D-035

クエリログから抽出した関連語集合を用いたウェブページ検索

A Method for improving Web search ranking using relevant terms extracted from query log

藤田 尚樹　高橋 大和　宮原 伸二　片渕 典史　片岡 良治
Naoki Fujita　Yamato Takahashi　Shinji Miyahara　Norifumi Katafuchi　Ryoji Kataoka

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

1. はじめに

近年，検索エンジンのウェブページ検索はインターネットを利用する際に欠かせなくなっている．検索エンジンは，検索者のクエリに含まれる検索キーワードとの適合度やウェブページ自体の重要度が高い順に検索結果を表示する．適合度は検索キーワードのページ内頻度（tf）等から，重要度はページ間のリンク情報やドメイン名などから算出される．

一般的に適合度はtfが高ければ高いとされるが，検索者は検索キーワードを多く含むページではなく，検索キーワードに関する情報を求めており，tfが高くても検索者の求める情報を含まない場合もある．そのため，検索エンジンはtf以外にも被リンクアンカーテキスト情報の考慮や，検索キーワードの絞り込み語や同義語によるクエリ拡張等の検索精度向上策を行っている．

我々は検索精度を向上させるため，tfに検索キーワードの関連語のページ内頻度も考慮する手法を提案しており，Wikipedia から抽出した関連語集合を用いた評価実験により検索精度の向上を確認した[4]．Wikipedia はインターネット上の百科事典であり，ある語に対して一つ，もしくは複数の意味が記述され，リンク構造から意味に基づいた関連語集合を抽出できる．この関連語集合を用いることで検索キーワードの意味を的確に表現したウェブページを検索上位に提示することが可能となる．

しかし，検索者が必ず検索キーワードの意味を的確に表現したウェブページのみを求めているとは限らない．例えば著名な俳優 A を検索キーワードとした場合，「A とは誰か」という詳細な情報を含んでいるページを求める検索者以外に，A の画像や最近の出演映画，ニュースなど「A の何か」を知りたい検索者も多い．その際に慣れている検索者は「画像」や「映画」などの検索キーワードをクエリに追加することで効果的な検索を行っている．この様な効果的な検索となるように補助する手法として，安川ら[5]は「関連語とは，Web 検索エンジンに対してユーザが入力する 1 組の検索クエリに含まれる語のことである」とし，クエリログから得られた関連語を用いてクエリを修正することで検索者の満足度を高める手法を提案している．我々も複数のキーワードからなるクエリを用いる検索者は検索意図を明確に表現しているとし，クエリ情報から検索精度の向上に寄与する関連語集合を抽出することが可能と考える．

本論文では，1 つのクエリに含まれる検索キーワード同士を関連語と考え，検索エンジンのクエリログから抽出した関連語集合を用いた検索ランキング手法を提案する．評価実験では提案手法と Wikipedia から抽出した関連語集合を利用した従来手法[4]及び，関連語を考慮しない場合とを比較し，提案手法が検索精度向上に有効であることを示す．

2．提案手法

我々はページの内容を評価する際に検索キーワードとその関連語が多数共起されていれば，検索キーワードについて書かれている可能性が高いと考える．そこで，関連語の頻度を考慮した適合度評価手法として，関連語の頻度を考慮した重み付tfを「tfに重み付係数を乗算した関連語の頻度を加算した値」とし，それを従来のtfと置き換えてBM25[2]の計算に利用することで，関連語を考慮した適合度を算出する．以下，関連語の抽出，重み付tfの計算，BM25 を用いた適合度計算手法について述べる．

2.1 クエリログから関連語抽出

クエリログから関連語を抽出する手法として，安川ら[5]と同様に，我々もクエリに含まれる検索キーワード同士は，検索者の検索意図に対して関連を持つと考え，1 つのクエリに含まれる複数の検索キーワード同士を関連語の関係とする．これに基づき，検索キーワードtの関連語は全てのクエリログから検索キーワードtと共起している語を関連語の集合W_tとして抽出する．

2.2 重み付tfの計算

重み付tfの計算は従来手法[4]を拡張し，式(1)により検索キーワードt_iのドキュメントd_nにおける重み付tfである$wtf(t_i, d_n)$を算出する．ここで，t_iのd_nにおける頻度を$tf(t_i, d_n)$とし，αを関連語の頻度を反映する関連語反映度とする．

$$wtf(t_i,d_n)=\begin{cases}(1-\alpha)tf(t_i,d_n)+\alpha\sum_{t_j\in W_{ti}}tf(t_j,d_n), & tf(t_i,d_n)>0\\ 0, & tf(t_i,d_n)=0\end{cases} \quad (1)$$

2.3 BM25 を用いた適合度計算

式(1)で求めた$wtf(t_i, d_n)$を用いて BM25 のスコアを計算する．m 個の検索キーワードからなるクエリ$\boldsymbol{q}=(t_1,..,t_m)$によるドキュメント$d_n$の BM25 のスコア$BM25(\boldsymbol{q}, d_n)$を計算する式(2)を以下に示す．ここで，$N$は検索対象のドキュメントの総数，$df(t_i)$は検索対象の中で$t_i$を含むドキュメント数であり，$dl(d_n)$は$d_n$のドキュメント長，$avdl$は検索対象の平均ドキュメント長である．$b$と$k_1$は計算用パラメータであり，評価実験では$b=0.75$，$k_1=1.2$として利用する．

$$BM25(\mathbf{q},d_n)=\sum_{i=1}^{m}w_{dn}(t_i) \quad (2)$$

$$w_{dn}=\frac{(k_1+1)wtf(t_i,d_n)}{k_1(1-b+b\frac{dl(d_n)}{avdl})+wtf(t_i,d_n)}\times\log\frac{N-df(t_i)+0.5}{df(t_i)+0.5}$$

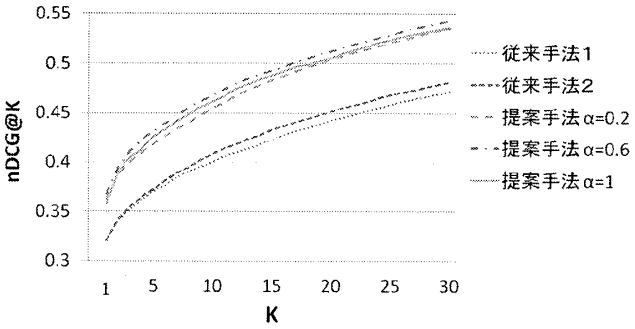

図 1. nDCG@K

3. 評価実験

3.1 実験概要

提案手法により得られた適合度を用いた検索ランキングの精度評価実験を実施した．実験におけるデータセットとして商用検索エンジンのクエリログから抽出した 1 つの検索キーワードで構成されたクエリ 420 件とクエリ毎に評価値を付与した商用検索エンジンの検索結果上位 200 件のウェブページ(合計 84,000 件)を用いた．評価値は 3 人の評価者により 6 段階の評価を行い，平均値を評価値とした．ウェブページは JTAG[3]を用いた形態素解析を行い，転置インデックスを作成して検索に用いた．形態素解析の際は検索キーワード及び関連語は簡易化のため，ユーザ定義語句として 1 語として扱った．

関連語を抽出するためのクエリログは，商用検索エンジンのクエリログを用い，提案手法に基づきデータセットの 420 件のクエリに対する関連語約 130 万語を抽出した．

検索精度評価指標は情報検索で一般的に用いられている nDCG[1]を採用した．6 段階で付与された評価値はより精度の変化を明確にするため(4, 3, 2, 1, 0, 0)の多値適合判定データとして用いた．ただし，検索結果として得られたページの中で，当該クエリに対する評価値が付与されていないページは検索結果から除外して nDCG の計算を行った．

提案手法の有効性を確認するための比較対象として，関連語を考慮しない通常の BM25 による検索ランキングを従来手法 1，Wikipedia から抽出した関連語情報を用いての提案手法と同様の検索ランキングを従来手法 2 として，提案手法との検索精度の比較を行った．比較の際に従来手法 2 の関連語反映度αは最も検索精度が高い結果を得られた $\alpha = 0.6$ とした．Wikipedia のデータはデータベースダウンロードサイトから取得した．

3.2 実験結果

図 1 に実験結果として従来手法 1，従来手法 2 及び関連語反映度 $\alpha = 0.2, 0.6, 1.0$ の場合の提案手法の検索結果上位 K 件の nDCG を示す．図 1 から提案手法の nDCG は従来手法 1 及び従来手法 2 に比べ，高い値となっており検索精度が向上していることがわかる．

また，提案手法と従来手法 1 及び従来手法 2 との統計的有意性を検証するため，ウィルコクソンの符号付き順位和検定を検索結果上位 10 件，20 件，30 件の精度に対して実施した．検定の結果，すべての場合について有意差あり（有意水準 1%）という結果となり，従来手法 1 及び従来手法 2 に対する提案手法の有効性が確認された．

表 1. nDCG@20 の変化量毎のクエリ数(従来手法 1 比較)

変化量⊿	-30%未満	-30%以上-5%未満	-5%以上+5%未満	+5%以上+30%未満	+30%以上
従来手法 2	5	59	244	90	22
提案手法（α=0.6）	10	70	94	128	118

4. 考察

提案手法がどのようなクエリに対しても平均的に検索精度向上を実現できているか，引き続き検証した．表 1 に，提案手法と従来手法 2 の各クエリの検索結果上位 20 件の精度(nDCG@20)を従来手法 1 と比較した場合に精度が変化した量ごとのクエリ数を示す．変化量⊿が「+5%以上 30%未満」，「+30%以上」のクエリの占める割合が従来手法 2 に比べて提案手法は大幅に増加しているため，全体としての検索精度は向上している．しかし，変化量⊿が「-30%未満」，及び「-30%以上-5%未満」のクエリは従来手法 2 と提案手法ともに約 2 割存在しており，わずかに提案手法の方が多い．したがって，提案手法が有効に機能せず，適用した結果逆に検索精度が低下してしまう検索クエリがあることがわかった．

5. まとめ

本論文では，クエリログから抽出した関連語情報を用いて，ページ内の関連語頻度を考慮した検索ランキング手法を提案した．評価実験では商用検索エンジンのクエリログを元に作成したデータセットを用いて，従来手法である通常の BM25 及び Wikipedia から抽出した関連語情報を用いた検索手法との比較を行った．評価実験を通じて，提案手法は従来手法より高い検索精度が得られ，クエリログから抽出した関連語情報を利用することで検索精度向上が可能であることを示した．今後は精度が低下した一部のクエリへの対応を検討し，さらなる精度向上に取り組む予定である．

参考文献

[1] K.Järvelin, and J.Kekäläinen, “IR evaluation methods for retrieving highly relevant documents”, *Proc. 23th SIGIR Conf. on Research and Development in Information Retrieval*, pp.41-48 (2000).

[2] S.E.Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. “Okapi at TREC-3”, *Proc. TREC 1994*. Gaithersburg, USA, November 1994.

[3] T. Fuchi and S. Takagi, “Japanese Morphological Analyzer using Word Co-occurrence —JTAG—”, *Proc. of COLING-ACL*, pp.409-413, 1998.

[4] 藤田尚樹, 高橋大和, 小長井俊介, 片岡良治: “関連語を考慮した重み付 tf を用いた BM25 による検索精度向上”, 第 17 回 Web インテリジェンスとインタラクション研究会,WI2-2010-20.

[5] 安川美智子, 横尾英俊: “クエリログから獲得した関連語のクラスタリングに基づく Web 検索”, 電子情報通信学会論文誌 D, Vol.J90-D, No.2, pp.269-280 (2007-02).