質問キーワードの近接性と密度分布に基づくウェブ検索の改善手法

Web Search Improvement Based on Proximity and Density Distribution of Multiple Keywords

田 馳† 手塚 太郎‡ 小山 聡‡
田島 敬史‡ 田中 克己‡

Chi TIAN Taro TEZUKA Satoshi OYAMA
Keishi TAJIMA Katsumi TANAKA

ウェブ検索エンジンに入力されるキーワード数は，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キーワード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傾向がある．本稿ではキーワード数が二つの場合を対象に，文書内における異なるキーワード間の距離，ならびにその密度分布を用いて，ウェブ検索結果のリランキングを行う手法を提案する．さらに，複数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．

One or two keywords are most commonly used as Web search queries. While most Web search engines perform very well for a single-keyword query, their precisions decrease as the number of keywords increases. In this paper, we propose a meta-search system that re-ranks Web search results based on the distances between keywords and the density of their distributions. The user interface of the system enables the user to dynamically change the weight put on the original rank and our proposed measures. The experiment showed that our method improves the precision of the Web search results.

1. はじめに

ウェブ検索エンジンに入力されるキーワード数は，一つあるいは二つである場合が高い割合を占める[1]．一方，既存のウェブ検索エンジンの多くはキーワード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，適合率が低下する傾向がある．

例として，二つのキーワード「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，複数キーワード間の出現場所の距離，ならびにその密度分布に着目することである．

本研究では，質問キーワードが二つの場合を対象に，文書内におけるキーワード間の距離，ならびにその密度分布を用いて，検索結果を改善する手法を提案する．開発されたシステムでは，1) キーワード間の初出距離，2) キーワード間の最小距離，3) キーワードの密度分布，の三つの指標を用いて，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，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．

2. 関連研究

2.1 質問キーワードの単語距離と密度分布の利用

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．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[2]．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[3]．

一方，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．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[4]．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[5]．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[6]．

本研究では，ウェブ検索の精度改善に関しては，単語距離と密度分布の両方を用いる点が異なっている．

2.2 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[7]．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

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 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．

† 学生会員 京都大学大学院情報学研究科博士前期課程
tianchi@dl.kuis.kyoto-u.ac.jp
‡ 正会員 京都大学大学院情報学研究科
{tezuka, oyama, tajima, tanaka}@dl.kuis.kyoto-u.ac.jp

3.2 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．

また，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．

4. 提案手法

4.1 提案手法の概要

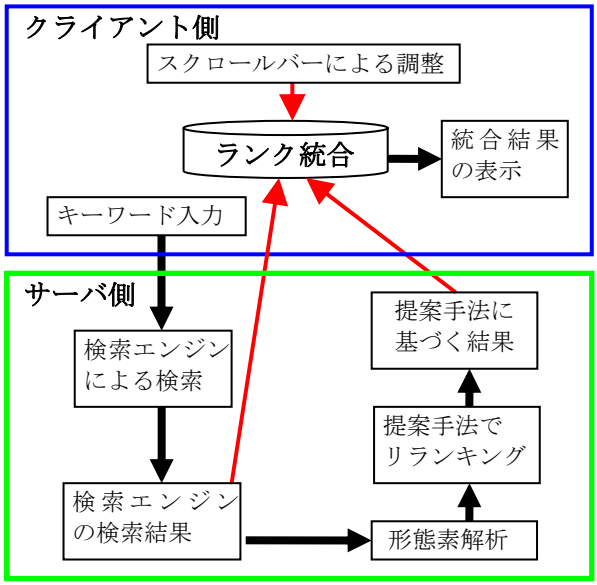


図 1　提案手法の概要
Fig.1　Flow of our method

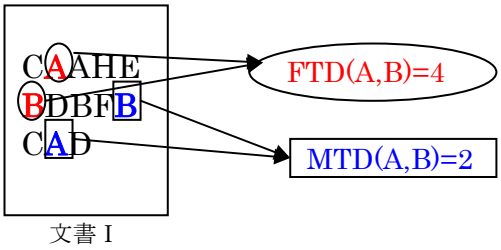
提案手法では，以下の順序でウェブコンテンツの検索が行われる．図 1 に提案手法の概要を示す．

(1) 検索クエリを検索エンジンに送り，結果を取得する．
(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単語距離，密度分布を求める．
(3) 各検索結果に対して，(2)で求めた単語距離，密度分布に基づき，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) 統合結果を表示する．

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．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 2　初出単語距離と最小単語距離
Fig.2　FTD and MTD

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．

4.2.1 初出単語距離(FTD)

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

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 2 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)

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

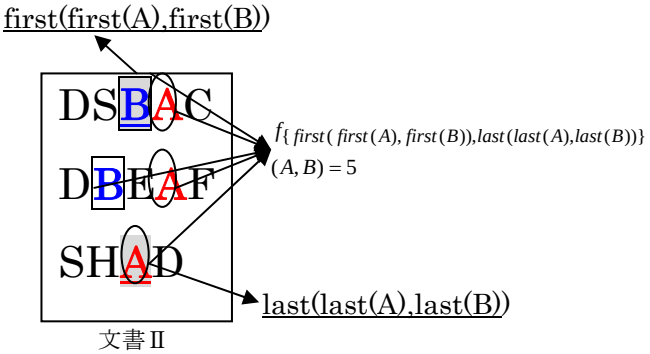
最小単語距離(Minimum Term Distance)は，文書の中で現れたあらゆる A と B の単語距離の中で最小のものである．例えば，図 2 の文書 I における A と B の最小単語距離は 2 となる．最小単語距離の使用は，関連する単語同士は近接して現れるという仮説に基づくことである．

4.3 単語の密度分布(DD)

キーワードの密度分布(Density Distribution)を式(3)のように定義する．ここで，キーワードを A，B と置いた．

$$DD(A,B) = \frac{f_{\{first(first(A),first(B)),last(last(A),last(B))\}}(A,B)}{TD(first(first(A),first(B)),last(last(A),last(B)))} \quad (3)$$

キーワードの密度分布は，最初に現れた質問キーワード(A 或いは B)と最後に現れた質問キーワード(A 或いは B)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する．例として，図 3 の文書 II でのキーワードペア A と B の密度分布は 0.5 となる．密度分布の使用は，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．



文書 II

$$\Rightarrow DD(A,B) = \frac{f_{\{first(first(A),first(B)),last(last(A),last(B))\}}(A,B)}{TD(first(first(A),first(B)),last(last(A),last(B)))} = 5/10 = 0.5$$

図 3　単語の密度分布　　Fig.3　DD

4.4 リランキング手法

4.4.1 初出距離法

キーワード間の初出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．初出距離が小さいものほど高いランクを得る．

4.4.2 最小距離法

キーワード間の最小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．最小距離が小さいものほど高いランクを得る．

4.4.3 密度分布法

質問キーワードの密度分布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．密度分布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．密度分布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リランキング手法におけ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 はリランキング手法によるランクである．$S(S \in [0,1])$ は Y の重みで，$(1-S)$ 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，統合ランクはリランキング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．

4.6 ユーザインタフェース

提案手法の有効性を示すため，システムのプロトタイプ SPDD(Search by Proximity and Density Distribution)を実装した．図 4 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

図 4 SPDD　　Fig.4 SPDD

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. 評価実験

5.1 実験結果

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に対して評価実験を行った．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．

図 5 と図 6 は，それぞれ主題修飾型の 20 組の質問キーワードと主題並置型の 10 組の質問キーワードに対しての平均適合率の比較結果を示す．表 1 と表 2 は，それぞれ主題修飾型と主題並置型の質問キーワードに対して，各リランキング手法における，Google に対しての平均適合率の改善ポイント数を示す．

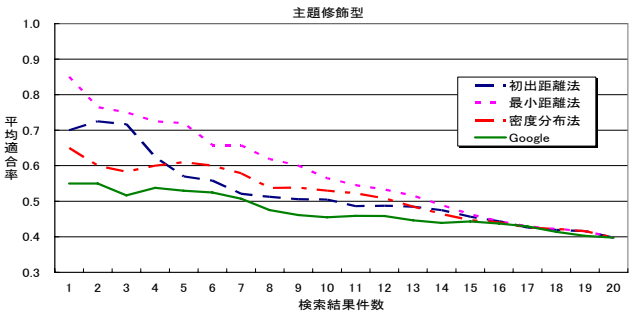


図 5 主題修飾型の比較結果
Fig.5 Results of subject modifying type

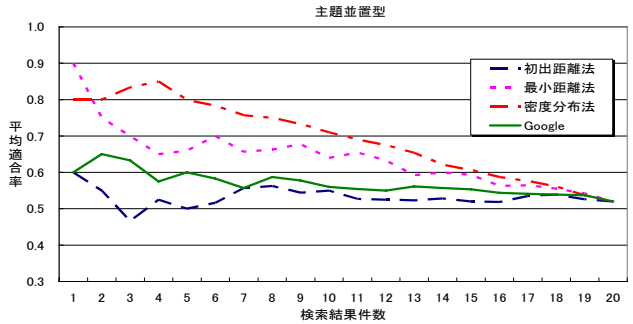


図 6 主題並置型の比較結果
Fig.6 Results of subject juxtaposing type

表 1 平均適合率の改善ポイント数（主題修飾型）
Table 1 Improvement in precision (subject modifying type)

主題修飾型	5 件	10 件	15 件	平均
初出距離法	4.0	5.0	1.4	5.0
最小距離法	**19.0**	**11.0**	**2.0**	**10.7**
密度分布法	8.0	7.5	0.3	4.6

表 2 平均適合率の改善ポイント数（主題並置型）
Table 2 Improvement in precision (subject juxtaposing type)

主題並置型	5 件	10 件	15 件	平均
初出距離法	-10.0	-1.0	-3.3	-3.7
最小距離法	6.0	8.0	4.0	7.2
密度分布法	**20.0**	**15.0**	**5.3**	**12.3**

5.2 考察

5.2.1 主題修飾型の質問キーワードに対する考察

表 1 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，最小距離法がもっとも優れている．最小距離法では上位 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．いずれも初出距離法，密度分布法より平均適合率の改善が顕著である．一方，初出距離法と密度分布法の場合では，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，最小距離法・初出距離法・密度分布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，最小距離法の効果は最も顕著である．

5.2.2 主題並置型の質問キーワードに対する考察

表 2 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，密度分布法が一番優れている．密度分布法では上位 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．いずれも初出距離法，最小距離法より平均適合率の改善が顕著である．

最小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，密度分布法に次ぐ結果を得ている．一方，初出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密度分布法と最小距離法が適合率を改善する効果があった．特に，密度分布法は最小距離法よりも効果を持つことが示された．

6. まとめと今後の課題

本研究では，ウェブ検索においてキーワード数が二つである場合を対象に，文書内における異なるキーワード間の距離，ならびにその密度分布を用いて，ウェブ検索結果のリランキングを行う手法を提案した．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，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．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，初出距離法，最小距離法及び密度分布法は効果があること，特に最小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密度分布法と最小距離法は効果があること，特に密度分布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，質問キーワード数が三つ以上の場合への対応が挙げられる．また，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] J.,Callan: “Passage-level evidence in document retrieval,” Proceedings of the 17th Annual International ACM SIGIR Conference, pp. 302-309, 1994.

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

[4] 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.

[5] 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.

[6] 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.

[7] 小山聡，田中克己: “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.

田 馳 Chi TIAN
京都大学大学院情報学研究科博士前期課程在学中．2006 年京都大学工学部情報学科卒業．複数キーワード検索の研究・開発に従事．日本データベース学会学生会員．

手塚 太郎 Taro TEZUKA
京都大学大学院情報学研究科社会情報学専攻助手.2005 年京都大学大学院情報学研究科博士後期課程修了.博士(情報学).主に地域情報検索システム，ウェブからの知識発見の研究に従事．情報処理学会，日本データベース学会各会員．

小山 聡 Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助手．2002年京都大学大学院情報学研究科博士後期課程修了.博士(情報学).主に機械学習，データマイニング，情報検索の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI各会員．

田島 敬史 Keishi TAJIMA
京都大学大学院情報学研究科社会情報学専攻助教授．1996年東京大学理学系研究科情報科学専攻博士後期課程修了.博士(理学)．主にデータベースプログラミング言語,Web 検索の研究に従事．IEEE Computer Society，ACM，情報処理学会，日本データベース学会等各会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976年京都大学大学院修士課程修了．博士（工学）．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．