文脈関連度による検索質問の関連語抽出

佐々木 稔†‡，北 研二‡
†SigMatics Inc.　‡徳島大学工学部
E-mail: sasaki@sigmatics.co.jp
{sasaki, kita}@is.tokushima-u.ac.jp

1　はじめに

近年，インターネットの普及とともに，個人でWWW (World Wide Web) を代表とするネットワーク上の大量の電子データやデータベースが取り扱えるようになり，テキストデータの山の中で必要な情報を取り出す機会が増加している．しかし，このようなデータの増加は必要な情報の抽出を困難とする原因となる．この状況を反映し，情報検索，情報フィルタリングやクラスタリングといった技術が注目を集め，過去数十年の間に新聞記事などの文書を対象とした研究が盛んに進められ，高速な文字列検索アルゴリズムや自動索引づけなどに多くの成果が得られている．

ユーザが自分の検索要求を表現するためによく使われるのが自然言語であり，LycosやGooのようなインターネット上にあるWWWサイトの検索エンジンなどで検索を行う場合，ユーザは自分の検索要求を少ない数の索引語からなる検索質問として表現している．しかし，その検索要求をユーザが正確に索引語として表現できる場合もあるが，時としてユーザの意図している索引語が見つからずに，言語化した検索質問の意味内容をユーザが表現できない場合もある．また，情報検索システムでは，検索質問と文書中の索引語が一致することにより検索が行われ，言い換え表現などのような概念に対して表現の多様性を考えることなしに，字面での検索が行われてしまうという問題が生じる．

このような問題を解決するために，与えられた検索質問に対して関連性のあるタームの集合を文書集合の中から自動的に抽出し，検索質問に拡張する検索質問拡張 (Query Expansion) の研究が盛んに進められている．たとえば，「減税」関連のあるタームとして，「所得」，「消費」，「税率」，「増税」などを考えることができる．このようなタームは「減税」の表す意味や概念的な考え方が同一であるとはいえないのであるが，何らかの関連性を持ち，減税について書かれている文書にこれらの関連語が含まれている可能性が高いと考えられる．

本報告では，このような関連語抽出手法のひとつとして提案されている文脈関連度 (Contextual Document Relevance)[4] を用いて検索質問を拡張する情報検索システムを構築し，検索実験を行い，この手法の改良点などの提案を行う．

2　文脈関連度

ユーザの与えた検索質問と適合する文書における，タームの出現頻度を解析した文脈関連度を計算するアルゴリズムを示す．文書 d_i は，その文書に出現するターム t_j の重み w_{ij} を要素とする文書ベクトル $d_i = (w_{i1}, w_{i2}, \cdots, w_{it})$ で表される．また，検索質問 Q も同様に，検索質問に出現するターム t_j の頻度 q_j を要素とする検索質問ベクトル $Q = (q_1, q_2, \cdots, q_t)$ として表される．

文脈関連度の計算は，まず検索対象となる全文書の検索質問に対する適合度，すなわち，類似度の計算を行い，検索質問に対する適合文書を検索する．検索質問と文書の類似度は2つのベクトルの余弦とし，次式により類似度が求まる．

$$rel(Q, d_i) = \frac{Q \cdot d_i}{|Q| \cdot |d_i|} = \frac{\sum_{j=1}^{t} w_{ij} q_j}{\sqrt{\sum_{j=1}^{t} {w_{ij}}^2 \sum_{j=1}^{t} {q_j}^2}} \quad (1)$$

次に，検索質問に対して各文書の持つ類似度がある一定の閾値を超えるとき，その文書に出現しているターム全てに，検索質問に対する関連性を与えるための重みを付与する．ターム t_j がある文書に出現している場合，以下のように，その文書におけるタームの重みと，その文書の検索質問に対する類似度の積を取ったものが，検索質問中に存在するひとつのタームの関連度となる．

$$cdr(Q,t_j)=\sum_{i=1}^{n} w_{ij} rel(Q,d_i) \qquad (2)$$

しかし，この計算はタームがどの文書にも出現するようなものであると，検索質問とタームが関連あるかどうかに関わらず，そのキーワードには高い重みが与えられることになる．このような場合を考慮し，正規化を行う必要があり，以下のようなキーワードの重みの和を考える．

$$df_j=\sum_{i=1}^{n} w_{ij} \qquad (3)$$

これより，検索質問に対するタームの正規化した関連度は以下のようになる．

$$ncdr(Q,t_j)=\frac{\sum_{i=1}^{n} w_{ij} rel(Q,d_i)}{\sum_{i=1}^{n} w_{ij}} \qquad (4)$$

この関連度が高いほど検索質問に関連のあるタームとなり，この値の高い順にいくつかのタームを取り出す事により検索質問の拡張が行われる．

3 実験

文脈関連度の関連語抽出性能を調べるために，評価用テストコレクションである BMIR–J2[3] を用いて関連語を抽出する実験を行った．BMIR–J2 は毎日新聞 94 年のデータの中から 5080 記事を検索対象に選び，60 個の検索要求とそれらの関連記事が用意されたテストコレクションで，検索対象の新聞記事全体で約 6 メガバイトの容量を持っている．

まず，前処理として検索対象のデータに対し，茶筅を用いて形態素解析を行い，名詞，または，アルファベットと判定された単語のみをタームとして用いた．名詞以外の品詞と判定されたもの，または，名詞であるが数として判定されたものについては，ベクトル作成には必要ない機能語であるものとして，今回の実験ではこれ以上扱わないことにした．この前処理の結果，15114 個の単語がタームとして抽出された．

これらのタームを要素とする文書ベクトルを作成するとき，タームの頻度に重みを加えた数値をベクトルの要素とする．数多く提案されている重みづけ手法で，今回の実験では以下の式で定義された対数エントロピー重み [1] を用いた．L_{ij} は j 番目の文書に対する i 番目のタームへの重み，G_i は文書全体に対する i 番目のタームへの重みを表す．

$$L_{ij}=\begin{cases} 1+\log f_{ij} & (f_{ij}>0) \\ 0 & (f_{ij}=0) \end{cases} \qquad (5)$$

$$G_i=1+\sum_{j=1}^{n} \frac{\frac{f_{ij}}{F_i}\log\frac{f_{ij}}{F_i}}{\log n} \qquad (6)$$

ここで，n は全文書数，f_{ij} は j 番目の文書に出現する i 番目のタームの頻度，F_i は文書集合全体における i 番目のタームの頻度を表す．これより，j 番目の文書から得られる文書ベクトルの i 番目の要素 d_{ij} は，

$$d_{ij}=L_{ij}\times G_i \qquad (7)$$

となる．

得られたターム·文書行列から先に述べた文脈関連度を計算するために，まず検索要求を検索対象と同様に形態素解析を行い，数以外の名詞やアルファベットと判定された単語のみをタームとして抽出する．抽出されたタームから検索要求ベクトルを計算し，検索対象となるそれぞれの文書ベクトルとの余弦を計算し，検索要求と文書との類似度を計算する．文脈関連度を計算する際の閾値として，類似度が 0.1 以上の文書に対し，検索要求に含まれるタームのひとつと検索対象に含まれるすべてのタームとの関連度を計算し，その和を正規化したものが検索質問に対するタームの文脈関連度となる．

例として，「航空大手 3 社」という検索質問に対して，文脈関連度の高いタームを順に並べたものを，表 1 に示す．この表から分かるように，正規化する，しないに関係なく，文脈関連度は検索要求に関連のあるタームをある程度抽出することができた．しかし，正規化を行った場合，

表 1: BMIR–J2 における検索質問「航空大手 3 社」と文脈関連度の高いターム

正規化あり		正規化なし	
ターム	関連度	ターム	関連度
夏型	0.13814	社	2.41222
ＡＮＡ	0.13260	航空	2.11729
キャセイ	0.11998	大手	1.52279
ＪＡＬ	0.11129	各社	0.96159
ＪＡＳ	0.11081	機	0.85183
ユナイテッド	0.10719	国内	0.77307
ボ	0.10543	決算	0.75304
既	0.09012	メーカー	0.74870
カンタス航空	0.09012	会社	0.74586
ミニストップ	0.08862	電機	0.73532
ノース	0.07966	航空機	0.69333
エコノミー	0.07927	スチュワーデス	0.66970
京王	0.07891	事業	0.62636
航空審議会	0.07600	価格	0.59208
全日本空輸	0.07553	期	0.55645

検索対象となる文書集合全体に出現するタームは検索要求に対する文脈関連度が小さくなるために，全体的に頻度の少ないカタカナ語やアルファベットによる略語表記に高い文脈関連度が与えられる結果となった．この場合，このような語が関連語として十分に抽出されているが，「航空機」などのような文書全体を通して一般的なタームは「航空大手 3 社」に関連のあるタームとしてあまり抽出されなかった．

これに対し，正規化を行わない場合，検索対象となる文書集合全体に出現するタームの重みが変らないため，「会社」，「電機」などといった一般的なタームを抽出することができた．しかし，正規化を行った場合に多く得られる略語やカタカナ語が関連語として抽出されていないので，正規化を行わない場合の方がもっともらしい関連語が得られると考えられる．また，BMIR–J2 のデータ数が少ないために正規化をしていない場合の結果がもっともらしく，正規化をした場合の結果ように，頻度の少ない数字のタームが関連語として抽出されているということも考えられる．

このようにして得られた関連語の内，文脈関連度の上位 10 タームを加えることによって検索質問を拡張し，再度検索を行った結果を図 1 に示す．このグラフにおいて，'VSM' は検索質問拡張を行わない場合の検索結果，'CDR' は正規化を行わない，式 (2) を用いて検索質問拡張を行った場合の検索結果，'NCDR' は正規化を行う，式 (4) を用いて検索質問拡張を行った場合の検索結果を表している．この実験の結果，検索質問に関連のあるタームを拡張する前後において，検索性能はほとんど変わらない結果であった．

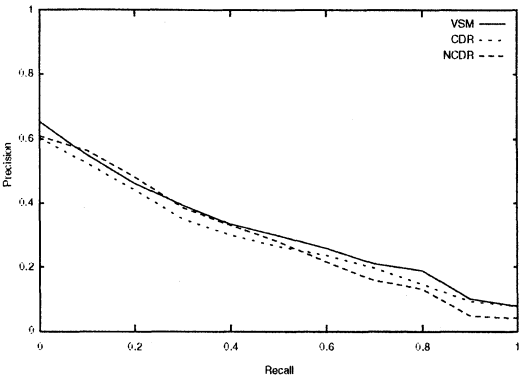

図 1: BMIR–J2 における再現率–正解率曲線

また，同様の実験を情報検索システムの評価用テストコレクションである MEDLINE についても行った．MEDLINE は医学・生物学分野における英文の文献情報データベースで，検索の対象となる文書の件数は 1033 件で，約 1 メガバイトの容量を持つテキストデータである．また，MEDLINE には 30 個の評価用検索要求文とそれらの関連記事が用意されている．まず，前処理として MEDLINE の記事全体から抽出した 1033 件の記事から一般的な 439 個の英単語をストップワードに指定して，文書の内容と関係のほとんどない単語は削除した．この前処理の結果，4329 個の単語がタームとして抽出された．これらのタームを要素とする文書ベクトルを作成するとき，タームの頻度に対する重みは BMIR–J2 と同じ重み付け手法を用いた．

このようにして得られたタームからターム·文書行列を作成し，検索要求ベクトルとの類似度が 0.1 以上の文書に対して，先に述べた BMIR–J2 と同様に，文脈関連度の上位 10 タームを加えることによって検索質問を拡張し，再度検索を行った結果を図 2 に示す．このグラフにおいて，'VSM' は検索質問拡張を行わない場合の検索結果，'CDR' は正規化を行わない，式 (2) を用いて検索質問拡張を行った場合の検索結果を

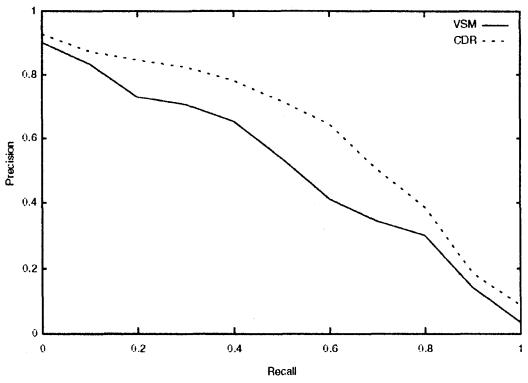

図 2: MEDLINE における再現率–正解率曲線

表している．MEDLINE の場合については，グラフからも分かる通り，検索質問を拡張する前に比べて大幅に検索性能が向上していることが分かる．また，BMIR–J2 の実験結果と比較して検索質問拡張の効果が顕著に表れていることも分かる．

4 おわりに

本報告では，検索要求と検索対象との類似度から検索要求に関連のあるタームを抽出する手法を用いた情報検索システムを提案し，その検索性能を調査するために，BMIR–J2 および MEDLINE を用いて検索実験を行った．その結果，見た目での判断であるが，ある程度まで検索要求に対する関連語を抽出できることからも，さらに改良することによってより精度の高い関連語を抽出できるのではないかと考えられる．また，この手法を用いて検索質問拡張実験を行った結果，テストコレクションとして BMIR–J2 を用いた検索では結果がほとんど変わらなかったが，MEDLINE を用いた場合については検索結果が改善されることが分かった．これにより，この手法が検索質問拡張に有効であると考えられる．

今後の課題としては，まずテストコレクションとして用いた BMIR–J2 からのターム抽出をより検討する必要がある．今回は数以外の名詞やアルファベットを索引語として用いたが，接尾語や接頭語など，抽出した名詞に接続した方がより検索に有効なタームになると考えられる．より有効なタームの選択，抽出に重点を置いて改良を加えていきたい．

また，検索要求と文書との類似度計算をより高い精度で計算する手法を用いて，関連文書に対する類似度を上げることで，関連語を抽出できる可能性が高いのではないかと考えられる．その方法として，Latent Semantic Indexing(LSI)[2] や Concept Projection[5] などを用いて，より関連性の高いタームにより高い重みを付けるといった処理が挙げられる．また，タームの関連度を求める際に計算するベクトル間の類似度にかけるタームの要素を，数字など関連のないタームが抽出されないような，より精度の高い重み付けが必要となる．考えられる方法としては，単語間の関連性を用いる事が挙げられる．この単語間の関連性を求める方法に関しては索引語抽出を目的とした種々の方法が提案されている．このような単語間の関連性を基にしたタームの重み付けを行うことにより，より関連のあるタームが得られるのではないかと考えられる．さらに，Local Context Analysis[6] など本手法と同様な検索質問拡張手法を用いて検索実験を行い，本手法との検索性能の違いを調べる必要がある．

参考文献

[1] Erica Chicholm, Tamara G. Kolda: "New Term Weighting Formulas for the Vector Space Method in Information Retrieval", Technical Memorandum ORNL-13756, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 1998.

[2] Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, Richard A. Harshman: "Indexing by Latent Semantic Analysis", Journal of the Society for Information Science, 41(6), pp. 391-407, 1990.

[3] 木谷 強ほか: "日本語情報検索システム評価用テストコレクション BMIR–J2", 情報処理学会研究会報告 98-DBS-114-3, pp. 15-22, 1998.

[4] Kai Korpimies and Esko Ukkonen: "Searching for General Documents", In Proc. of the 3rd International Conference on Flexible Query Answering Systems, FQAS '98, Lecture Notes in Artificial Intelligence 1495, pp. 203-214. Springer-Verlag.

[5] 佐々木 稔, 北 研二: "ランダム・プロジェクションによるベクトル空間情報検索モデルの次元削減", 自然言語処理, Vol. 8, No. 1, 2001.

[6] Jinxi Xu, W. Bruce Croft: "Query Expansion Using Local and Global Document Analysis", In Proc. of the 19th International Conference on Research and Development in Information Retrieval (SIGIR 96), pp. 4-11, 1996.