検索キーワードとコンテキストとの相関に基づく検索文書のリランキング

長谷川 隆明　　今村 賢治　　菊井 玄一郎
日本電信電話株式会社　NTT サイバースペース研究所
hasegawa.takaaki@lab.ntt.co.jp

１．はじめに

サーチエンジンに入力される検索キーワード数は１語もしくは２語が大半であり、ユーザの検索意図を明確にするのは難しい。パソコン向けの Web ページの検索では、PageRank[1]に代表される Web ページ間のリンク構造の情報を用いることにより、検索キーワード数が少なくても多くのユーザの検索意図を反映しているかのような実用的な検索精度を達成してきた。しかしながら、近年ニーズが高まっている携帯電話向けの Web ページの検索では、現状では Web ページ間のリンク情報が十分存在しているとは言えないために、リンク情報の効果は限定的であると考えられる。

そこで我々は、リンク情報を用いずに検索精度を上げるために、検索キーワードが文書中に出現するコンテキストに着目した。その動機は、検索キーワードと高い相関があるコンテキストを有する文書はその検索キーワードによく適合する文書ではないかという仮説に基づいている。検索キーワードを含む文書のうち、検索キーワードと相関の高いコンテキストを有する文書を上位にランキングすることによって、検索精度を向上させる方法を提案する。

本稿では、まず第２節でリンク情報を用いずに検索精度を向上させるための従来手法について述べ、第３節で検索キーワードと相関の高い単語や固有表現を用いた検索文書のリランキングの方法について説明する。第４節で評価実験の結果について報告し、第５節で結論を述べる。

Document Re-ranking using Correlations between Queries and Contexts by Takaaki Hasegawa, Kenji Imamura and Genichino Kikui. NTT Cyberspace Laboratories, NTT Corporation.

２．検索精度を向上させるための従来手法

検索精度を向上させるためにこれまで多くの手法が提案されている。中でも代表的な手法は適合性フィードバックである。適合性フィードバックは、検索文書の上位にランキングされた文書の中で、人手により適合／不適合を判定することにより、適合する文書に含まれる単語を新たに検索質問に追加する検索質問拡張と、検索質問中の単語の重みを修正する単語の再重み付けにより検索質問の修正が行われる[2]。

検索質問に含まれる単語が少ない場合には、検索質問を拡張することが有効である。検索質問拡張には、同じ概念を有する単語を検索質問に追加するためにシソーラスやオントロジーを用いる方法も提案されている。しかしながら、この方法はシソーラスやオントロジーを人手によって維持していくコストが大きいことが問題である。人手をかけない方法として、検索質問に含まれる単語とその単語を含むランキング上位の文書全体に含まれる単語のうち、相互情報量に基づく相関の高い単語に重み付けし、それを検索質問に追加して再検索する方法が提案されている[3]。しかしながら、この方法で対象にしている文書は論文抄録であるのに対し、サーチエンジンで検索される文書はブログなどの CGM（Consumer Generated Media）を含む。このため、１つの文書には様々なトピックについて記載されている可能性が高く、提案されている方法をそのまま適用することは難しいと思われる。また、サーチエンジンにおける人手を介した検索質問拡張は、事例が少ないものの、必ずしも正解ページのランキングを上げることにはならなかったという報告もある[4]。

我々は、サーチエンジンに対しては、文書全体を対象とした相関の高さではなく、検索質問に関連する部分に限定した相関の高さに基づいて検索文書をリランキングする方が有効ではないかと考え

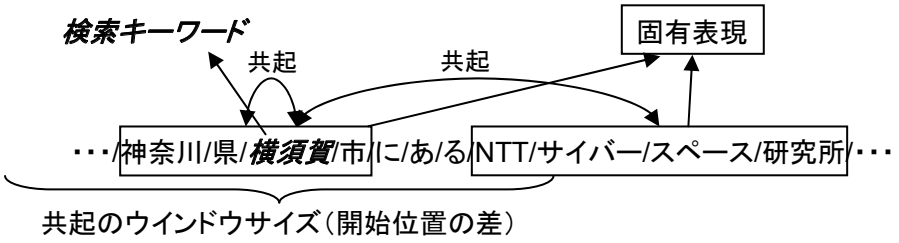

図１ 検索キーワードと固有表現の共起関係の例

ている。さらに、検索質問拡張の対象は一般的には単語であり、文書の内容を把握する手がかりとして有効な複合語や固有表現は用いられていない。複合語や固有表現を用いることは、検索質問に適合する文書の検索に有利であろうと考えている。

３．コンテキストの相関によるリランキング

本研究では、サーチエンジンのように検索質問が１語の検索キーワードでしか与えられない場合であっても、検索文書のランキング精度を向上させる方法を提案する。キーとなるアイデアは、検索キーワードと相関の高い単語や固有表現を用いて、検索キーワードを含む文書のリランキングを行うことである。特に、固有表現は文書の内容を規定する重要な手がかりとなるだけでなく、検索意図を明確にするためにも有利に働くと考えている。例えば、検索質問がある俳優を指す人名の場合には、その俳優と共演している俳優の人名、その俳優の出身地である地名、その俳優が所属している事務所の組織名、その俳優が出演している映画タイトル等の固有物名は、検索意図に適合する文書を検索する手がかりとして有利であろう。近年では固有表現抽出技術が実用的になってきており、固有表現抽出器を用いて固有表現を機械的に抽出することのメリットは大きい。

本稿で提案する手法は下記のステップからなる。

【ステップ１】検索質問で与えられる検索キーワードを含む文書の集合から、文書中に出現する検索キーワードとその周辺に共起する単語および固有表現との相互情報量を計算する。

【ステップ２】文書に含まれる検索キーワードの周辺に共起する各単語や各固有表現の頻度に応じて、検索キーワードと各単語や各固有表現との間の相互情報量を文書のスコアに加算する。

【ステップ３】検索キーワードの周辺に共起する単語や固有表現により文書に付与されたスコアと、検索キーワードの頻度に基づく文書のスコアとの線形和に従って、文書のリランキングを行う。

提案手法は厳密に言えば検索質問拡張ではないが、検索質問以外の単語や固有表現を文書のスコア付けに用いるという点で検索質問拡張に極めて近い。以下では、ステップごとに詳細を説明する。

３．１ 検索質問とコンテキストとの相関

固有表現は IREX の定義に従えば、人名、地名、組織名、固有物名、日付表現、時間表現、金額表現、割合表現というクラスに分類される。本研究では、対象を人名、地名、組織名、固有物名のインスタンスとするが、クラスの情報は用いない。

検索対象とする文書集合がブログなどの場合には、１文書に１つのトピックだけが記述されているとは限らず、様々なトピックが含まれることも想定した方がよいであろう。そこで、検索キーワードと固有表現の共起については、あらかじめ規定したウインドウサイズの範囲内に検索キーワードと固有表現が同時に出現する場合に限定する。検索キーワードも固有表現も１単語で構成されるとは限らないので、検索キーワードと固有表現のそれぞれの開始位置の差が、あらかじめ規定したウインドウサイズの範囲内であれば、これらは共起関係にあるとみなす。図１の例のように、あらかじめ規定したウインドウサイズの範囲に固有表現が完全には含まれていない場合や、検索キーワードが固有表現の内部に含まれている場合も共起関係にあるとみなす。

検索キーワードと固有表現の相関を測る尺度として、相互情報量 (Point-wise Mutual Information) を用いる。検索キーワード Q と固有表現 E との間の相互情報量を次式によって定義する。

$$PMI(Q,E) = \ln \frac{C(Q,E) * N}{C(Q) * C(E)}$$

ここで、$C(Q,E)$は検索キーワード Q と固有表現 E の共起頻度、$C(Q), C(E)$は検索キーワード Q の出現頻度と固有表現 E の出現頻度をそれぞれ表す。N は文書集合における全単語数である。
固有表現と同様に、検索キーワードの周辺に出現する単語についても検索キーワードとの相互情報量を計算する。ただし、検索キーワードと共起する単語は、あらかじめ規定したウインドウサイズの範囲に出現した単語とする。

３．２　相互情報量に基づく文書のスコア

検索キーワードの周辺に相関の高い単語や固有表現がたくさん共起する文書は、検索質問に適合するという仮説を立てた。この仮説に従って、検索キーワードの周辺に共起する単語および固有表現の相互情報量に基づいて文書にスコアを与え、文書の単語数で正規化する。検索キーワード Q に対する文書 D のスコアを次式によって定義する。

$$Score(D,Q) = \sum_i Co(D,Q,E_i) * PMI(Q,E_i) * 1/N_D$$

ここで、$Co(D, Q, E_i)$は文書 D において検索キーワード Q と単語または固有表現 E_i が共起する頻度を表し、N_Dは文書 D の単語数を表す。
ただし、固有表現に関してはウインドウサイズの範囲内のすべてを用いるのではなく、残差IDF(RIDF)[5]を用いてフィルタリングを行う。これは、残差 IDF が小さい固有表現はノイズになりやすいと考えたからである。残差 IDF は、ポアソン分布から推定される IDF 値と実際の IDF 値との差を用いて、単語の重要性を評価する尺度である。ポアソン分布からの推定値は単語の文書集合全体における出現頻度から求めることができる。文書の内容を把握する手がかりになりにくい一般語は、たとえ IDF 値が高くてもポアソン分布からの推定値は IDF 値に近くなるため、残差 IDF 値は小さくなるという性質を持つ。一方、文書の内容を把握する手がかりになる内容語は、文書内で繰り返し用いられるのでポアソン分布からの推定値とは乖離が大きくなるので、残差 IDF 値は大きくなることが知られている。この理由は、内容語は同一文書内で繰り返し用いられるので、事象が独立に生起するというポアソン分布の仮定が崩れるためとされる。残差 IDF 値が 0 以下になる固有表現は検索キーワードとの共起関係から排除することとする。固有表現 E_i の残差 IDF 値は次式により求める。

表１　評価実験におけるデータ

項目	データ数
検索質問数	93
正解文書数	436
全文書数	76639
平均単語数	329.9
平均固有表現抽出数	11.5

$$RIDF(E_i) = \log\frac{n}{n_i} + \log\left(1 - e^{-\frac{F_i}{n}}\right)$$

ここで、n は全文書数、n_i は固有表現 E_i を含む文書数、F_i は文書集合における固有表現 E_i の出現頻度をそれぞれ表す。

３．３　文書のリランキング

文書のスコアは、文書の単語数で正規化した検索キーワードの出現頻度に基づくスコア(Score1)と、検索キーワードと共起する単語の相互情報量に基づくスコア(Score2)と、検索キーワードと共起する固有表現の相互情報量に基づくスコア(Score3)から、各々の文書のスコアを正規化した上で次式に示すような線形和を求める。

$$Score = \alpha * Score1 + \beta * Score2 + \gamma * Score3$$

この値に従って検索キーワードを含む文書のリランキングを行う。

４．評価

本手法の有効性を検証するために、評価実験を行った。評価実験に用いるデータについて表１にまとめる。検索質問の内訳は、一般的な単語や複合語が 34 件、人物のフルネームが 29 件、特定の企業名やサイト名が 13 件、人名や組織名以外の固有表現が 7 件、その他 10 件であり、すべてキーワード１語である。各検索質問には人手により適合と判定された正解文書が 1～19 文書(平均 4.7 文書)が対応している。文書は携帯電話向け Web ページのうち、検索キーワードを含む文書のみを準備した。固有表現については、固有表現抽出器[6]を実行することにより機械的に抽出した。今回用いた文書集合における固有表現抽出の精度については未測定であるが、文書のスタイルや内容が近いと思われるブログにおいて、F 値 75.5(再現率 72.3、適合率 79.0) を達成している。
検索キーワードと単語および固有表現の共起を規

表２ 評価結果

	ベースライン(1)	周辺単語(2)	周辺NE	周辺NE(RIDF)(3)	(1)(2)(3)の線形和
MRR	0.15	0.11	0.14	0.14	0.16
MAP	0.095	0.071	0.085	0.086	0.098
成功率(3位以内)	0.14	0.12	0.14	0.14	0.18

定するためのウインドウサイズを経験的に検索キーワードの前後８単語に設定した。また、文書集合全体における出現頻度が５回以下の単語や固有表現は共起の対象から除外した。文書のスコアを最終的に決定する重みα，β，γは経験的にそれぞれ 0.45,0.25,0.3 と決定した。

検索キーワードと共起する単語や固有表現を考慮することの有効性を計るために、文書の単語数で正規化した検索キーワードの出現頻度によるスコア(Score1)のみを用いたランキングをベースラインとした。

評価尺度は情報検索の評価によく用いられる MRR(Mean Reciprocal Rank)と MAP(Mean Average Precision)を用いた。MRR は正解の中で最も上位にランキングされた正解のみに着目する評価尺度であるのに対し、MAP はすべての正解のランキングに着目する評価尺度である。また、携帯電話の小さな画面に検索結果を表示することを考慮し、ランキング上位３位以内の成功率という評価尺度も用いた。これは、ランキング上位３位以内にひとつでも正解が含まれていれば成功とし、全検索質問数に対する各検索質問の成功回数の割合を評価尺度として用いた。

ベースライン(1)、周辺の単語のみ(2)、周辺の固有表現のみ、残差 IDF を用いた周辺の固有表現のみ(3)、(1)(2)(3)の線形和による組み合わせの評価結果を表２に示す。ベースラインが MAP、MRR において周辺の単語のみによる方法や周辺の固有表現のみによる方法よりも良かったが、線形和による組み合わせでは、MAP、MRR、成功率ともベースラインを上回る結果になった。ウインドウサイズは検索キーワードの前後８単語の場合が最も精度が良かったので、検索キーワードとの共起関係をその周辺の単語や固有表現に限定することには意義があったと考えられる。また、残差 IDF 値が 0 以下の固有表現の除外により、僅かながら MAP が改善したのではないかと考えられる。

検索キーワードの周辺の固有表現のうち、残差 IDF 値が正の値のものの相互情報量を用いて文書にスコアを付けたときに、どの程度の割合の文書にスコアが付くのかを調べた。その結果、46.9%の文書にしかスコアが付いていないことがわかった。共起のウインドウサイズは検索キーワードの前後８単語が最も精度が良かったので、ウインドウサイズを広げることは精度の低下を招く。このため、さらに文書のランキングの精度を向上させるためには、固有表現抽出の再現率を上げることや、固有表現以外にも手がかりにできる複合語を選別することが今後の課題であると考えている。

５．おわりに

本稿では、検索精度の向上を目的として、検索キーワードとコンテキストとの相関に基づく文書のリランキングの方法を提案した。キーとなるアイデアは、1)コンテキストとして文書中の検索キーワードの周辺に着目すること、2)単語だけでなく固有表現との相互情報量も利用することである。今後はもっと大規模なデータを用いて評価・検定を行い、本手法の有効性を検証していきたい。

参考文献

[1] 山名、近藤：解説：サーチエンジン Google, 情報処理 42 巻 8 号、2001.

[2] 北、津田、獅子堀：情報検索アルゴリズム、共立出版、2002.

[3] 金谷、梅村：相関係数を用いた実証的重みの分析と検索質問拡張、情報処理学会　研究報告 IPSJ SIG Technical Report、 2003-FI-13、2003.

[4] 高見、田中：類似性を考慮したスニペットの再生成による検索結果のリランキング、日本データベース学会 Letters Vol.6, No.1, 2007.

[5] M. Yamamoto, K. W. Church: Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substring in a Corpus, Computational Linguistics, Vol.27, No.1, 2001.

[6] 齋藤、鈴木、今村：CRF を用いたブログからの固有表現抽出、言語処理学会第 13 回年次大会発表論文集、2007.