情 報 学 基 礎 68−10
自 然 言 語 処 理 151−10
（2002. 9. 18）

潜在的文脈関連度を用いた検索質問拡張

佐々木 稔† 新納 浩幸‡

†茨城大学 工学部 情報工学科 ‡茨城大学 工学部 システム工学科

†‡〒316-8511 茨城県日立市中成沢町 4-12-1

† sasaki@cis.ibaraki.ac.jp ‡ shinnou@dse.ibaraki.ac.jp

概要

情報検索システムのなかで代表的な情報検索モデルである関連性フィードバックを用いる場合，はじめの検索結果をもとに関連語を抽出し，元の検索質問に加えて再度検索を行う．これにより検索質問の中に含まれている意味の曖昧さが軽減される．一般的にこの手法を用いると，拡張なしの検索質問を用いた場合と比較して拡張後の検索質問を用いた場合検索精度が下がることが報告されている．その原因として，シソーラスに含まれている同義語や類義語の数が少ないことが挙げられる．また，意味の同じような単語を用いたとしても，結果として同じ文書が検索されてしまう可能性も存在している．本稿では，このような関連語に対する問題点を解決するために文書集合内で存在する索引語間の関連性を考慮した潜在的文脈関連度を提案し，提案した手法を用いた検索モデルを構築し，評価用テストコレクションである MEDLINE を利用した検索実験を行い，その有効性を示す．

Query Expansion Using Latent Contextual Document Relevance

Minoru Sasaki† Hiroyuki Shinnou‡

†Department of Computer and Information Sciences, Faculty of Engineering, Ibaraki University

‡Department of Systems Engineering, Faculty of Engineering, Ibaraki University

4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan

† sasaki@cis.ibaraki.ac.jp ‡ shinnou@dse.ibaraki.ac.jp

Abstract

When the relevance feedback, which is one of the most popular information retrieval model, is used in an information retrieval system, a related word is extracted based on the first retrival result. Then these words are added into the original query, and retrieval is performed again using updated query. Generally, Using such query expansion technique, retrieval performance using the query expansion falls in comparison with the performance using the original query. As the cause, there is a few synonyms in the thesaurus and although some synonyms are added to the query, the same documents are retireved as a result. In this paper, to solve the problem over such related words, we propose latent context document relevance in consideration of the relevance between index words in the document set.

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，ユーザが検索したい意味内容を持つ単語を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く，これらの語を補って再度検索をすることでより精度の高い検索をすることができる．

このような研究の中で最もよく知られているもののひとつに関連性フィードバックが存在する．これは入力された検索質問において一度検索を行い，その結果から上位に検索された関連文書を検索質問に加え，さらに関連がないと判断された文書を検索質問から差し引いて再度検索をするものである．この手法の他にもさまざまなものが存在する．例えば，Local Context Analysis などで用いられている疑似関連性フィードバック [12] や関連文書の中で頻繁に出現し，さらにその中で分散している索引語を抽出し，それを絞り込み語として再検索の支援を行う [7] といったものが挙げられる．

しかし，これらの手法は一度検索を行ってその結果をもとに関連語を抽出し，はじめに与えた検索質問に含まれている曖昧さの軽減を行っている．もともと関連性フィードバックは，ユーザが与えた検索質問に対し，検索結果からユーザの情報要求に近づける処理であるが，この処理を行わずに検索質問の曖昧さを軽減する手法として，一般的にシソーラス検索と呼ばれる類義語辞書を用いた検索質問の拡張 [4] が存在する．一般的にこの手法を用いると，拡張なしの検索質問を用いた場合と比較して拡張後の検索質問を用いた場合検索精度が下がることが報告されている．その原因として，シソーラスに含まれている同義語や類義語の数が少ないことが挙げられる．また，意味の同じような単語を用いたとしても，結果として同じ文書が検索されてしまう可能性も存在している．

このような問題点を解決するひとつの方法として，検索に有効な関連語をシソーラスとして用意することで，検索精度が向上するのではないかと考えられる．ユーザの要求する情報を効率よく見つける処理の第1段階として関連語シソーラスを用いることで，少ない数の検索質問に含まれる意味の曖昧さを解消することが可能となり，検索したい内容の特定などの効果が期待できる．しかし，このようなシソーラスを手作業で構築するのは索引語の数が非常に膨大なため手間がかかるという問題が生じる．このため，関連語シソーラスを自動的に構築するシステムが必要となってくる．

本稿では，このような関連語に対する問題点を解決するために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)[6] に対して改良を加え，文書集合内で存在する索引語間の関連性を考慮した，潜在的文脈関連度 (Latent Contextual Document Relevance) を提案する．また，あらかじめ予備検索を行なって関連語を抽出することがなく，最初の検索で関連語が抽出できるように，潜在的文脈関連度を用いて関連語のシソーラスを構築し，これをもとに検索質問を拡張する．このような関連語シソーラス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，検索性能の評価を行う．

2 潜在的意味インデキシング

潜在的意味インデキシング (Latent Semantic Indexing, LSI) はベクトル空間モデルの一種であり，検索対象となる文書集合より得られた索引語・文書行列に対して，特異値分解 (Singular Value Decomposition, SVD) などの行列変換手法を用いて，多次元空間における文書ベクトルの要素を抽象化する．この LSI によく用いられる SVD は，一般に制約条件のない線形最小二乗問題の解や行列の階数や相関を求めるために使われる手法で，直交基底を用いた同値変換に基づく行列の対角化が行われる．

本節では，LSI で用いられる SVD の定義を示し，LSI が索引語・文書行列に対してどのような効果が存在するのかを概説する．

2.1 特異値分解

n 個の文書からなる文書集合から，$m \times n$ である索引語・文書行列 A が得られたとする．このとき，行列 A の階数が r であるとすると，A の特異値分解は次のように定義される．

$$A = U\Sigma V^T \tag{1}$$

ここで，$U = (u_1, \cdots, u_m)$ と $V = (v_1, \cdots, v_n)$ は $U^TU = V^TV = I_n$ を満たすユニタリ行列，Σ は $\Sigma = diag(\sigma_1, \sigma_2, \cdots, \sigma_n)$，$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq \sigma_{r+1} = \cdots = \sigma_n = 0$ を満たす対角行列で，この σ_i $(i = 1 \cdots r)$ は A の特異値と呼ばれる．これらの特異値の値 σ_i により，左特異ベクトル u_i と右特異ベクトル v_i が導き出され，A の i 番目の 3 つ組 $\{u_i, \sigma_i, v_i\}$ が定義される．この 3 つ組を用いることで，行列 A は次のように表すことができる．

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \tag{2}$$

索引語・文書行列 A の特異値分解 $A = U\Sigma V^T$ が得られたとすると，行列 U に含まれる左特異ベクトル $u_1, \cdots, u_m$ が文書空間の正規直交基底となる．また，行列 V に含まれる右特異ベクトル $v_1, \cdots, v_n$ が索引語空間の正規直交基底となる．LSI は行列 Σ に含まれる特異値の小さいものを除くことにより行列 A の次元が削減され，元のランク r よりも低いランク k である近似行列 A_k が得られる．

このように索引語・文書行列 A から近似された索引語・文書行列 A_k が得られたとき，A_k には対象とする文書集合における索引語間の共起情報が計算されており，結果として意味的な関連付けが自動的に行われている．たとえば，「プログラミング」という索引語を含む文書集合を考えたとき，この中に「Java」という索引語が頻繁に存在し，共に関係が非常に深いとする．このとき，索引語「Java」を含む文書中に索引語「プログラミング」が存在していなくても，次元を削減した近似行列を計算すると索引語「Java」に引っ張られる形で，内容的に関係の深い索引語「プログラミング」に比較的高い重みが付けられる．

3 潜在的文脈関連度

ユーザの要求する情報を効率よく見つける処理の第 1 段階として，ユーザが与えた検索質問に対してある基準において概念の似ている索引語を抽出，追加することにより，少ない数の検索質問に含まれる意味の曖昧さが解消される．通常，概念の似ている語には同義語・類義語辞書などの語彙的ツールが用いられる．しかし，このようなシソーラスを手作業で構築することは，索引語の数が非常に膨大であるため手間がかかるという問題がある．このため関連語シソーラスを自動的に構築する方法として，潜在的意味インデキシングによる文脈関連度を提案し，

これを潜在的文脈関連度と名付けることとする．

本節では，まず文脈関連度の計算方法についての概説を行い，この文脈関連度の問題点を指摘し，改良点についての仮説を立てる．この仮説を元に，関連語シソーラスを構築するための尺度である潜在的文脈関連度の計算方法について概説する．

3.1 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，索引語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索引語 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索引語 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．

$$rel(Q, d_i) = \frac{Q \cdot d_i}{|Q| \cdot |d_i|} = \frac{\sum_{j=1}^{t} w_{ij} q_j}{\sqrt{\sum_{j=1}^{t} {w_{ij}}^2 \sum_{j=1}^{t} {q_j}^2}} \tag{3}$$

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索引語全てに，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．索引語 t_j がある文書に出現している場合，以下のように，その文書における索引語の重みと，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索引語の関連度となる．

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

しかし，この計算は索引語がどの文書にも出現するようなものであると，検索質問と索引語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．

$$df_j = \sum_{i=1}^{n} w_{ij} \tag{5}$$

これより，検索質問に対する索引語の正規化した関連度は以下のようになる．

$$ncdr(Q, t_j) = \frac{\sum_{i=1}^{n} w_{ij} rel(Q, d_i)}{df_j} \tag{6}$$

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

3.2 潜在的文脈関連度

上述した文脈関連度は検索質問が与えられたとき，一度類似度の計算を行っている．このように類似度計算をすることによって関連語抽出する手法は，これまでに提案されている関連語抽出に基づいた検索質問拡張手法のそのほとんどにおいて採用されている．この場合，関連語を抽出するためには検索要求と文書の類似度が非常に大きな影響を持っている．このため，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，より文書中の単語の分布を考慮した質の高い関連語を抽出できるのではないかと考えられる．

そのひとつの方法として我々が提案するのが，類似度を計算する際に元の文書ベクトルを用いるのではなく，索引語・文書行列に対して特異値分解を行い，次元を削減した近似行列から得られる文書ベクトルを用いることである．特異値分解を行うことにより，上述のように共起しやすい索引語に対して意味的な関連付けが自動的に行われる．このため文脈を考慮した索引語と文書の関連度をより高い精度で計算することが可能となるのではないかと考えられる．

このことを考慮して，関連語シソーラスを構築するための尺度である潜在的文脈関連度を計算するアルゴリズムについて概説する．前節と同様に，文書 d_i は，その文書に出現する索引語 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，検索対象となる文書ベクトル集合から成る索引語・文書行列を A とする．このとき，行列 A

を特異値分解すると階数が r の場合以下のように表される．

$$A = U\Sigma V^T \tag{7}$$

これより，行列 Σ に含まれる特異値の小さいものを除いて，元のランク r よりも低いランク k である近似行列 A_k を得る．

$$A_k = U_k\Sigma_k V_k^T \tag{8}$$

これをもとにして，文書集合に含まれる索引語 T の t_l に対する潜在的文脈関連度の計算を行う．まず，索引語 T に対応する要素の値のみ 1 でその他は 0 であるベクトル v と，近似行列 A_k から得られる文書ベクトル $d_i^{(k)}$ の余弦を次式のように計算する．

$$crel(v_T, d_i^{(k)}) = \frac{v_T \cdot d_i^{(k)}}{|v_T| \cdot |d_i^{(k)}|} \tag{9}$$

先に述べた文脈関連度と同様に，その文書における索引語の重みとその文書の検索質問に対する余弦の積が潜在的文脈関連度となる．

$$lcdr(T, t_l) = \sum_{i=1}^{n} w_{il} \cdot crel(v_T, d_i^{(k)}) \tag{10}$$

潜在的文脈関連度についても同様に，検索質問に対する索引語について正規化した関連度は以下のようになる．

$$nlcdr(T, t_j) = \frac{\sum_{i=1}^{n} w_{ij} \cdot crel(v_T, d_i^{(k)})}{\sum_{i=1}^{n} w_{ij}} \tag{11}$$

4 実験

本節では，潜在的文脈関連度を用いた情報検索モデルの構築を行い，その検索性能評価として，MEDLINE を用いた検索実験について述べる．

4.1 データ

実験で用いたデータは，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 を利用した．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 1Mbyte 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各要求文に対する正解文書が用意されている．

MEDLINE に含まれている 1033 件の文書全体から，前処理として，“a” や “about” などの一般的な 439 個の英単語を不要語リストに指定して，文書の内容と関係のほとんどない単語は削除した．この後，接辞処理を行い，残った英単語を語幹に変換する処理を行った．この前処理の結果，文書全体に 5526 個あった単語から，4329 個の単語が索引語として抽出され，実験データとして用いた．

4.2 検索実験方法

実験では，MEDLINE から前処理により得られた索引語を要素とする文書ベクトルと検索要求ベクトルを作成し，比較することで検索スコアを計算する．文書ベクトルを作成するとき，ベクトルの要素には局所的，大域的な索引語の分布を考慮するために，索引語の頻度に重み付けした数値が用いられる．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 番目の索引語への重み，G_i は文書全体に対する i 番目の索引語への重みを表す．

$$L_{ij} = \begin{cases} 1 + \log f_{ij} & (f_{ij} > 0) \\ 0 & (f_{ij} = 0) \end{cases} \tag{12}$$

$$G_i = 1 + \sum_{j=1}^{n} \frac{\frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i}}{\log n} \tag{13}$$

ここで，n は全文書数，f_{ij} は j 番目の文書に出現する i 番目の索引語の頻度，F_i は文書集合全体における i 番目の索引語の頻度を表す．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，

$$d_{ij} = L_{ij} \times G_i \tag{14}$$

となる．

このようにして得られた索引語から索引語・文書行列を作成し，その行列に対して特異値分解により次元削減を行い，近似行列を計算する．次に，関連度を求めたい索引語からなるベクトルと近似された

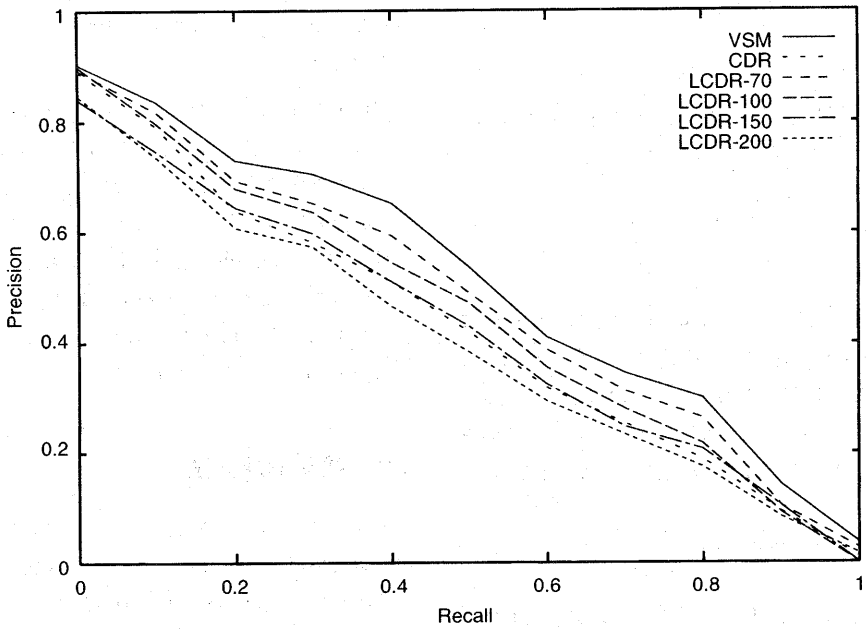

図 1: 正規化なしの場合における再現率–正解率曲線

文書ベクトルとの余弦が 0.1 以上の文書に対して，潜在的文脈関連度の上位に存在する 15 個の索引語を加えることによって検索質問を拡張した．これにより拡張した検索質問に対して再度検索を行い，ベクトル空間モデルにおける類似度である内積により計算された検索スコアのうち，上位 50 文書を検索結果として出力する．

検索システムの評価には，一般的に用いられている適合率 (Precision) と再現率 (Recall) を用いた [5][11].

$$\text{Recall} = \frac{\text{システムが出力した正解文書数}}{\text{全正解文書数}} \quad (15)$$

$$\text{Precision} = \frac{\text{システムが出力した正解文書数}}{\text{システムが出力した文書数}} \quad (16)$$

再現率と適合率は，それぞれ個別に用いて，システム評価を行うことができるが，本実験では，一般にランクづけ検索システムの評価に用いられる再現率・適合率曲線を用い，システムの評価を行った．この曲線は，各質問に対しひとつの曲線が作成されるが，本稿の検索システム評価には，全 30 個の質問に対する各再現率での平均を計算した再現率・適合率曲線を用いた．

4.3 実験結果

潜在的文脈関連度による検索質問拡張を行った検索実験の結果を図 1 に示す．ここで，このグラフは正規化を行わない，式 (10) を用いて潜在的文脈関連度の計算を行った場合の検索結果を表している．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 (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．その他の 'LCDR' とあるのが潜在的文脈関連度を用いた場合の検索結果で，その後ろにある数字は特異値分解を行って近似行列を求めたときの次元数である．

グラフからも分かる通り，文脈関連度を用いた場合の検索精度と比較すると次元数が 100 までにおいて潜在的文脈関連度の方が良い精度を示している．しかし，次元数が 150 とした場合，文脈関連度の検

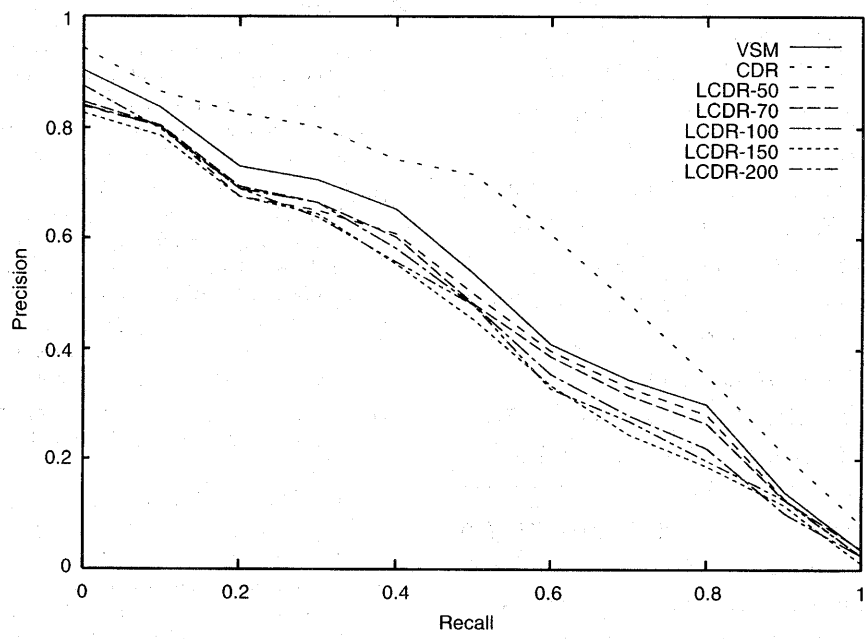

図 2: 正規化ありの場合における再現率–正解率曲線

索精度とほぼ同じようになり，次元数が 200 になると文脈関連度の方が検索精度が良くなっている．次元数を削減したことにより寄与率の少ない軸を無視することになるため，文書集合の中で関連のある索引語が効率良く抽出できているのではないかと考えられる．ただ，潜在的文脈関連度が検索質問拡張に有効であるという理論的な仮定に期待したにも関わらず，検索質問拡張を行わない元のベクトル空間モデルを用いた検索精度の方が高く，期待ほどの顕著な効果が表れない結果となった．

次に，式 (11) を用いて潜在的文脈関連度の計算を行ったときの検索結果を図 2 に示す．このグラフにおいても期待したほどの顕著な効果が出ず，それぞれの次元に対してほとんど同じ程度の検索精度となった．また，正規化を行った場合では文脈関連度の効果が顕著に表れ，元のベクトル空間モデルの検索精度を大幅に上回っている．特異値分解を行ってランクの低い近似行列を求めた場合，元の行列では 0 であった要素の値が，削減した特異値ベクトルがなす方向への成分だけ増減する．このため，内積計算や一文書に対する索引語の重みにかなりの影響が生じているのではないかと考えられる．

5 おわりに

本報告では，あらかじめ予備検索を行なって関連語を抽出することがなく，最初の検索で関連語が抽出できる潜在的文脈関連度を提案した．この潜在的文脈関連度の検索性能を調査するために，検索質問を拡張する情報検索システムを構築し，テストコレクションである MEDLINE を用いて検索実験を行った．その結果，単純なベクトル空間モデルを用いた情報検索モデルの検索性能と比較して検索精度が低下する結果となったが，正規化を行わない場合に関しては LSI を利用しない文脈関連度による検索モデルとの比較を行ったところ，潜在的文脈関連度を用いたモデルの検索精度が良くなっていることが分かった．また，正規化を行った場合については文脈関連度を用いた方が検索質問拡張の効果が顕著に表れ，元のベクトル空間モデルの検索精度を大幅に上

回っている．しかし，潜在的文脈関連度を用いたものは期待したほどの効果が表れず，それぞれの次元に対してほとんど同じ程度の検索精度となった．

本報告における検索実験では，関連度を求めたい索引語からなるベクトルと近似された文書ベクトルとの余弦が0.1以上の文書ベクトルのみを計算の対象としたり，拡張させる索引語の数を10と限定している．このため，さらに検索実験を行うことで，検索質問拡張を行うための尺度として潜在的文脈関連度が有効であるという可能性は，まだ残っているのではないかと考えられる．特異値分解によって次元を削減すると，削減前のスパースな行列から密な行列になり，その要素には大きく影響を受け数値が変化したものから少しだけ変化するものまでさまざま存在する．そのため，近似行列から隠れた意味情報を取り出すためには，余弦を0.1よりも小さく設定することで解決できるのではないかと考えられる．また，拡張させる索引語の数が10としたが，拡張させる索引語の数は5で検索精度のピークになるとの報告もある[8]．拡張させる索引語の数をもう少し少なくすることにより，どのように検索精度が変化するかも今後の課題のひとつである．

さらに，これまでLSIの他に提案されている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，これまでに我々が提案したConcept Projection[9]などを用いて，関連性の高いタームに高い重みを付ける改良を行いたい．さらに，Local Context Analysis[12]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．また，今回の実験では英語のテストコレクションのみで実験を行ったが，BMIR-J2[3]やIREX[10]などの日本語文書を対象とした大規模なテストコレクションを用いて，検索質問に対してどのような関連語が抽出されるのか調査したい．

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] 栗山 和子: "シソーラスを用いた検索式拡張の評価", 情報処理学会研究会報告 98-FI-52, pp. 1-8, 1998.

[5] D. D. Lewis. Evaluating text categorization. In *Proceedings of Speech and Natural Language Workshop*, pp. 312–318, 1991.

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

[7] 酒井 浩之, 大竹 清敬, 増山 繁: "絞り込み語提示による一検索支援手法の提案", 言語処理学会第7回年次大会, pp. 185–188, 2001.

[8] Mark Sanderson: "Word Sense Disambiguation and Information Retrieval", In *Proceedings of the 17th ACM SIGIR Conference*, pp. 142-151. 1994.

[9] 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.

[10] 関根 聡, 井佐原 均: "IREX プロジェクト概要", In *Proceedings of the IREX workshop*, pp. 1–5, 1999.

[11] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Van Nostrand Reinhold, New York, 1994.

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