重回帰分析を用いた近接クエリの重み推定によるWeb情報検索

柴田 鉄也† 江口 浩二†
† 神戸大学大学院工学研究科情報知能学専攻

1 はじめに

本研究では，クエリを構成する複数の語の組み合わせである近接クエリ要素に対して，その重みを重回帰分析を用いて推定する手法を提案する．まず，訓練データを用いて最適な近接クエリ要素の重みを決定し，それを目的変数とする．次に，相互情報量をはじめとした近接クエリ要素に関する様々な特徴量を説明変数とした線形重回帰モデルの係数を推定する．新たなクエリに対しては，その特徴量とすでに推定した回帰係数を用いることで，近接クエリの重みを推定できる．このように独立した語だけでなく近接クエリの重みを推定することで，自然言語文クエリによるWeb情報検索の高精度化を目指す．

2 回帰を用いた近接クエリ要素の重み推定手法

本節では，回帰を用いて近接クエリ要素の重みを推定する手順について述べる．本研究はRegression Rank[1]の拡張となるもので，特に回帰の部分はその手法を基盤としている．[2]において，語間依存性モデルとRegression Rank[1]を組み合わせた手法が提案されているが，近接クエリの重みについては一定値を一律に与えている．そこで本研究では，課題となっていた近接クエリの重みを推定する手法を提案する．

(1)訓練データを用いて最適な近接クエリ要素の重みを決定

(2)近接クエリ要素に関する様々な特徴量を説明変数とした線形重回帰モデルの係数を推定（近接クエリ要素に関する特徴量の重みが決定）

(3)新たなクエリに対しては，そのクエリの特徴量とすでに推定した回帰係数（特徴量の重み）を掛け合わせ，近接クエリの重みを推定

このような手順で近接クエリ要素の重みを推定する．また，これらの手順をwindow sizeごとに行い，最終的には複数のwindow sizeを持つ近接クエリを用いて検索を行う．以降，順にそれぞれの詳細について述べる．

2.1 最適な近接クエリ要素の重みの決定

回帰を行うには，最適な近接クエリ要素の重みが必要であるため，まずはその重みを特定する．その方法を述べる前に，前提として，検索は重みつきの単語クエリ要素と重みつきの近接クエリ要素を足し合わせて行う．また近接クエリ要素の最適な重みを特定するために用いる単語クエリ要素の最適な重みは，既に特定済みであると仮定している．

さて本題に戻り，訓練データを用いた近接クエリ要素の重みの最適化方法であるが，平均精度（MAP）を最大化するよう，単純な勾配法を用いた．まず，単語クエリ要素のみに対して重みを最適化し，次に，近接クエリ要素の重みを最適化した．

Estimating proximity query weights with multiple regression for Web Retrieval

Tetsuya SHIBATA† and Koji EGUCHI†, †Department of Computer Science and Systems Engineering, Kobe University

2.2 近接クエリの特徴量

近接クエリの最適な重みがわかったら，次は近接クエリを表現するための特徴量を定義する必要がある．表1に我々の用いた近接クエリ要素に関する特徴量を載せる．本研究では，近接クエリ要素は2つの語から成る物に限定している．よってすべての特徴量は2つの語 (q_1, q_2) に関するものとなる．この表中でtf,dfはそれぞれ単語頻度，文書頻度を表す．また，各特徴量は $[0,1]$ になるように，クエリごとに正規化している．

表1: 近接クエリの特徴量

特徴量名 (タイプ)	定義
品詞 (boolean)	各品詞を含むか？
ストップワード (boolean)	ストップワードを含むか？
位置関係 (boolean)	同じ文節に属するか？ 単語が隣り合っているか？ 連続する文節に属するか？ 係り受け関係があるか？ 複合名詞であるか？
term frequency (integer)	$\sum(tf_1,tf_2)$ $\max(tf_1,tf_2)$ $\min(tf_1,tf_2)$
document frequency (integer)	$\sum(df_1,df_2)$ $\max(df_1,df_2)$ $\min(df_1,df_2)$
google term frequency (integer)	$\sum(tf_1^{google},tf_2^{google})$ $\max(tf_1^{google},tf_2^{google})$ $\min(tf_1^{google},tf_2^{google})$
共起文書数 (integer)	共起する文書数 100語以内に共起する文書数
語間距離 (real)	全文書内での最小距離 全文書内での最小距離の平均 全文書内での最小距離の平均（100語以内に現れる場合）
相互情報量 (real)	$\log \frac{p(q_1,q_2)}{p(q_1)p(q_2)}$
Dice係数 (real) Jaccard係数 (real) Overlap係数 (real) Cosine距離 (real)	$\frac{2\times count(q_1\cap q_2)}{df_1+df_2}$ $\frac{count(q_1\cap q_2)}{df_1+df_2-count(q_1\cap q_2)}$ $\frac{count(q_1\cap q_2)}{\min(df)}$ $count(q_1\cap q_2)/\sqrt{df_1\times df_2}$

2.3 回帰による特徴量の重みの推定

訓練データにより，最適な重みが特定され，さらにそれぞれの近接クエリ要素を表現する特徴量が得られた．次は，これらを用いて重回帰により，特徴量の重みを特定する．重回帰において，本手法では近接クエリのwindow sizeごとに特徴量の重みを推定する．またクエリごとに回帰を行い，各クエリの特徴量の重みを平均する．

以下，回帰の詳しい説明を行う．あるクエリには，N つの近接クエリ要素があるとする．そこで $Y = \{y_1,...,y_N\}$ はそれらの最適な重みである．さらに $\mathbf{X} = \{X_1,...,X_N\}$ を特徴量ベクトルとする．次に，d を特徴量の数であるとし，$X_i = \{x_i^0, x_i^1,...,x_i^d\}$ を i 番目の特徴量ベクトルとする．また $W = \{w_0, w_1,...,w_d\}$ を特

徴量の重みベクトルとする．我々は，この特徴量の重みベクトルを求めたい．ここで既存手法と同じように L2 線形回帰を用いて，$\sum_i^N L(f(X_i,W),Y_i) = \sum_i^N (y_i - \sum_{j=1}^d w_j x_i^d)^2 = (Y-\mathbf{X}W)^T(Y-\mathbf{X}W) + \beta X^T W$ で近似し，最適な重みベクトル W^* を見つけるために最小化する．ここで，既存手法と同じように $\beta = 1$ をパラメータとして用いた．結果的に L2 線形回帰は閉解：$W^* = (\beta I + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$ を持つ．I は恒等行列を表す．

2.4 新しいクエリに対する重みの推定

回帰によって，特徴量の重みが得られた．これを用い，新しいクエリに対して単語の重みと複数の window size の近接クエリの重みを列挙すると次の式になる．

$$Y_t, Y_{w_2}, ..., Y_{w_n} = \mathbf{X}_t W_t, \mathbf{X}_w W_{w_2}, ..., \mathbf{X}_w W_{w_n} \quad (1)$$

ここで，Y_t, Y_{w_n} はそれぞれ単語クエリと近接クエリの重みベクトル，$\mathbf{X}_t, \mathbf{X}_{w_n}$ を単語と近接性の特徴量ベクトル，W_t, W_{w_n} を推定した特徴量の重みベクトルとする．ここで $\mathbf{X}$ は，各クエリ要素ごとに特徴量ベクトルを持つ，ベクトルのベクトルである．n を window size とし，これらを用いて検索を行う．例として，Y_t は (0.40 "サルサ" 0.20 "踊れる" 0.05 "方法" ...)，Y_{w20} は (0.003 #uw20 ("サルサ" "を") 0.05 #uw20 ("サルサ" "踊れる")...) のようになる．ここで，語または語の組の直前の数値は重みを示し，「#uw20()」は 20 語の window 幅を条件とする近接演算を示す．

3 実験と評価

提案手法の有効性を検証するため，評価実験を行う．実験では，提案手法により重みを推定した近接クエリと既存手法の Regression Rank を用いて重みを推定した単語クエリを組み合わせて用いる．また本研究では，検索システム Indri[1] をインデクシングなどの検索プラットフォームとして利用している．

データセット 実験に用いたデータセットは，NTCIR-3 WEB（訓練用），NTCIR-4 WEB（テスト用）[2] という Web 検索評価用テストコレクションである．これらはそれぞれ，100GB の文書データ，検索課題 47, 35 件（日本語），適合判定データからなる．本論文では，上述のテストコレクションの検索課題の DESC フィールドに記述された自然言語文を，クエリとして使用した．

近接クエリの選択 回帰によって，すべての近接クエリ要素の重みが推定されるが，実際に有用な近接クエリは [2] によると 1 – 6 つだけである．よって本研究では，window size ごとに近接クエリの中から重み上位 $Rank = N$ 件のクエリを用いる．

実験結果 まず 1 つ目の実験として，提案手法で近接クエリ上位何件使用するかについての実験を行った．結果として $Rank = 10$ のとき，最適値をとることがわかったため，これ以降の実験を $Rank = 10$ として行う．また単一の window size のみを用いた実験結果を表 2 に示す．実験では，平均精度（MAP）を評価基準として評価を行っている．

表 2: 単一の window size による実験結果

Rank	2	5	10	20	30	50
MAP	0.1630	0.1653	0.1680	0.1704	0.1685	0.1687

[1] http://www.lemurproject.org/indri/
[2] http://research.nii.ac.jp/ntcweb/

表 3: window size の組み合わせごとの結果

	2	5	10	20	30	50
2		0.1666	0.1703	**0.1775**	0.1731	0.1743
5	0.1666		0.1647	0.1694	0.1656	0.1695
10	0.1703	0.1647		0.1700	0.1678	0.1713
20	**0.1775**	0.1694	0.1700		0.1713	0.1734
30	0.1731	0.1656	0.1678	0.1713		0.1691
50	0.1741	0.1695	0.1713	0.1734	0.1691	

次に，window size の最適な組み合わせを決定するための実験を行う．ここで window size の候補としては，2,5,10,20,30,50 を用いる．3 つ以上も可能ではあるが，今回は 2 つの window size の全組み合わせで実験を行った．このとき $(Y_t, Y_{w_2}, Y_{w_{20}}) = (0.8, 0.1, 0.1)$ としている．その結果を表 3 に示す．これによると，window size の 2 と 20 を組み合わせた場合が最適値をとることがわかる．

次に，単語クエリと近接クエリの重みの比率を決めるための実験を行う．今回，各 window size の近接クエリの合計値には均等に重みを割り当てている．その結果，$(Y_t, Y_{w_2}, Y_{w_{20}}) = (0.8, 0.1, 0.1)$ のときに評価値（MAP）が最適となり，0.1775 となった．

次に訓練データで，名詞と動詞を重み均等で使った場合，Regression Rank のみを使った場合，そして提案手法を使った場合の MAP の最適値（window size を単一で使用，window size を組み合わせて使用）について比較を行った．その結果を表 4 に示す．ここで RR は Regression Rank を示す．結果として，提案手法は Regression Rank のみを使った場合と比べて，10.80% の改善が見られた．

表 4: 訓練データによる実験結果の比較

Method	名詞+動詞	RR	window20	window2,20
MAP	0.1428	0.1602	0.1704	0.1775

最後に訓練データで決定したパラメータを用いてテストデータで実験した結果を表 5 に示す．この結果，提案手法は Regression Rank のみを使った場合と比べて，8.77% の改善が見られた．これで新しいデータに対しても提案手法が有効であることが示された．

表 5: テストデータによる実験結果の比較

Method	名詞+動詞	RR	window20	window2,20
MAP	0.1327	0.1379	0.1478	0.1500

本研究で，回帰を用いた近接クエリ要素の重み推定と window size の組み合わせによって，自然言語文クエリの検索精度が改善されることが示された．

謝辞 本研究の一部は，科学研究費補助金基盤研究 (B)（20300038）の援助による．

参考文献

[1]Lease, M., Allan, J. and Croft, W. B.: Regression Rank: Learning to Meet the Opportunity of Descriptive Queries, *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, Berlin, Heidelberg, Springer-Verlag, pp. 90–101 (2009).

[2]Lease, M.: An improved markov random field model for supporting verbose queries, *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, ACM, pp. 476–483 (2009).