリア　　Ⅳ.表示エリア

ユーザは入力エリアに質問キーワードを入力し，スクロールバーで重み付けの値を設定し，検索を行う．各リランキング手法の解析結果は解析エリアに出力される．また，要約エリアでは検索結果のスニペットとコンテンツのテキスト部分が表示される．表示エリアでは，画像も含めたウェブページ全体が表示される．

5.4 システムの実行例

図 6 は，質問キーワード「四条　中華料理」で検索した状況を示す．入力エリアにおいては，質問キーワードを入力し，スクロールバーの値を 1 に設定してから，検索を行う．本システムでは，スクロールバーの値を 0 から 4 までの五つのレベルを用意した．それぞれ，以下のように示す基準に基づいて統合ランクを求めている．

レベル 0 (S=0)：検索エンジンのランキングのみを利用．

レベル 1 (S=0.25)：検索エンジンでの検索結果を重視．

レベル 2 (S=0.5)：検索エンジンと提案手法での結果を同等に重視．

レベル 3 (S=0.75)：提案手法での検索結果を重視．

レベル 4 (S=1)：提案手法によるランキング．

次に，システムがテキスト解析を行い，解析結果を解析エリアに出力する．解析エリアでは検索結果のタイトル，URL，スニペット，ウェブコンテンツ，初出単語距離・最小単語距離・局所的出現密度の値，ならびにそれらに基づくランク順位を表示している.また，解析エリアでは，ユーザが選らんだリランキング手法によって順序付けされた検索結果を表示している．さらに，ユーザが選んだ検索結果のスニペットとコンテンツのテキスト部分が要約エリアで表示され，画像も含めたウェブページ全体が表示エリアで表示される．

6. 評価実験

本章では，本研究によって提案された三つのリランキング手法を Google によるランキングと比較した評価実験を述べる．実験の概要及び評価尺度に続いて，結果と考察を述べる．

6.1 実験の概要

質問キーワード間の意味的関連ごとに，主題修飾型，主題並置型に対してそれぞれ 20 組，10 組の質問キーワードで比較実験を行った．ただし，Google からの検索結果の取得件数は 20 件に設定する．

評価実験では，Google で得られる上位 20 件の検索結果に対して，リランキング手法を適用する．また，リランキングした検索結果と Google の検索結果での適合率で各手法の優劣を比較します．実験で使われた質問キーワードをキーワード間の意味的関連別で表 1 と表 2 に示した．

主題修飾型		
Q1	京都	名所
Q2	京都大学	情報学科
Q3	ケーキ	作り方
Q4	Java	語源
Q5	日本	歴史
Q6	ロボット	歴史
Q7	四条	中華料理
Q8	木屋町	和食
Q9	寺町	フランス料理
Q10	密度	定義
Q11~Q20：省略		

表 1　評価実験用質問キーワード（主題修飾型）

主題並置型		
Q1	食欲	季節
Q2	年齢	睡眠時間
Q3	就職率	景気
Q4	給料	学歴
Q5	身長	寿命
Q6	適合率	再現率
Q7	カロリー	ダイエット
Q8	結婚年齢	子供の数
Q9	記憶力	年齢
Q10	性格	血液型

表 2 評価実験用質問キーワード（主題並置型）

6.2 実験の評価尺度

評価は，適合率を用いて行った．適合率とは，検索結果における適合文書の割合を示す値である[14][15]. 検索結果の文書の総数を M，そのうちの適合する文書数を N とすると，適合率 P は式(6)で表される．

$$P = \frac{N}{M} \qquad (6)$$

本研究では,検索結果の適合･不適合の判定を行い，適合率を比較することで提案手法の優劣の判定基準としている．

6.3 実験結果

次に，表 1 と表 2 の質問キーワードを用いて本研究の初出単語距離法・最小単語距離法・局所的出現密度法と，Google での適合率について比較を行う．図 7 は表 1 の主題修飾型の 20 組の質問キーワードに対しての上位 K（K＝1,2,...,20）件検索結果の平均適合率の比較結果を示す．図 7 から求めた各リランキング手法が Google に対しての平均適合率の改善ポイント数を検索結果の件数ごとに表 3 で示した．図 8 は，表 2 の主題並置型の 10 組の質問キーワードに対しての上位

K(K=1,2,...,20)件検索結果の平均適合率の比較結果を示す．図 8 から求めた各リランキング手法が Google に対しての平均適合率の改善ポイント数を検索結果の件数ごとに表 4 で示した．

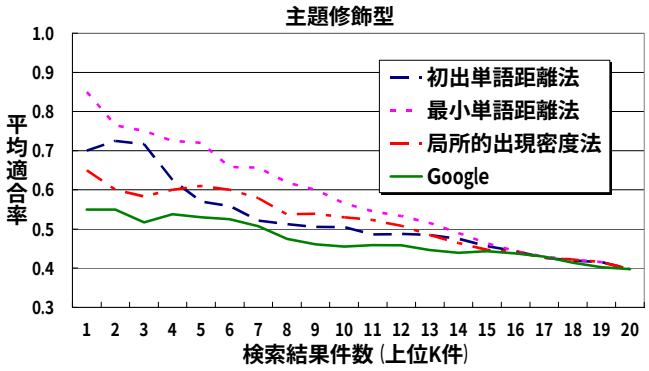


図 7 主題修飾型の比較結果

主題修飾型	5 件	10 件	15 件	平均
初出単語距離法	4.0	5.0	1.4	5.0
最小単語距離法	**19.0**	**11.0**	**2.0**	**10.7**
局所的出現密度法	8.0	7.5	0.3	4.6

表 3 平均適合率の改善ポイント数（主題修飾型）

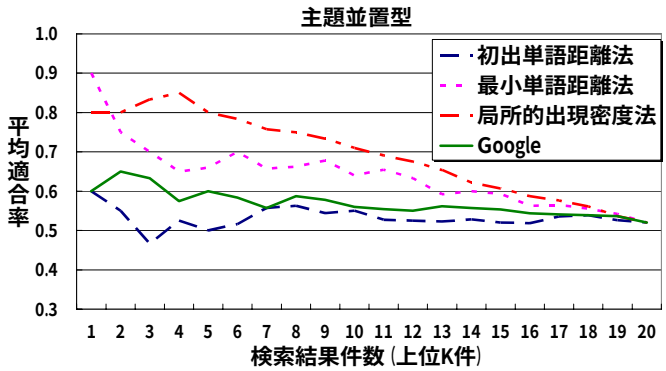


図 8 主題並置型の比較結果

主題並置型	5 件	10 件	15 件	平均
初出単語距離法	-10.0	-1.0	-3.3	-3.7
最小単語距離法	6.0	8.0	4.0	7.2
局所的出現密度法	**20.0**	**15.0**	**5.3**	**12.3**

表 4 平均適合率の改善ポイント数（主題並置型）

6.4 考察

6.4.1 主題修飾型の質問キーワードに対する考察

表 3 で示したように，主題修飾型の質問キーワードに対しては，最小単語距離法がもっとも優れている．最小単語距離法では上位 5 件，10 件，15 件の検索結果に対しての平均適合率の改善はそれぞれ 19.0，11.0，2.0 ポイントとなる．また，全体での平均適合率の改善の平均は 10.7 ポイントとなる．いずれも初出単語距離法，局所的出現密度法より平均適合率の改善が顕著である．

初出単語距離法と局所的出現密度法の場合では，それぞれ全体での平均適合率の改善の平均は 5.0 ポイント，4.6 ポイントとなり，大きな改善は見られないが，ある程度効果があると言える．

結論として，主題修飾型の質問キーワードに対しては，最小単語距離法・初出単語距離法・局所的出現密度法のいずれの手法とも適合率を改善する効果があった．特に，最小単語距離法の効果は最も顕著である．

6.4.2 主題並置型の質問キーワードに対する考察

表 4 で示したように，主題並置型の質問キーワードに対しては，局所的出現密度法が一番優れている．局所的出現密度法では上位 5 件，10 件，15 件の検索結果に対しての平均適合率の改善はそれぞれ 20.0，15.0，5.3 ポイントとなる．また，全体での平均適合率の改善の平均は 12.3 ポイントとなる．いずれも初出単語距離法，最小単語距離法より平均適合率の改善が顕著である．

最小単語距離法は全体での平均適合率の改善の平均は 7.2 ポイントとなり，局所的出現密度法に次ぐ結果を得ている．一方，初出単語距離法は全体での平均適合率の改善の平均は 3.7 ポイントの低下となり，Google による本来のランキングよりも劣っている．

結論として，主題並置型の質問キーワードに対しては局所的出現密度法と最小単語距離法が適合率を改善する効果があった．特に，局所的出現密度法は最小単語距離法よりも効果を持つことが示された．

結論	主題修飾型	主題並置型
初出単語距離法	○	—
最小単語距離法	◎	○
局所的出現密度法	○	◎

表 5 評価実験の結果

6.5 結論

本章では初出単語距離法・最小単語距離法・局所的出現密度法の三つのリランキング手法を評価するため，主題修飾型及び主題並置型の質問キーワードをそれぞれ 20 組，10 組を用い，Google によるランキングとの比較実験を行った．結果として，表 5 で示したように，主題修飾型の質問キーワードに対しては最小単語距離法，初出単語距離法及び局所的出現密度法が共に効果があること，特に最小単語距離法はもっとも効果が大きいことを示した．一方，主題並置型の質問キーワードに対しては局所的出現密度法と最小単語距離法が共に効果があること，特に局所的出現密度法はもっとも効果が大きいことを示した．

本研究に基づくリランキング手法は多様な質問キーワードに対して適用できると考えられるが，効果の現れない質問キーワードも存在すると考えられる．例として，主題修飾型の質問キーワードに対して最小単語距離法を使用した場合,適合率が Google より低くなる場合が見られる．このような場合には，ユーザはインタフェース上でランキング結合の重み付けを変更することにより, Google 本来のランキングで検索結果を表示させることができる．

7. まとめと今後の課題

本研究では，ウェブ検索において質問キーワードの数が二つである場合を対象に，質問キーワード間の意味的関連を分類した上で，文書内における質問キーワードの近接性を用いて，検索結果を改善する手法を提案した．

また，複数の指標に対する重み付けをクライアント側で調整し，ウェブ検索エンジンによる本来のランキングと結合することで，ユーザの要求に応じてリランキングの結果を動的に変更できるユーザインタフェースを実装した．さらに，評価実験の結果として，主題修飾型の質問キーワードに対して，初出単語距離法，最小単語距離法及び局所的出現密度法は効果があること，特に最小単語距離法はもっとも効果があること，主題並置型の質問キーワードに対しては局所的出現密度法と最小単語距離法は効果があること，特に局所的出現密度法がもっとも効果があることを示した．

今後の課題としては，質問キーワードの数が三つ以上の場合への対応が挙げられる．また，主題修飾型の質問キーワードは，さらにいくつかのグループに分けることが可能である．評価実験の結果より，地域名と店舗，組織と人名，などにおいては本提案手法が特に有効であることが推測される．それぞれのグループに関して適合率がどの程度向上するかを調べることで，本手法がどのような検索クエリに対して有効であるのかを詳細に示すことができる．さらに，システムがユーザの入力した質問キーワードの暗黙的な意味的関連を判定し，もっとも効果的なリランキングを行わせることがこれからの課題である．

謝辞

本研究は，一部，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」による．また，本研究は一部，文部科学省科学技術振興費知的資産プロジェクト「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）による．ここに記して謝意を表します．

文献

[1] B. J. Jansen, A. Spink, J. Bateman and T. Saracevic: “Real life information retrieval: A study of user queries on the web” ACM SIGIR Forum, Vol. 32, No. 1，pp. 5-17，1998.

[2] Trellian, http://www.trellian.com/

[3] Google, http://www.google.co.jp

[4] 山本毅雄，橋爪宏達，神門典子，清水美都子: 全文検索－技術と応用，丸善，1998.

[5] 佐良木昌，新田義彦: 正規表現とテキストマイニング，明石書店，2003.

[6] J. Callan: “Passage-level evidence in document retrieval,” Proceedings of the 17th Annual International ACM SIGIR Conference, pp. 302-309, 1994.

[7] K. Sadakane and H. Imai: “On k-word Proximity Search,” IPSJ SIG Notes 99-AL-68, 1999.

[8] 黒橋禎夫，白木伸征，長尾眞: “出現密度分布を用いた語の重要説明箇所の特定,” 情報処理学会論文誌, Vol. 38, No. 04, pp.845-854, 1997.

[9] 佐野綾一，松倉健志，波多野賢治，田中克己: “部分グラフを基本単位とした Web 文書検索：単語の出現密度分布の適用,” 情報処理学会研究報告, Vol.99, No.61, pp.79-84, 1999.

[10] 中谷圭吾，鈴木優，川越恭二: “文書間類似度とキーワードを用いた Web リンク自動生成手法,” 日本データベース学会 Letters, Vol. 4, No. 1, pp.89-92, 2005.

[11] 小山聡，田中克己: “j 質問の階層的構造化を用いた Web 検索手法の提案,” DBSJ Letters, Vol.1, No.1, pp.1-4, 2001.

[12] 形態素解析システム『茶筌』，
http://chasen.naist.jp/hiki/chasen

[13] ウェブテキスト解析システム『SlothLib』，
http://www.dl.kuis.kyoto-u.ac.jp/~ohshima/wiki/index.php?SlothLib

[14] C. J. van Rijsbergen: Information Retrieval (Second Edition), Butterworths, 1979.

[15] 徳永健伸: 情報検索と言語処理 (言語と計算 5), 東京大学出版会, 1999.