Vol.2014-DBS-159 No.10
Vol.2014-IFAT-115 No.10
2014/8/1

単語の意味概念行列を用いたキーワード生成による関連論文検索システム

林 佑磨[1]　奥野 峻弥[1]　山名 早人[2,3]

概要：研究者は，研究意義や既存手法を知るために，自らの研究分野に関連する論文の調査を行う．論文の調査に広く用いられる論文検索システムは，ユーザがキーワードをクエリとして与えるキーワード検索が一般的である．専門用語の多い技術分野などでは，特に研究分野にまだ精通していない研究者が，適切なキーワードを与えて検索を行い，満足な結果を得ることは難しい．この問題を解決するため，我々は論文の概要を入力とする関連論文検索システムを提案した．同システムでは，入力の概要に含まれる単語が持つ意味を意味概念行列として表現し考慮することで，検索に用いるクエリの自動生成を行っている．本稿では，我々が以前提案したシステムの拡張を行う．具体的には，1) 日本語論文検索への対応，および 2)RSM による論文クラスタリングを用いてより質の高いキーワード生成を実現する．日本語に対応している既存の論文検索システムとの比較により，p@10 を平均で 0.17 向上させることに成功した．

1. はじめに

自らの研究分野に関連する論文の調査は，研究意義や既存手法を把握するために，研究者にとって最も重要な作業の一つである．近年，インターネットの普及に伴い，関連論文の調査はしばしば Google Scholar[1] のようなインターネット上の論文検索システムを用いて行われる．

論文検索システムを用いると，データベースに登録されている過去の膨大な論文に対する検索を行うことができる．多くの論文検索システムでは，ユーザ自らが検索に用いる複数のキーワードを考え，クエリとして発行するキーワード検索が主流である．しかし，まだ研究分野に精通していない研究者が関連論文検索を行う際や，専門用語の多い学術分野の論文検索を行う際には，適切なキーワードの想起が難しい場合が多い．そのため，目的とする分野の論文を発見するために多大な時間を浪費したり，論文の発見自体ができないこともある．

この問題点を解決するため，我々は以前，研究分野に精通していない研究者がキーワードを考える必要のない関連論文検索システムを提案した [2]．同システムは，「ユーザである研究者が，自らの研究に関連する幾らかの論文を保持している」という状況を前提としている．そして，システムへの入力を論文の概要本体とすることで，入力された概要，および同システム最大の特徴である意味概念行列を用いたクエリの生成による関連論文検索を実現した．

意味概念行列（SCM: Semantic Concept Matrix）とは，行と列が共に単語であるような対称行列である．SCM の各行（あるいは列）は，ある 1 単語を全単語の分布とするベクトルとして表現している．ベクトル内の要素は，単語と意味的な関連の強い語の重みは大きく，関連の弱い語の重みは小さくなっている．これは，国語辞典の見出し語を，その語釈文に含まれる語で説明することと等価である．

本論文では，以前のシステムに対して次の二点の拡張を行う．

本稿における主な改善点

1) 日本語論文検索への対応
2) RSM（Replicated Softmax Model）[3] を用いた論文のクラスタリング

具体的には，英語論文のみに対応していたものを日本語論文に対応させるために，論文データベース CiNii[7] が提供する OpenSearch API で収集した日本語論文の概要を用いて SCM の構築を行う．さらに，前システムではクエリ生成の際に不要な一般語が混ざるという問題があった．これは，SCM の構築に用いる論文群をカテゴリにクラスタリングしなかったため，SCM を作る元となる単語-文書行列を作る際に大域的な重み付けがうまく機能しなかったことが原因である．そこで本稿では，RSM を用いた論文の

[1] 早稲田大学基幹理工学研究科　〒 169–8555 東京都新宿大久保 3-4-1
[2] 早稲田大学理工学術院　〒 169–8555 東京都新宿大久保 3-4-1
[3] 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2-1-2

クラスタリングを行うことでこの問題を解決する．

本論文では以下の構成をとる．2 節では，関連研究の紹介を行う．3 節では，提案する関連論文検索システムの説明を行い，4 節では，既存研究との比較評価実験を行う．最後に 5 節でまとめる．

2. 関連研究

2.1 高野らによる連想技術を用いた検索エンジン

高野らは，大規模文書に対する類似度の高速計算が可能である連想計算エンジンを用いることで，「Webcat-plus」[4] や「想一 IMAGINE」[5] に代表される連想検索システムを作成している．

連想検索システムとは，文章あるいは文書そのものをクエリとして，文書間の類似度を計算することにより関連する文書を提示するような検索システムである．

連想検索は，入力に関連する文書を広く提示できるという点，そして利用者に気づきを与えるような結果が得られるという利点があるが，これらの利点は，研究分野のように狭い範囲で深く調べたい場合は逆に関係が薄い文書を抽出するという欠点になってしまう．

2.2 高久らのふわっと関連検索

高久らは，2010 年に「ふわっと関連検索」を提案した．「ふわっと関連検索」の入力は任意の文書である．システムは入力から特徴語を抽出することで，その文書に関連する論文の検索を行う [6]．高久らの手法（[6] より引用）を以下に示す．

入力：任意の文書または文書が掲載された web ページの URL
出力：入力文書に関連する論文集合
1) 本文抽出
与えられた文書，あるいは web ページのリンク先から本文を抽出する．
2) 特徴語抽出
本文を単語で分割し，各単語について出現頻度と単語の生起確率の積による重み付けを行い，特徴語群ベクトルを作る．
3) 検索クエリの発行
特徴語群ベクトルから，重みの大きい上位 10 単語を，その順に AND 条件で結合し，論文データベースに検索クエリとして発行する．検索結果の合計が 100 件を超えるまで，クエリから最も重みの低い特徴語を除外し，検索を繰り返す．
4) 検索結果の表示
検索で得られた結果から重複の除去を行い，最終的な出力として表示する．

上記の手法以外に，論文 [6] の中で異なるいくつかの手法に対する評価が行われているが，高久らが CiNii[7] と連携して提供している「ふわっと CiNii 関連検索」[8] という論文検索サービスでは，上記の手法が採用されている．

高久らの手法では，与えられた文書に含まれる単語のみから検索に用いるクエリの生成を行う．そのため，入力文書に含まれない専門用語をクエリに混ぜることはできない．また，頻度と生起確率の積による重み付けで重要語を決定するため，文書によってはクエリに含む単語として意味的にふさわしくない単語を混ぜてしまう可能性がある．

そこで，我々は意味概念行列 SCM を用いることで，文書には含まれないが意味的に分野と強く関連する単語も混ぜたクエリ生成を実現した．

3. 単語の意味概念行列を用いたキーワード生成による関連論文検索システム

本節では，我々が以前提案した関連論文検索システム [2] の改良点として，意味概念行列 SCM の構築方法と SCM を用いたクエリの生成方法，およびシステム構築に用いたデータセットとパラメータに関する説明を行う．

3.1 システムの概要

本システム内部の処理に関する概略図を図 1 に示す．

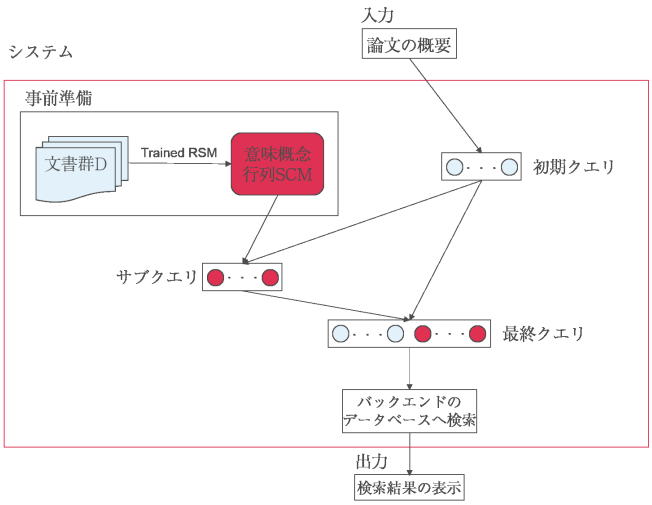

図 1 提案システムの概略図

本システムは，稼働する前の事前準備として，複数の論文の概要を用いた意味概念行列 SCM の構築を行う．この SCM の構築には，訓練済みの RSM[3] を利用する．ここで，RSM とはトピックモデルの一つである．RSM の学習，および SCM の構築に関する詳細は次節にて述べる．

事前準備を終え，システムを稼働した後は，ユーザにより入力された論文の概要に含まれる重要語を初期クエリとして取り出す．取り出された初期クエリに SCM を用いることで，意味的に強く関連する単語を取り出し，サブクエリとする．得られた初期クエリとサブクエリを組み合わせ

て最終的なクエリを作成し，バックエンドのデータベースに検索をかけることで関連論文の取得を行う．

3.2 意味概念行列の作成方法

単語と単語の間には明らかに意味的な相関が存在する [2]．この事実に基づき，全ての単語の持つ意味の概念を行列として定義したものが，意味概念行列（SCM: Semantic Concept Matrix）である．SCM は以下の 2 つの手順により作成される．

SCM 作成の手順

1) 複数文書から，単語-カテゴリ行列 $\mathcal{M}_{wc}$ の作成
2) $\mathcal{M}_{wc}$ の自己相関行列 $\mathcal{M}_{acr}(=\mathcal{M}_{wc}\mathcal{M}_{wc}^T)$ の作成，および $\mathcal{M}_{acr}$ の斜交化による，意味概念行列 $\mathcal{M}_{scm}$ の作成

以下では SCM の作成方法について，詳細を説明する．

3.2.1 単語-カテゴリ行列の生成

扱う対象である単語群を $\mathcal{W}=\{w_i|1<i<M\}$, カテゴリ群を $\mathcal{C}=\{c_j|1<j<N\}$，文書群を $\mathcal{D}=\{d_k|1<k<K\}$ とする．単語-カテゴリ行列 $\mathcal{M}_{wc}$ は，式 (1) に示す，行を単語 w_i，列をカテゴリ c_j とするような行列である．

$$\mathcal{M}_{wc}=\begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_M \end{array}\begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \cdots & c_N \end{array} \\ \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,N} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ m_{M,1} & m_{M,2} & \cdots & m_{M,N} \end{pmatrix} \end{array} \quad (1)$$

ここで，$\mathcal{M}_{wc}$ の要素 $m_{i,j}$ は，カテゴリ c_j に属する文書 d_k に含まれる，単語 w_i の頻度を全て加算することにより得られる．ただし，任意の文書 $d_j\in\mathcal{D}$ は必ず $\mathcal{C}$ に属する 1 つ以上のカテゴリに属するものとする．

文書 d_k が属するカテゴリを特定するために，RSM（Replicated Softmax Model）[3] という RBM（Restricted Boltzmann Machine）[9] に基づくトピックモデルを利用した．RSM は複数のトピックをその掛け合わせで表現することを可能とする，積モデルと呼ばれるトピックモデルの代表である．トピックモデルを用いることで，各文書を潜在的なトピック（今はカテゴリ）にクラスタリングすることが可能となる．

RSM は図 2 に示す通り，潜在ユニットおよび観測ユニットからなる．ここで，観測ユニットの各ノードは，単語群 $\mathcal{W}$ に含まれる全単語であり，潜在ユニットの各ノードは何らかの特徴を抽出したものである．今回は潜在ユニットの意味付けとして，各ノードが $\mathcal{C}$ に属するカテゴリを表現していると解釈した．

RSM はトピックが生成する確率が「和」ではなく「積」の形で表現されるため，一般に LDA[10] に代表される混合モデルと比べて良い精度でのクラスタリングが可能である [3]．

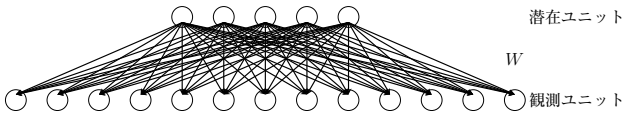

図 2 RSM の概念図

RSM の訓練は，複数の文書を与えることで，図 2 における観測ユニットと潜在ユニットとのつながりの重み W を学習する．学習を終えた後は，RSM に新規文書が与えられると，1 つ以上の潜在ノードが，ある確率値で発火する．発火したノードに対応するカテゴリに，発火確率と新規文書に含まれる単語の頻度を乗じたものを加算していくことで，単語-カテゴリ行列 $\mathcal{M}_{wc}$ を作成する．

3.2.2 自己相関行列の作成と斜交化

単語-カテゴリ行列 $\mathcal{M}_{wc}$ の要素は，RSM によるカテゴリの発火確率を乗じているが，基本的には各単語の出現頻度と同等のものである．ここで，複数のカテゴリに共通して頻出する語の重みを下げるために，RIDF（Residual Inversed Document Frequency）[11] による大域的な重み付けを行う．

RIDF は，ポアソン分布を導入した大域的重み付け手法である．ポアソン分布は，カテゴリに寄らず多くの文書に現れる一般語に対してはその文書頻度の良い近似となり，文書を特徴付ける内容語には当てはまらないという特性をもつ [11]．したがって，RIDF を用いることで，特徴的な内容語の重みを下げずに，一般語の重みだけを下げることが可能となる．単語 w_i に対する RIDF は式 (2) により表される．

$$\begin{aligned} RIDF_{w_i} &= IDF-\widehat{IDF} \\ &= \log\frac{|\mathcal{D}|}{|\{d|w_i\in d\wedge d\in\mathcal{D}\}|}+\log\left(1-\exp^{-\frac{F_{w_i}}{|\mathcal{D}|}}\right) \end{aligned} \quad (2)$$

ここで，F_{w_i} は単語 w_i の大域的な頻度，すなわち全文書を通しての出現頻度である．単語-カテゴリ行列 $\mathcal{M}_{wc}$ において，各単語 $w_i\in\mathcal{W}$ に対応する行にその単語の RIDF 値を乗じる．こうして得られる新たな単語-カテゴリ行列を $\mathcal{M}'_{wc}$ とする．

つぎに，$\mathcal{M}'_{wc}$ の自己相関行列 $\mathcal{M}_{acr}$ を式 (3) により作成する．

$$\mathcal{M}_{acr}=\mathcal{M}'_{wc}\mathcal{M}'^T_{wc}=\begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_M \end{array}\begin{array}{c} \begin{array}{cccc} w_1 & w_2 & \cdots & w_M \end{array} \\ \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M,1} & a_{M,2} & \cdots & a_{M,M} \end{pmatrix} \end{array} \quad (3)$$

このようにして作られる自己相関行列 $\mathcal{M}_{acr}$ は行と列が共に単語であるような単語-単語行列である．$\mathcal{M}_{acr}$ の要素 $a_{i,j}$ は，行の単語 w_i と列の単語 w_j があるカテゴリにおいてよく共起する場合に大きな値をとり，どのカテゴリにおいてもあまり共起しない場合には小さな値をとる．

$\mathcal{M}_{acr}$ は直交座標に基づく行列である．すなわち，各単語は互いに直交し，独立であるとみなすことができる．しかしながら，本来，単語と単語には意味的な相関が存在する [2]．そこで，$\mathcal{M}_{acr}$ の各成分にその成分を含む行と列のコサインを乗じる斜交化を行い，斜交座標へ変換する．この斜交化を行うことで，単語間の意味的な相関を表現する．

斜交化を行なって得られる最終的な意味概念行列 $\mathcal{M}_{scm}$ は式 (4) となる．

$$\mathcal{M}_{scm} = \begin{array}{c} \\ w_1 \\ w_2 \\ \vdots \\ w_M \end{array} \begin{array}{c} \begin{array}{ccc} w_1 & \cdots & w_M \end{array} \\ \begin{pmatrix} a_{1,1}\cdot\cos\theta_{1,1} & \dots & a_{1,M}\cdot\cos\theta_{1,M} \\ a_{2,1}\cdot\cos\theta_{2,1} & \dots & a_{2,M}\cdot\cos\theta_{2,M} \\ \vdots & \ddots & \vdots \\ a_{M,1}\cdot\cos\theta_{M,1} & \dots & a_{M,M}\cdot\cos\theta_{M,M} \end{pmatrix} \end{array}$$
$$, cos\theta_{i,j} = \frac{\mathcal{M}_{wc}[i,*]\cdot\mathcal{M}_{wc}[*,j]}{\|\mathcal{M}_{wc}[i,*]\|\|\mathcal{M}_{wc}[*,j]\|} \qquad (4)$$

3.3 意味概念行列を用いたクエリ生成

本システムは，入力として与えられた論文の概要，および意味概念行列 SCM を用いて，検索に用いるクエリの自動生成を行う．システムが生成するクエリは，1) 入力される概要に含まれる重要単語として抽出される「初期クエリ」と，2)SCM を用いて抽出される初期クエリの近傍単語である「サブクエリ」の 2 つから構成される．以下では初期クエリ，およびサブクエリの作成方法を説明する．

3.3.1 初期クエリの生成

初期クエリ $\mathcal{IQ}$ は，入力された概要本体に含まれる n 個の重要単語を抽出することで作られる単語の集合である．具体的な手順を以下に示す．

1) 入力された概要に対して，MeCab[12] を用いた形態素解析を行い，単語群 $\mathcal{W}$ に属する単語の集合として表現する．すなわち，Bag-of-words 表現にする．

2) Bag-of-words 形式化された入力に含まれる各単語に対して，出現回数と RIDF による重み付け（tf-$ridf$）を行う．

3) tf-$ridf$ の値が大きい上位 n 単語を抽出し，初期クエリ $\mathcal{IQ}$ とする．

3.3.2 サブクエリの生成

サブクエリ $\mathcal{SQ}$ は，初期クエリ $\mathcal{IQ}$ に含まれる単語と意味的に強く関連する単語である近傍単語から構成される単語集合である．$\mathcal{SQ}$ の作成は，意味概念行列 SCM を用いて行う．具体的な手順を以下に示す．

1) 候補となる近傍単語の抽出
   i) 初期クエリ $\mathcal{IQ}$ に含まれる単語 $w_i \in \mathcal{IQ}$ に対応する SCM の行ベクトル $\mathcal{M}_{scm}[i,*]$ を取り出す．
   ii) $\mathcal{M}_{scm}[i,*]$ と，単語群 $\mathcal{W}$ に含まれる w_i 以外の全単語の，SCM における行ベクトル $\forall j \neq i, \mathcal{M}_{scm}[j,*]$ とのコサイン類似度を計算する．
   ここで，コサイン類似度は式 (5) により表される．
   $$sim_{cos}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{\|\boldsymbol{v}_1\|\|\boldsymbol{v}_2\|} \qquad (5)$$
   iii) コサイン類似度が大きいものから上位 m 単語を近傍単語として抽出する．
   iv) i) から iii) までの操作を，$\mathcal{IQ}$ に含まれる全単語に対して行い，候補となる全ての近傍単語 $\mathcal{W}_{nh}$ を取得する．

2) $\mathcal{W}_{nh}$ に属する単語を，以下の順で並び替える．
   i) 出現回数
   ii) 出現順位
   ここで，出現順位とは，全近傍単語の中で，対象の近傍単語が何番目に抽出されたかを意味する．全ての近傍単語は $\mathcal{IQ}$ に含まれる単語 w_i の近傍として取り出されるので，w_i の tf-$ridf$ が高いものほど早くとり出されることになる．

3) 並べられた，候補となる近傍単語の，上位 m 単語を抽出し，サブクエリ $\mathcal{SQ}$ とする．

3.4 関連論文検索

検索に用いるクエリの作成と発行の手順は，高久らの「ふわっと関連検索」[6] で用いられている手法に従った．具体的には，システムが生成した初期クエリ $\mathcal{IQ}$，およびサブクエリ $\mathcal{SQ}$ に含まれる語を順に AND 条件で結合することによりクエリを作成し，検索を行う．なお，クエリに含まれる単語数を高久らの手法で生成されるものと等しくするため，3.3.1 における初期クエリ数 n および 3.3.2 におけるサブクエリ数 m はそれぞれ 5 と設定した．実際の検索は以下の手順で繰り返し行われる．

1) バックエンドのデータベースへクエリの発行

2) 最も重要度の低い単語を除き，新たなクエリとする．

3) 検索により取得した論文の合計数が 100 を超えていなければ，1) と 2) を繰り返す．

このようにして得られた検索結果を，出現回数の順で並び替え，本システムの最終出力とする．

Vol.2014-DBS-159 No.10
Vol.2014-IFAT-115 No.10
2014/8/1

3.5 使用したデータとパラメータおよびバックエンドの論文データベースについて

本システムでは，用いた文書集合 $\mathcal{D}$ として，CiNii に登録されている，出版年が 2010 年から 2013 年の間である論文の抄録を用いた．ただし，抄録が記載されていないものに関しては取り除いたため，最終的には合計 48,110 本の論文抄録を用いた．

これらの 48,110 文書を形態素解析ツールである MeCab[12] にかけることで，単語集合 $\mathcal{W}$ を得た．ただし，抽出する単語は名詞あるいは形容詞に限定した．また，分野によらずあまりに一般的な単語と稀な単語を除くため，経験および頻度分布に基づいて，総出現頻度が 8 回以上 3,000 回以下の合計 16,013 単語を扱う対象とした．

論文のクラスタリングに用いた RSM（Replicated Softmax Model）[3] の学習には，新規文書に対する汎化性能を持たせるため，全文書数 $|\mathcal{D}|$ の約半分である 20,000 個の文書を $\mathcal{D}$ からランダムに抽出して用いた．この際，RSM のパラメータとして，観測ユニットのノード数は総単語数 $|\mathcal{W}| = 16,013$ である．また，潜在ユニットのノード数は，perplexity が低かった $|\mathcal{C}| = 32$ と設定した．また，単語-カテゴリ行列の生成に関しては，学習後の RSM に $\mathcal{D}$ に属する全ての文書を入力することにより行った．

3.6 システムのインタフェース

本システムのインタフェースを図 3 に示す．

図 3 システムのインタフェース

システムへの入力は，任意の論文の概要となっている．図 3 の入力欄に論文の概要を貼り，検索ボタンを押すと，入力と関連する論文の 1) タイトル，2) 著者，および 3) 発行年が提示される．

4. 評価実験

4.1 評価実験手法

本評価実験では，提案システムと既存システムである高久らによる「ふわっと CiNii 関連検索」[8] の両方に複数の論文の概要を入力し，両システムが検索結果として提示する上位 10 件の論文に関して，3 人の評価者が評価を行った．実験に関する詳細な条件を以下に示す．

評価実験の条件

- 評価者：情報理工分野を専門とする大学院生 3 人．
- システムへの入力：CiNii に登録されている，出版年が 2009 年以前である論文で，抄録が付いている無作為に取り出された情報理工分野の論文 10 本．なお，入力として用いる抄録は，本システムを構築する際に用いたデータセットには含まれていない．
- 評価内容：評価者は，論文執筆者になったという設定で，両システムの出力結果上位 10 件が，入力に用いた論文の参考文献としてふさわしいものかどうかの適合度を A（適合），B（部分適合），C（不適合）の 3 段階で評価する．なお，評価者は，出力された論文のタイトルを見て適合度を判定し，タイトルのみからの判定が不可能な場合は概要も参考にして判定を行う．
- 適合度の基準：適合度の基準は表 1 に示す．

表 1 適合度の基準

適合度	基準
A（適合）	重要なトピックについて言及しており，論文の参考文献として記載するのにふさわしい論文である
B（部分適合）	論文の参考文献として記載するのにふさわしくはないが，論文を執筆する際にある程度は参考になる
C（不適合）	内容が完全に異なり，まったく参考にならない

上記の条件下で 3 人の評価者が付けた適合度に対し，1)P@10，2)MAP@10 の評価指標を用いて最終評価を行う．

ここで，1) と 2) どちらの評価指標においても，適合度における A のみを適合とみなした評価 $\mathrm{P@10}_A$ および $\mathrm{MAP@10}_A$ と，B 以上を適合とみなした評価 $\mathrm{P@10}_{AB}$ および $\mathrm{MAP@10}_{AB}$ を算出した．

4.2 実験結果

3 名の評価者による評価結果をそれぞれ表 2，表 3，表 4 に示す．また，3 名の評価者が付けた評価結果の平均をとったものを表 5 に示す．

表 2　評価者 1 が行った評価実験の結果

手法	$P@10_A$	$P@10_{AB}$	$MAP@10_A$	$MAP@10_{AB}$
提案	0.23	0.62	0.51	0.83
既存	0.27	0.49	0.67	0.65

表 3　評価者 2 が行った評価実験の結果

手法	$P@10_A$	$P@10_{AB}$	$MAP@10_A$	$MAP@10_{AB}$
提案	0.33	0.75	0.57	0.89
既存	0.24	0.50	0.46	0.70

表 4　評価者 3 が行った評価実験の結果

手法	$P@10_A$	$P@10_{AB}$	$MAP@10_A$	$MAP@10_{AB}$
提案	0.25	0.54	0.58	0.81
既存	0.23	0.41	0.51	0.69

表 5　3 名の評価者が行った評価実験の平均

手法	$P@10_A$	$P@10_{AB}$	$MAP@10_A$	$MAP@10_{AB}$
提案	0.27	0.64	0.55	0.84
既存	0.25	0.47	0.55	0.68

4.3　実験結果の考察

4.2 の結果から，表 1 に示した適合度の A と B を適合とした場合，本システムは既存システムと比べて p@10 が平均で 0.17 高いという結果を得た．このことから，出力結果上位 10 件に関連論文を平均で 1 本以上は多く提示できており，既存手法と比べて有用であることが示された．

本システムは既存システムと比べ，記述量が少ないような概要に対して特に良い結果の出力を行うことができた．これは，意味概念行列 SCM を用いて本文に含まれない関連語である近傍単語を追加しているためである．なお，今回は入力として情報理工分野の論文概要を用いたが，この分野内における得意・不得意は見受けられなかった．

ここで，実際にある単語に対して，意味概念行列 SCM を用いて抽出した上位 5 個の近傍単語の例を以下に示す．

SCM を用いて抽出した近傍単語の例

- 「意味」の近傍単語：
  概念，解釈，考究，想像，言及
- 「遺伝子」の近傍単語：
  突然変異，マーカー，変異，ポリメラーゼ，発現
- 「ベイズ」の近傍単語：
  マルコフ，系列，識別，サポートベクトルマシン，確率

このように，SCM を用いることで，ある単語と意味的に強く関連する語である近傍単語の抽出が可能であることがわかる．この近傍単語をクエリに混ぜることで，入力に関連する分野の検索を幅広く行うことが可能となる．

一方で，入力の概要に含まれる特徴的な語が少ない場合，本システムはうまく関連論文の提示を行えなかった．これは，一般的な単語の近傍単語を加えたことで検索の方向性が定まらなくなったためであると考えられる．

5.　おわりに

本稿では，我々が以前提案した関連論文検索システム [2] の改良を行った．具体的には，1) 日本語対応を行い，2) 意味概念行列 SCM の質を高めた．1) は，CiNii[7] から取得した論文の概要を用いた SCM の構築を行い実現した．2) は，SCM の構築に用いる論文の概要を RSM によりカテゴリにクラスタリングし，各単語に RIDF による重み付けを行うことで一般語の重みを下げることにより実現した．

SCM の改善により，より質の高い近傍単語の抽出が可能となった．これにより，最終的なクエリに入力された概要に含まれていない専門用語も混ぜることができ，分野に関連する論文を広く検索することを可能とした．

本手法は幾らかのヒューリスティクスを用いている．したがって，今後は更なる実験を行い，各種最適なパラメータを見つけることが必要である．また，クエリの生成は単語を単純に AND でつなげているため，追加された近傍単語が検索に十分に活かされていない．そこで，今後は複数単語からのより良いクエリ作成について考える必要がある．

参考文献

[1] Google: “Google Scholar”, http://scholar.google.co.jp/, 2014 年 6 月 23 日アクセス．
[2] 林佑磨，奥野峻弥，山名早人: “意味概念に基づいた関連論文検索システム ～ 近傍文書からのキーフレーズ抽出を用いた自動クエリ生成 ～”，第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM2014），2014.
[3] Ruslan Salakutdinov and Geoffrey E. Hinton: “Replicated Softmax: an Undirected Topic Model”, Advances in Neural Information Processing Systems, Vol. 22, pp.1607-1614, 2009.
[4] 国立情報学研究所: “Webcat-plus”, http://webcatplus.nii.ac.jp/，2014 年 6 月 23 日アクセス．
[5] 連想出版: “想-IMAGINE”, http://imagine.bookmap.info, 2014 年 6 月 23 日アクセス．
[6] 高久雅生, 江草由佳: “簡易類似文書検索手法 「ふわっと関連検索」 の予備的評価と分析”, 情報処理学会研究報告, データベース・システム研究会報告, Vol. 2010, No. 14, pp.1-6, 2010.
[7] 国立情報学研究所: “CiNii”, http://ci.nii.ac.jp/, 2014 年 6 月 23 日アクセス．
[8] 高久雅生: “ふわっと CiNiin 関連検索”, http://fuwat.to/cinii, 2014 年 6 月 28 日アクセス．
[9] Ruslan Salakhutdinov and Geoffrey Hinton: “A practical guide to training restricted Boltzmann machines”, Technical report 2010-003, Machine Learning Group, University of Toronto, 2010.
[10] David M. Blei, Andrew Y. Ng and Michael I. Jordan: “Latent dirichlet allocation”, the Journal of machine Learning research, Vol. 3, pp. 993-1022, 2003.
[11] 北研二，津田和彦，獅々堀正幹: “情報検索アルゴリズム”, 共立出版株式会社，初版，2005.
[12] 工藤拓: “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.sourceforge.net/, 2014 年 6 月 23 日アクセス．