E-017

単語の適合性フィードバックを用いたクエリの拡張手法の提案

Proposal of a Query Expansion Method Using Relevance Feedback for Words

鈴木　永史郎†　　杉本　徹†
Eishiro Suzuki　　Toru Sugimoto

１．研究背景と目的

事実や出来事について調べる時，Web 検索を用いることがある．この時，ユーザは調べたい情報に関するクエリを入力するが，クエリとして与える情報が少ないと目的に合った情報が見つからない場合がある．この場合，ユーザはより詳しいクエリの入力を試みるが，必ずしも検索結果を絞り込めるクエリを入力できるとは限らない．

これに対して，クエリに加えるべき単語をシステムによって生成する研究がいくつか行われてきた．加える単語の一般的な生成手法として，知りたい情報と関連する文書をユーザに選択してもらう適合性フィードバック[1]が用いられる．適合性フィードバックを用いた研究として金子らの研究[2]がある．金子らはユーザが選択した関連文書を教師データとして，Random Forest により学習を行うことで，ユーザが評価していない文書に対しても適合・不適合の判定を行った．このデータを用いた再検索によって，高精度な検索結果のランキングが作成できることを示した．一方で，関連していない文書の中にも関連のある単語が出現する場合があるため，有用な情報が切り捨てられてしまうことがある．

本研究では，ユーザの知りたい情報が単語，特に固有表現である状況を対象とし，単語の適合性フィードバックを利用したクエリの拡張手法を提案する．また，システムによって回答の候補を提示することを目指す．

２．システムの概要

本研究で提案するシステムの処理の流れを図１に示す．

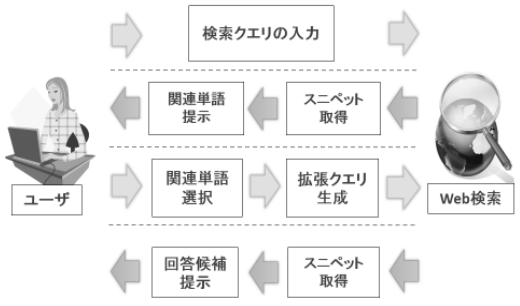

図 １　システムの概要

まず，ユーザに知りたい単語の固有表現の種類を人名，地名，組織名，固有物名の 4 種類から選択してもらう．次に，クエリを入力してもらい Web 検索によって得られた検索結果の上位 100 件から，スニペットを得る．スニペットからユーザに提示する単語を抽出して提示し，ユーザは提示された単語から関連単語を選択する．選択された関連単語を用い，元のクエリに新たに単語を加えた拡張クエリを生成する．最後に，拡張クエリを用いて再び検索を行い，検索結果の上位 100 件のスニペットから回答候補を抽出し順位付けを行いユーザに提示する．本研究では，Web 検索で用いる検索エンジンとして Yahoo!を用いた．

３．システムの処理

３．１　単語の抽出

３．１．１　整形

Web 検索によって得られたスニペットから抽出した単語に対し，整形処理として関連単語及び回答として適さない不要な語の除去を行う．不要な語は，助詞などの機能語や，ユーザが入力したクエリに含まれる単語とする．

３．１．２　固有表現の抽出

ユーザの知りたい単語を固有表現としているため，これを抽出する必要がある．よって，係り受け解析ツール CaboCha[3]を用いて抽出を行う．CaboCha では IREX にて定義される 8 種類の固有表現を抽出可能だが，本研究ではその内の人名，地名，組織名，固有物名の 4 種類のみを対象として抽出を行う．

３．１．３　複合名詞の抽出

「携帯電話」のように「携帯」と「電話」が連なって出現することで意味を成す単語があるため，これも抽出する．よって，名詞が 2 語以上連続して出現した場合，名詞を連結する処理を行うことで複合名詞として新たに単語を得る．この時，連結途中の単語及び連結前の単語も 1 つの単語として扱うこととした．

３．２　関連単語の提示

2 つの方法で行う．1 つは単語ごとにスコアを算出し，これが高い単語を関連性の高い単語として順位付けを行ったものを提示する．スコアの算出には，単語の出現頻度と出現したスニペットの順位を考慮し，出現頻度が高く上位のスニペットに出現した単語のスコアが高くなるようにする．スニペット中の単語 w のスコア score(w) は，得られたスニペットの数を n，そのうち w を含むスニペットの数を k，それらのスニペットの検索された順

† 芝浦工業大学，Shibaura Institute of Technology

位をそれぞれr_1, r_2, …, r_kとして以下の式によって得られる.

$$\mathrm{score(w)} = \sum_{i=1}^{k} (n - r_i + 1)$$

2 つ目の方法は，ユーザが選択した知りたい単語の固有表現の分類を用いて，これが同じ種類の単語がスニペット中に出現し，この後に助詞が続いた時，助詞の後に続く単語を関連単語としてユーザに提示する．これは，選択される関連単語について調査したところ，正解の後に助詞が続いた時，その後に続く単語が関連単語として選択される傾向にあるという結果に基づく．

３．３　拡張クエリの生成

ユーザが入力した元のクエリに，選択された関連単語の１つを追加することで拡張クエリを生成する．関連単語が複数選択された場合，その関連単語の数だけ拡張クエリの生成を行う．

３．４　回答候補の提示

生成された拡張クエリごとに検索を行う．回答候補の抽出には 3.2 節で述べたスコアを求める式を用いた．検索によって得られたスニペットを用いてスコアを求め，順位付けを行った回答候補を提示する．

４．評価実験と考察

提案手法の有効性を検証するため，テストデータを用いて本手法の回答精度を求める実験を行った．また，比較のためフィードバックを行わない通常の検索と既存のフィードバック手法による検索についても精度を求めた．既存のフィードバック手法は文書に対する適合性フィードバックとした．クエリに追加する単語は得られた関連文書に対し形態素解析を行い，単語の出現頻度を求めることでこの値が最も高い単語を用いた．通常の検索及び既存手法の回答候補の提示には 3.2 節におけるスコアを求める式を用いて順位付けを行った．また，単語の抽出条件は 3.1 節で述べた処理に準ずる．提案手法となる単語のフィードバックでは，ユーザが最も関連性が高いとした単語を１つ用いて回答候補を得ることとした．

各手法の精度は MRR(Mean Reciprocal Rank)を求めることで，この値が大きいほど精度が高いとして評価を行う．MRR は質問数 q，i 番目の質問に対する回答候補において正解が出現した順位r_iから，以下の式によって得られる．

$$\mathrm{MRR} = \frac{1}{q}\sum_{i=1}^{q}\frac{1}{r_i}$$

また，実験で使用するテストデータは，知りたい単語とこれを検索するためのクエリのセットの 20 個を用いた．知りたい単語は，ユーザが選択する 4 種類の固有表現のいずれかに分類される単語とし，固有表現の種類ごとに知りたい単語とクエリのセットを 5 個ずつ用意した．使用したデータの一部を表 1 に示す．

表 1. 使用したデータの例

固有表現の分類	回答	クエリ
人名	伊藤博文	初代、総理大臣
地名	甲子園	野球、大会
組織名	Apple	iPhone、企業
固有物名	シリコンバレー	IT、企業、半導体、アメリカ

上記のデータを用いた検証結果として，表 2 に各手法における MRR の値を示す.

表 2. 各手法における回答精度

手法	MRR
フィードバックなし	0.261
文書フィードバック	0.393
単語フィードバック(提案手法)	0.414

表 2 より，提案手法は既存手法より高い精度が得られた．この要因の１つとして，提案手法ではクエリに追加する単語として特定性の低い単語が選ばれづらい傾向にある点が挙げられる．例として，表 1 における「シリコンバレー」では，既存手法における追加する単語は「メーカー」となっており，正解の順位にほとんど影響を与えていなかった．提案手法では，「ベンチャー」といった特定性が高い単語が選択されており，正解の順位を大きく上げることができた．一方で，提案手法では正解が関連単語として選択されることはなかったが，既存手法では正解が追加される単語となることもあり，後者の方が正解の順位を大きく上げる場合もあった．また，ユーザが選択した関連単語において，その単語が提示された順位は必ずしも高い順位ではない場合もあるため，出現頻度と出現位置によるスコアの算出ではユーザが考える関連の度合いを反映することが難しいこともあった．

５．まとめと今後の展望

本研究では，単語の適合性フィードバックを利用したクエリの拡張手法を提案し，評価実験を行うことで，既存手法よりも高い精度で回答を提示できると確認できた．

今後の展望としては，ユーザがより関連していると考える単語が回答候補の上位となるよう，特定性が高いと考えられる単語に対し，求めたスコアに重みを付加するよう式を改良したい．

参考文献

[1] Christopher D. Manning et al., “Introduction to Information Retrieval”, Cambridge University Press, 2008
（岩野和生他（訳）“情報検索の基礎”，共立出版，pp.157-172，2012）

[2] 金子弘明，梅澤猛，大澤範高：適合性フィードバックにおけるユーザ負荷軽減手法，情報処理学会研究報告，2013-NL-214(3)，pp.1-8，2013

[3] CaboCha:http://code.google.com/p/cabocha/