E-038

文書構造を考慮した近接度スコアを用いた文書検索結果ランキング方式
A Ranking Method Using Proximity Scoring Based on a Document Structure

鈴木 克典† 湯川 高志† 戸田 浩之§ 数原 良彦‡ 片岡 良治‡
Katsunori Suzuki Takashi Yukawa Hiroyuki Toda Yoshihiko Suhara Ryoji Kataoka

1. はじめに

近年，ニュース記事・広告・個人の思想など，ありとあらゆる情報がインターネット上に溢れており，ユーザーがその中から必要な情報を得る手段として，Web 検索サービスが普及している．

Web 検索サービスとは，ユーザーが自分の知りたい事項を表すキーワードを問い合わせ (クエリ) として入力すると，クエリに基づいて Web 文書を検索・提示するものである．その際に，インデクスされた Web 文書からクエリに適合する文書を探し出すメカニズムは，検索エンジンと呼ばれる．与えられたクエリに基づき，検索エンジンは Web 上の各文書に対して，大きく分けて２つのタイプのスコア付けを行う．ひとつは，PageRank[1]や被リンク数・URL 長など，その文書固有のスコアであり，クエリに依存しないタイプである．もうひとつはクエリに依存するタイプのスコアで，これには TF-IDF[2]や BM25[3]に代表されるキーワード関連度，また文書内における，クエリ中キーワードの近接性に基づくスコア (近接度スコア)[5][6][7] などが挙げられる．検索エンジンが最終的に提示する検索結果は，上に挙げたような複数のスコアを元に文書スコアを算出し，それをランク付けしたものである．著者らはそれら各種スコア中でも，クエリ中のキーワード近接性に着目している．

ここでまず，著者らが着目する近接性について詳述する．近接性とは，与えられたクエリに含まれるキーワードがごく近くに共起する文書は，キーワード同士が単に含まれる文書よりも，ユーザーの要求する文書に近いだろうという考えに基づいた尺度である．例えば“横浜 ラーメン”というクエリがユーザーから与えられたとき，それぞれのキーワードの類似度を個別に評価するだけでは，“横浜のランドマークタワー”と“北海道のラーメン”について記述された文書も，検索結果に含まれてしまう可能性がある．そこで近接性という尺度を評価に導入することにより，このような文書はランキングを下げることができる．このように，近接性は文書検索において，ヒューリスティックな判断基準となる．

また一般的な文書の構造について着目すると，各文書は１つの表題を持ち，それに続いて章・節・本文の集合体として構成される．ここで，文書における表題や，章・節などの見出しは，それに後続する文の内容を端的に表現するものと考えられる．その表題や見出しをタイトルとする．それにより，そのタイトルに後続する文章に現れるキーワードは，タイトルに含まれるキーワードと関連性を持つと考えられる．この場合，文書中で見かけ上の単語間距離から判断する近接性と，キーワード同士の意味的な関連から判断する近接性は異なる可能性がある．

一方，Web 文書を作成するのに用いられている HTML 言語は，文書の論理構造を指定するためのタグと見栄えを指定するためのタグが混在している．しかし，見栄えに関する記述は css ファイルに記載し，HTML ファイルには文書の論理構造のみを記述するように，近年は W3C から推奨されている．すなわち HTML 文書は，文書の論理構造を記述したタグのみが含まれることになり，上記のような構造に基づく近接性評価を取り入れることが可能となる．よってそれを利用することにより，Web 検索の精度を向上させることが可能と考えられる．

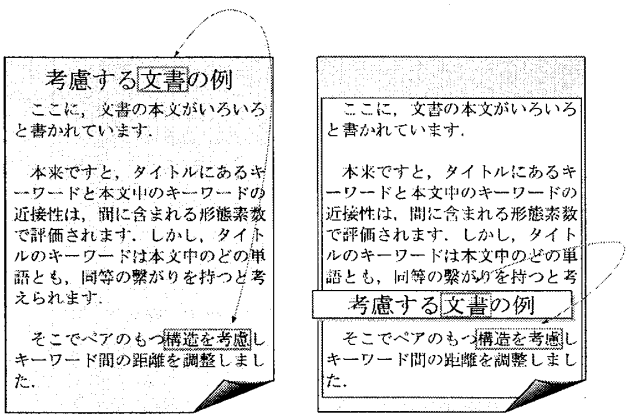

図 1 文書構造を考慮した単語間距離の算出

従来提案されている近接性評価は，キーワード間の形態素数のみを考慮するのが一般的であり，文書構造に起因する近接性の齟齬が解消されていない．そこで本稿では，Web 文書が HTML 言語で記述された明示的な構造を有する点に着目し，より有意味な近接性評価を提案するものである．特に著者らは，タイトルとその本文の構造に着目し，その物理的距離から算出されるスコアを調整する事で検索精度の向上を目指す (図 1)．またクエリの近接性について，上記の構造を考慮した近接度スコアの与え方に加え，その近接度スコアと既存のキーワード関連度との組み合わせ方について新たな提案を行う．

以下，2 章で関連研究について，3 章で文書構造や用語について述べる．次に 4 章で事前検証による構造を考慮した近接性の必要性について，5 章では提案手法について，6 章で評価と考察を述べ，7 章でまとめる．

2. 関連研究

Web 検索をはじめとする文書検索においては，適合文書をランキングのより上位に表示させることが重要となる．その適合率を上げるために，文書のスコアを算出す

† 長岡技術科学大学 Nagaoka University of Technology
‡ 日本電信電話株式会社 NTT サイバーソリューション研究所 NTT Cyber Solutions Laboratories, NTT Corporation
§ NTT コミュニケーションズ株式会社 NTT Communications Corporation

る要素が多く研究されているがその手法の 1 つとしてユーザーにより与えられた検索クエリ中のキーワードが，検索対象の文書内で持つ近接性を，文書スコアの 1 つの指標にすることが提案されている．すなわち，文書内におけるキーワードの近接性が高いほど，その文書により高得点を与えようという考えである．

Keen らはこの考えに基づき，文書検索用の演算子として NEAR を提案している[4]．NEAR 演算子で指定されたキーワード同士の，文書内における近接性を評価し文書スコアに反映させる試みである．さらに Rasolofo らは，キーワード関連度と，文書中のキーワードペアの距離及びそのキーワードの重要性を組み合わせたスコアを提案しまたそれを用いて Web 検索における近接性の有効性を示した[5]．また Tao らは，キーワードの近接性の種類 (キーワードペアの最小距離，平均距離，全てのペアを含む最小範囲などについて) や，近接性をスコアに換算する式について詳細に評価している[6]．その中で，キーワードペアの最小距離を指標として用いたときに，もっとも検索精度が高くなることが実験的に示されている．

以上の研究では文書の構造は考慮されていないが，近接性評価に文の構造を考慮しようという試みは Andreas らによっても提案されている[7]．この手法は，構造が異なる場合に存在するキーワードペアの距離を，実際のものより長く評価しようというものである．たとえば，HTML においてパラグラフは<p>タグで示されるが，もし 2 つのキーワードが異なるパラグラフに存在すれば，距離をより大きく評価する．筆者らの提案は，指定した構造を持つペアの距離を小さく評価する試みであり，この試みとは相補関係にあると捉えることができる．

3. 本稿で扱う文書構造，および用語の定義

テキスト検索分野において，近接性とは一般にキーワードペアの間にある形態素数を基準にして与えられる．ここで，間の形態素数を距離と呼び，間に含まれる形態素が多いほど，距離が長くなると表現する．

ここで，Web 文書の構造の識別法を提案する．Web 文書は<head>タグで指定される要素（head 要素）と<body>タグで指定される要素（body 要素）とから構成され，さらに head 要素の中には title 要素が記述される．ここから，<title></title>間を表題，<body></body>間を本文と考えたとき，**タイトルとその本文**という構造を得ることができる．

また body 要素内には，heading タグと呼ばれる章題を与えるタグが書かれる．W3C では heading タグは<h1>から<h6>までを制定している．ここでは，h1～h3 までを章や節の見出しとして扱い，h4～h6 は本文の一部とみなす．章･節の見出しとそれに後続する文章からなる構造もまた，**タイトルとその本文**ととる事ができる．よって**タイトルとその本文**という構造は，タイトルが表題の場合と，見出しの場合の 2 種類が与えられる．

ここで，2 つのキーワードのうち一方がタイトル，もう一方がそのタイトルに属する本文に現れる場合に，**構造を持って出現**したキーワードペアと呼ぶ．図 2 において，矩形で囲まれた中にタイトルとその本文という構造を捉えることができ，またそれが全体で入れ子状になっている事がわかる．この，構造を持って出現したキーワードペアに対し，本手法では距離数値をディスカウントして距離の補正を行う．

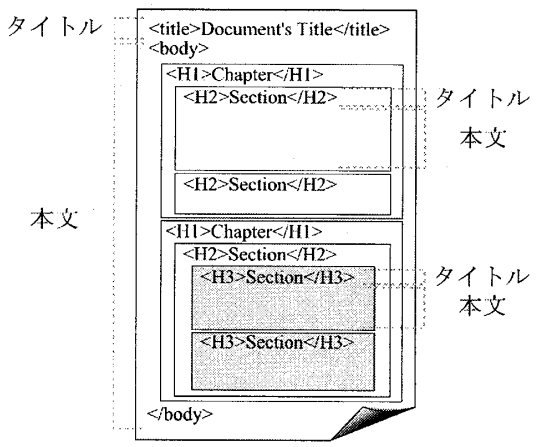

図 2 提案する HTML の構造

またキーワードが複数ある場合は，文書に存在するすべての異なる組み合わせについて，網羅的にペアを取得する方法をとった．すなわちクエリがキーワード 3 つ以上で構成される場合，あらゆる 2 つのキーワードペアについてその組み合わせを解析する．

4. 構造を考慮した近接性の有効性

まず著者らが着目する文書構造が検索結果へ与える影響を把握するために，予備調査を行った．使用したデータは，TREC-8 テストコレクションで使用されたアドホック検索タスクのうち，TREC により適合性判定結果が提供されている文書である[8]．さらにこのうち，クエリが複数キーワードで構成されたタスクを対象に解析を行った．

4.1 予備調査 I

はじめに，文書内におけるキーワード間の距離が，適合文書と不適合文書とでどの程度の差異があるのか調査を行った．ここでは構造を考慮せず，単純なキーワード間の距離を測定した．結果を表 1 に示す．

表 1 文書分類ごとの平均最短距離の差

	適合	不適合	(適/不適)
トピック数	45		
文書数/トピック	40.8	917.1	4%
ペア最短距離	242	538	45%

表 1 より，適合文書と不適合文書とにおいて，キーワードペア最短距離の平均は，適合文書が 45%短くなる事が認められた．これにより適合文書と不適合文書を判定する指標の 1 つとして，近接性が有効である可能性が示された．

4.2 予備調査 II

次に著者らが提案する**タイトルとその本文**という構造を持つペアの存在について，その程度を把握するために適合文書/不適合文書それぞれについて，キーワードペアが**タイトルとその本文**の構造を持って出現する文書数について調査を行った．なお，キーワードの一方が表題にあるペアが上段，見出しにあるペアが中段，その論理和が下段である．

表 2 文書分類ごとの構造解析

	適合	不適合	(適/不適)
ペアが構造を有す文書（表題）	2.98	24.71	
	7.10%	2.66%	266.9%
ペアが構造を有す文書（見出し）	3.62	26.4	
	9.03%	2.88%	313.5%
論理和	4.53	37.86	
	10.94%	4.11%	266.2%

表 2 に示した結果より，指定した構造を持って出現するキーワードペアが，適合文書内に含まれる割合は決して高くはないが，しかし適合文書と不適合文書においての含有率には大きな差が認められる．

4.3 予備調査Ⅲ

最後に，適合文書中において指定した構造を持って出現したキーワードペアが，どの程度の距離を持つのか調査を行った．結果を表 3 に示す．

表 3 適合文書の距離解析

	キーワードペア距離
文書全体	242
表題の構造	407
見出しの構造	277

表 3 より，文書全体の平均距離と比較して，構造を持って出現するキーワードペアの平均距離は小さいとは言いがたい．これより，タイトルと本文の構造を持って出現するペアはその共起距離が大きいため，単純な近接性評価では構造を持って出現したペアは文書内の共起距離のランキングに現れない，または評価が低すぎる可能性が示唆される．このため，今回ターゲットにしているタイトルと本文の構造が持つ関連性を考慮して，近接性に適切な評価を与えることで，検索精度の向上が期待できると考えられる．

5. 構造を考慮した文書スコア算出手法

前述したように，Web 検索において検索対象となる Web 文書は構造を有しており，単純な近接性評価では適切な評価を与えられない可能性がある．そこで近接性を評価する場合に文書の構造を考慮し，キーワードペアが持つ関連性を考慮した評価を与えるために，著者らの提案する近接性評価式，またキーワード関連度との組み合わせ方について次に述べる．

5.1 Tao の近接性

最初に，著者らがベースとした Tao らによる近接性の評価方法について述べる．Tao らは，近接性を測定する基準としてキーワードを含むスパンや平均距離などいくつかの測定法を提案した．その中で最も有効と評価されたキーワードペア最小距離 (Minimum pair Distance，以下 MinDist) について説明する．

MinDist は，次式により与えられる．

$$\delta_{(Q,D)} = \min_{q1,q2 \in Q \cap D, q1 \neq q2} \{Dis_{(q1,q2;D)}\} \quad (1)$$

これは，与えられクエリ Q に含まれるキーワード集合の可能な組み合わせのうち，文書 D 内において最短距離で出現するキーワードペア q1,q2 がもつ形態素数を現す．

また，この MinDist 値をスコア化し近接度スコアとするのに，Tao らは以下の式を用いた．

$$\pi_{(Q,D)} = \log(\alpha + \exp(-\delta_{(Q,D)})) \quad (2)$$

この関数は指数と対数を用いることで，近接性をスコア化する際に理想的に働くとされる．

また Tao らは，キーワード関連度 $BM25_{(Q,D)}$ と近接度スコア $\pi_{(Q,D)}$ を，以下の式を用いて組み合わせた．

$$R_{(Q,D)} = BM25_{(Q,D)} + \pi_{(Q,D)} \quad (3)$$

これはすなわち，単純にキーワード関連度 (BM25) と近接度スコアを加算した値を文書スコアとしている．

5.2 構造を考慮した近接性のスコア化

次に筆者らが提案する，構造を考慮した近接性のスコア化について述べる．まず，3 章で述べた構造を持って出現したペアについて，その距離の調整法を与える．そのペアがもともと持つ形態素数を $\chi_{(Q,D)}$，距離を短くするためのパラメータを γ とすると，そのキーワードペア距離は以下で与えられる．ここで，Q,D は Tao の近接性と同様，クエリと文書を指す．

$$\zeta_{(Q,D)} = \begin{cases} \chi_{(Q,D)} & \text{(構造を持たずに出現する時)} \\ \chi_{(Q,D)} \cdot \gamma & \text{(構造を持って出現する時)} \end{cases} \quad (4)$$

次にこのキーワードペア距離を，Tao らの式を改良した計算式を用いてスコア化する．

$$\psi_{(Q,D)} = \log(\alpha + \exp(-\zeta_{(Q,D)}/\beta)) \quad (5)$$

ここで α と β は任意の正の数である．変数 α は，近接度スコアとして有効な形態素数を定める．また筆者らは，Tao らの提案したスコア化式だと，形態素数の増加に対して近接度スコアの減衰が急すぎると判断した．そのため変数 β を導入し，形態素数の増加に対するスコアの減衰カーブを緩やかにした．

また Tao らは，近接度スコアとして先述の MinDist のみ，つまり 1 つの最短キーワードペアのみを近接度スコアに用いていたが，本稿では最短距離だけでなく，距離の短いほうから N 個のキーワードペアをスコア化に用いることを提案する．これは，そのクエリに関連する文書であるならば，複数のキーワードペアが出現するであろうという仮定に基づいている．またこれにより，偶然的にキーワード同士が近い位置に存在する，全く関係ない文書が高いスコアをつけられる，などの誤った近接度スコアの付与を低減させる狙いもある．

5.3 文書スコア算出法

ここでは，キーワード関連度と近接度スコアを組み合わせた文書スコアの算出法について述べる．Tao らは単純にそれらを加算していたが，本稿ではキーワード関連度と近接度スコアをそれぞれ正規化し，機械学習を用いてその線形和を求め，文書スコアを決定する．図 3 に，本稿で提案する文書スコアの算出法について示す．まず各文書に対しクエリが与えられると，そのクエリに対するキーワード関連度と近接度スコアが算出される．この時，異なるパラメータ設定 (図中の個々の○) により近接度スコアは異なった出力値を持つ．このうち，m 種のパラメータ設定を用いる手法を，**提案法 m** と呼ぶことにする．たとえば図 3 では，3 種のパラメータ設定に基づく近接度スコアを用いており，これは提案法 3 に相当する．これらの，

キーワード関連度と近接度スコアを機械学習によって得られた重みに基づいて加算したものが，本稿における文書スコアとなる．

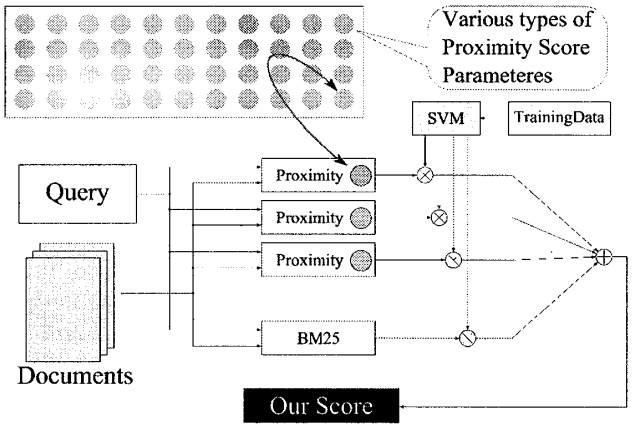

図 3　提案するスコア算出のフロー

本稿では機械学習の手法に RankingSVM を用いた[9]．線形カーネルを用いた RankingSVM は，訓練データから抽出される順序誤りを最小化するように，重みを設定する．

6. 評価

本章では，提案した手法を用いて検索精度の評価を行った．まず，実験に使用したキーワード関連度，データセット，パラメータなどについて述べた後，その条件で実験した結果について述べる．

6.1 キーワード関連度

キーワード関連度とは，要求されたクエリに対して Web 検索エンジンが文書のスコアの算出を行うための，1 つの代表的な検索方式として知られている．本稿ではキーワード関連度に BM25 を用いる[3]．なお BM25 のパラメータについては，一般的によく用いられる値（$k_1 = 2$, $b = 0.75$）を用いた．

6.2 データセット

今回はデータセットとして Glasgow 大学から提供されている WT2G と WT10G を用いた[12]．これらは Web 上の文書を大規模に構造化・集積したものであり，このうち TREC の WebTrack である TREC-8, TREC-9, TREC-2001 から，プーリングが行われ適合性判定データが付与された文書に対して解析を行った．表 4 に，それぞれのコーパス全体の概要を示す[13]．そのうち，クエリが 2 語以上のキーワードで構成されているタスクについて評価を行った WT2G では 47 件，WT10G では 77 件となる．

表 4 使用コーパス概要

	WT2G	WT10G
トピック	401-450	451-550
文書数	247491	1692044
平均単語数	2009.3760	2303.4063

また評価の指標として，Mean Average Precision (MAP) と Precision@n (P@n) を用いた[10]．これの値の計算には TREC 提供の評価ツール trec_eval を利用した[11]．MAP は平均適合率，また P@n (n=5, 10, 15･･･) は，ランキングの上位 n 件の文書適合率となる．

6.3 パラメータ α,β の決定

本稿で提案する近接性をスコア化する式は，5.2 節で述べたように設定すべきパラメータが多く存在する．本稿ではそのうち，近接性をスコア化する式のパラメータ α と β を，実験的に決定した．

コーパス WT2G の一部のタスクを用いて，それぞれの値を変化させた時の正規化適合率の変動について調査を行った．この調査においては，構造を考慮していない．また正規化適合率とは，MAP 値･P@n 値をそれぞれの最大値で除したものである．α についての結果を図 4，β についての結果を図 5 に示す．まず，変数 α で有効とする携帯素数を定め，次にその α の範囲内で，スコアの減衰カーブを定めるために変数 β を決めた．

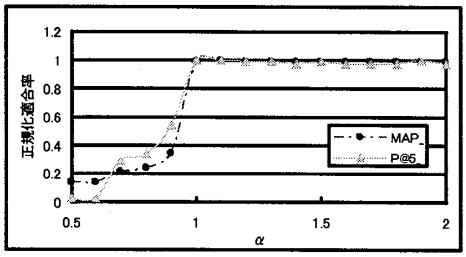

図 4　α とスコアの変動

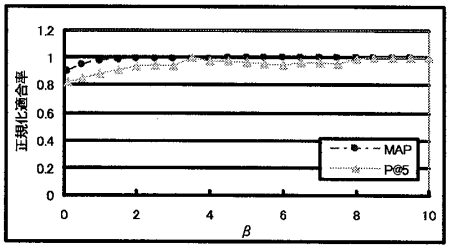

図 5 β とスコアの変動

この結果より，α=1.1，β=8.6 に決定した．これによって得られるスコアカーブは図 6 の通りである．Tao らのカーブと比較し，著者らのカーブは緩やかな曲線を描く．これは，やや距離が離れたキーワードペアでも，近接度スコアに影響を与える事を示す．

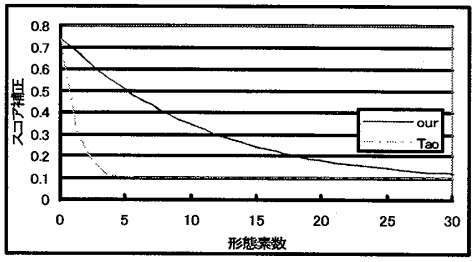

図 6 形態素数によるスコアカーブ

6.4 結果と考察

ここでは，5.2 節で述べたパラメータのうち構造に対するディスカウント値 γ・スコア化に使用するペア数 N，そして RankingSVM のパラメータについて示す．

最初に，構造に対するディスカウント値 γ について述べる．ここで，2 つのキーワードのうち片方がタイトルに存在する場合を考える．タイトルが表題の場合を**表題とその本文**，タイトルが見出しの場合を**見出しとその本文**とした時に，それぞれの事象について異なるディスカウント値を与えた．ただし，heading タグの種類の違いは考慮しなかった．

次に，スコアに使用するキーワードペア数について述べる．Tao らは MinDist，つまり N=1 のみを用いたが，提案法では文書内に存在するペアのうち，距離が短い方から上位 N 個を用いた．具体的には，N=1(MinDist)から N=ALL (文書に出現する全てのペアを使用) までである．

最後に，キーワード関連度と近接度スコアを組み合わせるときの，RankingSVM のパラメータについて述べる．RankingSVM には多くのアルゴリズムやパラメータがあるが，本稿ではカーネルには線形カーネルを，ソフトマージンについては 0.01～1.0 までを，その他のパラメータは初期設定を用いた．また機械学習に対する評価は，5-fold cross validation を用いた．まずコーパスを処理したデータセットを用意する．それをトピックにより 5 分割し，うち 3 つを training data，1 つを validation data，残りの 1 つを evaluation data とした．まず training data で学習を行い，validation data で RankingSVM の最適なソフトマージンを半自動で決定し，最後に evaluation data で評価を行った．これを，異なる組み合わせで 5 回繰り返すことで，データセット全てのクエリに対し評価を行った．

ここで，近接度スコアに関するパラメータをまとめて表 5 に示す．

表 5 パラメータ一覧

要素		値の範囲
ディスカウント値 γ	表題 (T)	1.0, 0.5, 0.3, 0.1, 0.05, 0.03
	見出し (H)	1.0, 0.3, 0.2, 0.1, 0.07, 0.05, 0.03
スコア化パラメータ	α	1.1
	β	8.6
利用するペア数（上位 N）		1, 5, 10, 15, 20, 25, All

表 5 のパラメータ設定を用いて，WT2G，WT10G それぞれのコーパスにおいて，以下の組み合わせで実験を行った．

表 6 手法とその表記法

名称	Feature
キーワード関連度のみ	BM25
＋既存の近接性	BM25 + Proximity(MinDist)
提案法 1	BM25 + Proximity
提案法 2	BM25 + Proximity*2
提案法 3	BM25 + Proximity*3

ここで Proximity (MinDist)とは，(2)式において，利用するキーワードペア数 N=1 の時の近接度スコアを指す．+既存の近接性とは，5.1 節 (2) 式とキーワード関連度を機械学習により組み合わせた検索システムである．また Proximity は 5.2 節 (5) 式において，上記パラメータをあてはめた近接度スコアである．

提案法 m (m=1～3) については，5.3 節で述べたとおりである．また，文書のスコア化に使用する，適切な近接度スコアを feature として選出するために，以下の操作を行った．まず，すべてのパラメータの組み合わせによる近接度スコアから feature をランダムに選び出し，キーワード関連度と組み合わせた．そこから算出された MAP 値と feature の重みを指標として，各パラメータの影響を解析した．その結果を考慮して，ヒューリスティクスに使用する feature の絞込みを行い，10～30 ほどを残した．そこから先ほどと同様にランダムに feature を選出し，BM25 と組み合わせスコアのよいものを選出した．

このように上記で決定したパラメータを用い，総合的な検索性能について評価を行った．データセット WT2G についての結果を表 7,8 に，WT10G についての結果を表 9,10 に示す．なお feature 項の記号は，T が表題に対するディスカウント値，H が同様に見出しに対するディスカウント値，N が利用したペア数となる．また表 8･10 における ratio とは，RankignSVM により算出された重みを指す．

表 7 WT2G における検索精度

	MAP	P@5	P@10	P@15
BM25	0.2562	0.4426	0.3957	0.3518
+既存	0.3000	0.4493	0.3967	0.3610
提案法 1	0.3093	0.4240	0.3907	0.3726
提案法 2	0.3113	0.4156	0.3867	0.3682
提案法 3	0.3178	0.4480	0.4062	0.3782

表 8 WT2G で使用された feature とその重み

	feature			ratio
	T	H	N	
+既存	BM25			1.5548
	1.0	1.0	1	1.9876
提案法 1	BM25			1.2023
	0.1	0.2	5	3.1657
提案法 2	BM25			1.1431
	0.1	1.0	5	2.7905
	0.03	0.07	15	0.7535
提案法 3	BM25			1.1050
	1.0	1.0	5	2.4656
	0.03	0.1	10	0.3444
	0.5	1.0	10	1.1414

表9　WT10Gにおける検索精度

	MAP	P@5	P@10	P@15
BM25	0.2038	0.3091	0.2610	0.2364
+既存	0.2156	0.3028	0.2718	0.2416
提案法1	0.2125	0.2923	0.2737	0.2418
提案法2	0.2228	0.3235	0.2773	0.2475
提案法3	0.2222	0.3158	0.2695	0.2528

表10　WT10Gで使用されたfeatureとその重み

	feature			ratio
	T	H	N	
+既存	BM25			2.3959
	1.0	1.0	1	0.8224
提案法1	BM25			1.9137
	1.0	0.1	5	1.5424
提案法2	BM25			1.9877
	0.01	0.3	ALL	-0.1064
	0.1	0.2	1	0.9605
提案法3	BM25			2.0010
	0.01	0.2	ALL	-0.6850
	0.5	0.1	ALL	0.5584
	0.03	0.07	1	0.9730

結果より，WT2GおよびWT10G双方のコーパスにおいて，提案法でMAP値が上る事が確認できた．キーワード関連度をそれぞれ基本とすると，＋既存の近接性および提案法1で4～20%程度の向上が見られた．また使用する近接度スコア (feature) の数を増やすことで，スコアが向上する傾向があるのがわかる．

次に，提案する近接度スコアのパラメータの傾向について述べる．まず構造を有して出現したペアに対するディスカウント値は，表題・見出しにかかわらず，有効な値の範囲は広くなった．しかし，WT2G・WT10G双方において，有効であると示されたfeatureのディスカウント値は1.0以外が多い．これより，タイトルとその本文という構造を有して出現するペアに対しては，その距離を縮める本提案法が有効だと言える．またディスカウント値が一定とならない原因として，headingタグの種類や文書長に関わらず，ディスカウント値が一定であることが挙げられる．この場合，大きい文書単位では適当なディスカウント値が，小さい文書に対しては過大評価となってしまうなど，細かな調整をしていなかったためである．

また利用するペア数は，コーパスにより異なるものとなった．WT2GではN=5～15の範囲が有効であったが，WT10GではN=1～全体まで，使用するペア数の範囲は広い．しかし，featureの組み合わせでN=1のみの結果がない事より，複数ペアをスコア付けに用いる手法は有効であると言える．

またキーワード関連度と近接度スコアの重みは，コーパスにより傾向が異なった．WT2Gでは，キーワード関連度より近接度スコアが重みが大きく，提案法1では近接度スコアが，キーワード関連度に対して2.6倍の重みをrankignSVMにより与えられた．しかしWT10Gでは，同じ提案法1であっても，キーワード関連度に対する近接度スコアの重みは0.8にしかならなかった．これは近接度スコアの有効性が，文書やクエリにより大きく異なることを現していると考えられる．

7. おわりに

本稿では，構造を考慮した近接性のスコア化・キーワード関連度との組み合わせ方を提案した．その手法についてWT2G・WT10G2つのコーパスで評価を行った結果，提案した手法によりMAP値が向上することを確認した．

本稿では，Web検索において文書のスコア付けの要素となるキーワードの近接性評価について，文書の構造，特にタイトルとその本文を取り上げ，スコア算出法を提案した．評価の結果，MAP値において提案手法を用いることで，検索精度が向上する事を確認した．

本稿は今回ターゲットとした構造の有効性を実験的に評価したものであり，より系統立てた手法を確立することで，MAP値やP@nは改善すると思われる．以下に，今後の課題を述べる．

最初に，近接度スコアに関して，文書サイズを考慮した近接性パラメータの検討をしたい．今回は文書サイズにかかわらず，距離のディスカウント値を一定にしてしまったが，これを文書サイズを考慮した値にすることで改善できると考えられる．また，HTMLタグのh1とh3は，それをタイトルとした時に後続する文章のサイズが異なることが考えられるが，本実験ではタグの種類による値を考慮してはいない．また，近接性をスコア化する際の計算式の改良，特にクエリ中にキーワードが3つ以上で構成される場合などを柔軟に対処できる手法を考案したい．

次に，ランキング学習の利用に関して，今回は提案法で用いたfeatureは，絞込みをヒューリスティクスで行い，そこからランダムに組み合わせた．それを絞込み・組み合わせともに自動で最適化を可能にしたいと考えている．

参考文献

[1] Sergey Brin, Lawrence Page, The anatomy of a large-scale hypertextual web search engine, Proceedings of the seventh international conference on World Wide Web 7, p.107-117, April 1998, Brisbane, Australia

[2] Salton,G. and Yang, C.G.: "On the Specification of Term Values in Automatic Indexiing." Journal of Documentation 29, 1973.

[3] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at Trec-3. In D. K. Harman, edhitor, The Third Text Retrieval Conference (TREC-3), 1995.

[4] Keen, E.M., 1992 Some aspects of proximity searching in text retrieval systems, Journal of information Science, vol. 18, No. 2, 89-98 (1992)

[5] Rasolofo, Y. and Savoy, J.: Term Proximity Scoring for Keyword-Based Retrieval Systems, ECIR '03, pp.207-218(2003)

[6] Tao, T. and Zhai, C.: An exploration of proximity measures in information retrieval, in SIGIR '07, pp.295-302 (2007)

[7] Rasolofo, Y., and Savoy, J. 2003 Proximity-aware scoring for XML retrieval ACM SIGIR conference on Research and development in information retrieval}, Singapore, Singapore New York, NY, USA: ACM, 845-846.

[8] D. Hawking, E. Voorhees, N. Craswell, P. Bailey: Overview of the TREC-8 Web Track, 2000,2

[9] Joachims, T.: Optimizing search engines using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002.

[10] Manning, Christopher D. and Raghavan, Prabhakar and Schtze, Hinrich, Introduction to Information Retrieval, 2008, Cambridge University Press

[11] trec site, http://trec.nist.gov/trec_eval/

[12] University of Glasgow, http://ir.dcs.gla.ac.uk/test_collections/

[13] Ben.H, Iadh,O:: Term Frequency Normalisation Tuning for BM25 and DFR Models, 2005, Lect Notes Comput Sci