DEWS2004 2-B-02

数式データを対象とした複合連想検索の実現

岸本　貞弥† 　中西　崇文†† 　櫻井　鉄也††† 　北川　高嗣†††

† 筑波大学大学院 理工学研究科　〒 305–0006 茨城県つくば市天王台 1–1–1 数値解析研究室
†† 筑波大学大学院 システム情報工学研究科　〒 305–0006 茨城県つくば市天王台 1–1–1 数値解析研究室
††† 筑波大学 電子・情報工学系　〒 305–0006 茨城県つくば市天王台 1–1–1
E-mail: †kishimoto@nalab.is.tsukuba.ac.jp, ††takafumi@nalab.is.tsukuba.ac.jp,
†††{sakurai,takashi}@is.tsukuba.ac.jp

あらまし　現在，Mathematical Markup Language(MathML) の仕様が公表され，web 上の数式を含む文書における数式が利用できる状況にある．これまで我々は Latent Semantic Indexing (LSI) を用いて MathML で記述された数式を問い合わせとして類似数式検索機能を実現してきた．今回この類似数式検索機能と，数学用語等の言葉を適用した意味の数学モデルによる単語間の関連連想検索機能を連結する方式を提案する．これによって複数のメディアからなる数式データに対する検索が行える．本稿では数式と言葉からなる数式データを対象とした複合連想検索について示し，線形代数の数式データを用いた実験結果も示す.

キーワード　MathML, 類似数式検索, 意味の数学モデル, 情報検索, 異種 DB

An Implementation Method of Composite Association Retrieval for Data of Mathematical Formulas with Words

Sadaya KISHIMOTO†, Takafumi NAKANISHI††, Tetsuya SAKURAI†††, and Takashi KITAGAWA†††

† Master's Program in Science and Engineering, University of Tsukuba
†† Graduate School of Systems and Information Engineering, University of Tsukuba
††† Institute of Information Sciences and Electronics, University of Tsukuba
E-mail: †kishimoto@nalab.is.tsukuba.ac.jp, ††takafumi@nalab.is.tsukuba.ac.jp,
†††{sakurai,takashi}@is.tsukuba.ac.jp

Abstract　Now Mathematical Markup Language(MathML) was released, and the use of mathematical content on the Web is technically possible. We have implemented a function of similarity-based retrieval for formulas with Latent Semantic Indexing (LSI), using formulas encoded by MathML as queries. This time, we are going to suggest linking the function of similarity-based retrieval for formulas to a function of words-related associative search applied to mathematical terms. In this paper, we describe an implementation method of composite association retrieval for data of mathematical formulas with words, and results of experiments in which the data of formulas with words from linear algebra were used.

Key words　MathML, Similarity-based formulas retrieval, Mathematical Model of Meaning, Information retrieval, Heterogeneous database

1. はじめに

いま web 上には文書や画像をはじめとしたメディアデータが大量に散在している．そして，それらは複合メディアであることが多い．例えば，文書においては言葉と画像からなるものが多く存在する．同様に，多くの数式を含んだ技術ドキュメントが web 上に散在している．これらもまた複合メディアであることが多い．例えば，数学の論文においては言葉と数式からなるものが多く存在する．このような状況の中で，検索者の要求により近いメディアデータを検索する方式を確立することは重要な課題となっている．

現在，Mathematical Markup Language(MathML) の仕様が

公表され，web 上の数式を含む文書における数式が利用できる状況にある．これまで我々は Latent Semantic Indexing(LSI) [1] を用いて Mathematical Markup Language(MathML) [2] で記述された数式を問い合わせとして類似数式検索機能 [3] を実現してきた．しかしながら，数式から数式を検索するのみでは応用範囲が極めて狭い．

一方，我々は検索者の与える文脈に応じた間接検索方式として，意味の数学モデルを提案している．これによって数学用語等の言葉を適用した単語間の関連連想検索機能が実現できる．

これら手法を統合して適応することにより，複合メディアである数式を含む文書に対する検索も可能であると考えた．そこで，数式を含む文書に対する検索の前段階として，言葉と数式からなる問い合わせによる言葉と数式からなる数式データを対象とした複合連想検索を提案する．

独自のインデックス付けを行った数学データベースに対してパターンマッチングによる検索を行う研究 [4] はすでに行われている．これに対して，本研究は World Wide Web Consortium(W3C) [5] による仕様である MathML を用いて，Web 上の数式データを対象とした類似検索を行うものであり，上記の研究とは対象とするデータ形式及び，検索方法が異なる．

本稿では，まず類似数式検索機能について述べ，次に単語間関連連想検索機能 [6] について述べる．さらにそれらの統合方式を示し，数式データを対象とした複合連想検索について示す．

2. Mathematical Markup Language (MathML)

MathML は数式の構造と内容の両方を書き表すことを可能とする XML ベースのマークアップ言語である．MathML ファイルは単独で使用されるほか，他の XML 文書に埋め込んで使用することが可能である．MathML は特に XHTML で記述された Web ページに数式を埋めこむ際に使われることを強く意識されている．

MathML は Maple や Mathematica などの数式処理アプリケーションで扱うことができるほか，Web ブラウザでも対応が進んでいる．Mozilla や Netscape7 では既に対応しており，Internet Explorer ではプラグインの MathPlayer を使用することで MathML に対応できる．また，TeX で書かれたドキュメントを MathML を含む文書に変換するソフトウエアもある [7]．

MathML では数式の表記を表す Presentation Markup と数式の意味を扱う Content Markup の 2 種類のタグが用意されている．以下でそれぞれについて説明する．本稿では，数式の内容を記述する Content Markup を検索に利用する．

- Presentation Markup

Presentation Markup は数式の持つ意味は表現しておらず，少数のタグセットから成る．これは主に web ブラウザなどでの数式表示を行うために用いられる．Presentation Markup で表現された数式例を図 1 に示す．

```
<math>
 <mrow>
  <msup>
   <mfenced>
    <mrow>
     <mi>a< /mi>
     <mo>+< /mo>
     <mi>b< /mi>
    < /mrow>
   < /mfenced>
   <mn>2< /mn>
  < /msup>
 < /mrow>
< /math>
```

図 1　$(a+b)^2$ の Presentation Markup による表記.
Fig. 1　Notation by Presentation Markup of $(a+b)^2$.

- Content Markup

Content Markup は省略された積など数式の表示には明確に現れてこない構造も含め，数式の内容を正確に記述するためのものであり，約 150 個のタグが用意されている．Content Markup で表現された数式例を図 2 に示す．

```
<math>
 <apply>
  <power/ >
  <apply>
   <plus/ >
   <ci>a< /ci>
   <ci>b< /ci>
  < /apply>
  <cn>2< /cn>
 < /apply>
< /math>
```

図 2　$(a+b)^2$ の Content Markup による表記.
Fig. 2　Notation by Content Markup of $(a+b)^2$.

3. 類似数式検索の実現方式

ここでは，類似数式検索の実現方式について概要を述べる．本方式は MathML で書かれた数式を対象として，与えられた数式とタグの構成が類似した数式を検索するシステムである．本方式の特徴は，数式の演算子に注目して検索を行うことにより，添え字や変数に使う文字の違いなどによる，記述方法が異なる数式においても同様の意味と捉えて検索可能な点にある．

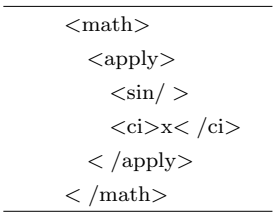
3.1 類似数式検索方式の概要

（1） 検索対象の数式群よりデータ行列を自動作成

まず，検索対象の MathML で記述された数式から，その数式の特徴を表すメタデータを抽出する．次にそれらを並べて構成するデータ行列を生成する．この行列により，検索対象となる数式データ群の類似度を計量する空間に表現することができる．メタデータ自動抽出方式については 3.2 節で示す．

（2） 問い合わせの数式よりメタデータを抽出

検索対象の数式データと同様に，問い合わせとして与えられた MathML で記述された数式から，その数式の特徴を表すメタデータを抽出する．



$\sin x$ の MathML 表現

	···	apply	sin	cos	cn	ci	plus	···
$\sin x$	···	1	1	0	0	1	0	···

MathML 中のタグの種類とその頻度をカウント

図 3　$\sin x$ の例.

Fig. 3　Example of $\sin x$.

（3）　**類似度を計量**

上記項目 (1),(2) により抽出されたメタデータから，類似度を計量し，その値の大きい順にソートする．これにより，問い合わせの数式とタグの構成が類似した数式が検索される．本方式では，類似度の尺度としてコサイン尺度を用いている．

3.2　MathML で表現された数式を対象としたメタデータ自動抽出方式

本節では，MathML で記述された数式からメタデータを抽出する方式について述べる．本方式は，MathML のタグ情報に注目し，数式の特徴として抽出することにより，数式の演算子に依存した検索を実現するものである．具体的には以下の手順で実現される．

（1）　**MathML 表現の数式が構成するタグの種類とその出現頻度を導出**

対象となる MathML 表現の数式データ $\boldsymbol{d}_i(i = 1, 2, \cdots, n)$ のタグの種類とその出現数をカウントすることで特徴づけする．

$$\boldsymbol{d}_i = (t_{1i}, t_{2i}, \cdots, t_{mi})^{\mathrm{T}}. \tag{1}$$

$t_{1i}, t_{2i}, \cdots, t_{mi}$ は対応する MathML のタグの出現頻度を表す．例として図 3 のように行う．

（2）　**tf・idf による重み付け**

抽出したタグの頻度によってその数式の特徴を表しているが，タグの中には，どの数式にも多く含まれるタグが存在し，各数式の特徴を表す際にノイズとなる可能性がある．本方式では，全文検索においてよく用いられている tf・idf [8], [9] を用いて重み付けを行う．

4.　単語間関連連想検索の実現方式

ここでは，数学用語等の言葉を適用した単語間関連連想検索の実現方式について概要を述べる．特定分野を対象とした連想検索のためのメタデータ空間生成し，意味の数学モデルに適用することでこれを実現している．この検索機能によって，問い合わせの語に関連する語を検索することができる．

4.1　意味の数学モデルの概要

4.1.1　意味の数学モデルの基本構成

本節では，人間が様々な印象を表す際に用いられる単語 (以下，印象語) によって表現した問い合わせに対応した情報群を検索することを目的とした意味の数学モデルの概要を示す．詳細は文献 [10]〜[12] に述べられている．

（1）　**メタデータ空間 $\mathcal{MDS}$ の設定**

検索対象となる情報群をベクトルで表現したデータにマッピングするための正規直交空間 (以下，メタデータ空間 $\mathcal{MDS}$) を設定する．メタデータ空間生成方式については，4.2 節で示す．

（2）　**検索対象データのメタデータをメタデータ空間 $\mathcal{MDS}$ へ写像**

設定されたメタデータ空間 $\mathcal{MDS}$ へ，検索対象データのメタデータをベクトル化し写像する．これにより，検索対象データのメタデータが同じメタデータ空間上に配置されることになり，検索対象データ間の意味的な関係を空間上での距離として計算することが可能となる．

（3）　**メタデータ空間 $\mathcal{MDS}$ の部分空間の選択**

利用者は与える文脈を複数の印象語を用いて表現する．ユーザが与える印象語の集合をコンテクストと呼ぶ．このコンテクストを用いてメタデータ空間 $\mathcal{MDS}$ に各コンテクストに対応するベクトルを写像する．これらのベクトルは，メタデータ空間 $\mathcal{MDS}$ において合成され, 意味重心を表すベクトルが生成される．意味重心から各軸への射影値を相関とし, 閾値を超えた相関値 (以下，重み) を持つ軸からなる部分空間が選択される．

（4）　**メタデータ空間 $\mathcal{MDS}$ の部分空間における相関の定量化**

選択されたメタデータ空間 $\mathcal{MDS}$ の部分空間において，検索対象データベクトルと検索語列との相関を計量する．メタデータ空間に写像された検索対象データベクトルの部分空間におけるノルムを求めることにより，文脈に対応した検索対象データの探索を行う．部分空間における検索対象データベクトルのノルムの大きさをその文脈と検索対象データとの関連の強さとする．これにより，与えられたコンテクストと各メディアデータとの相関の強さを定量化している．この部分空間における検索結果は，各メディアデータを相関の強さについてソートしたリストとして与えられる．

4.2　メタデータ空間生成方式

本節では，特定分野を対象としたメタデータ空間を，語とページの関係が記述されている書籍の索引を用いて生成する方式を示す．本方式はある言語が表現可能な空間全体を作成するのではなく，検索対象となるデータが包含されている特定分野に関する空間を生成することを目的としている．このような空間の作成を前提とすることにより，ある言語の空間を作成するよりも比較的少ない労力で，特定分野に関連するドキュメントの語と語の関係をより適切に表現できると考えられる．本方式では，検索対象が包含する特定分野について書かれた書籍が存在することを前提としている．

本方式は以下の流れで実現する．

（1）　初期行列の設定

まず，対象とする特定分野について書かれた書籍の索引を参照する．

索引とは以下の性質を持つものとする．

- 索引はキーワードとなる語とその語が関係するページと

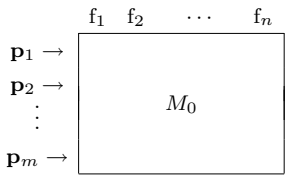


図 4　初期データ行列 M_0 によるメタデータの表現.
Fig. 4　Metadata represented in first data matrix M.

の組である.

- キーワードは，異なるページに何度出てきても良い.
- ページは基本的に複数個のキーワードを含むが，必ずしもキーワードを含んでいる必要はない.

索引に出現するキーワードとなる語を特徴語とみなし，索引情報から各ページ数を用いて特徴付ける.

$$\mathbf{p}_i = (f_{i1}, f_{i2}, \cdots, f_{in}) \tag{2}$$

ここで i はページ数，f_{ik} は特徴語に対応したページ数について特徴付けた値である．特徴付ける f_{ik} の値は，以下のように決定される.

- 索引中で特徴語がそのページ数を参照している場合 : "1"
- 索引中で特徴語がそのページ数を参照してない場合 : "0"

文献 [12]〜[14] のような用語辞典や辞書からデータ行列を生成する方式では，特徴付けのとりうる値として，"1"，"0"，"−1" の 3 値となっている．これは，用語辞典や辞書の内容から説明で「…である」などの肯定的な用法で用いられている場合は"1",「…ではない」,「…を伴わない」などの否定的な用法で用いられている場合は"−1"と意味を読取ってデータ行列に反映させている.

本方式は索引を用いるため，索引にはキーワードとしてあらわされている特徴語とそのページの関係しか記述されておらず，そこから，肯定の意味か否定の意味かを読取るには，本文をいちいち参照しない限り，不可能である．しかしながら，語が肯定の意味に使われているか否定の意味に使われているかに関わらず，その語がそのページに出現するということは，そのページで示されている事象を説明するために使われていることから，なんらかの関係を持っているということが考えられる．このことから，本方式では，"1"，"0"の 2 値を用いる.

以上から，$\mathbf{p}_i$ を用いて，$(\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m)^T$ とすることによって，図 4 のような m 行 n 列の初期データ行列 M_0 を作成する.

（2） 初期データ行列の修正によるデータ行列の生成

(1) で作成した初期データ行列 M_0 には，ページと語の関係を表す行列となっており，ページ同士の関係が反映されていない．そのため，ある概念を複数ページにわたって書かれている場合，索引に記述されているキーワードとして表される語とページの関係だけでは表現しきれず，精度を悪化させる原因となりうる．初期データ行列 M_0 にページ同士の関係を反映するように修正してデータ行列 M を生成する.

一般的に，書籍には目次が付いており，目次には章，節とその題名，そしてページ数が付与されている．章，節は，ある概念を説明するための論理的な枠であることから意味的な関係のあるページのかたまりとして捉えることができる．また，章，節に付与された題名は，説明された概念を端的に表すのに適切な語，フレーズである．これらの情報を反映することにより，ページ同士の関係を反映したデータ行列 M が生成可能となる.

まず，章，節の題名を語に分解し，接続詞など直接特徴を表さない語を排除する．その章，節に属するページ全てについて，題名を分解してできた語を特徴語として初期データ行列 M_0 を修正，追加する.

もし，題名を分解してできた特徴語が索引に使われていて，特徴として示されていた場合は，該当ページの特徴を全て"1"に修正する．もし，題名を分解してできた特徴語が索引に使われておらず，特徴として示されていない場合は，その特徴語を特徴として追加し，該当ページを全て"1"，それ以外のページを"0"と特徴付ける.

以上により，m 行 $n+\alpha$ 列のデータ行列 M を生成できる．ここで，α は特徴を追加した場合の要素の増加分を表す.

（3） 相関行列 M^TM からメタデータ空間生成

(2) で生成されたデータ行列 M の相関行列 M^TM を計算すると，$n+\alpha$ 行 $n+\alpha$ 列の行列となる．これは特徴語と特徴語の関係を示す行列となる．よって，この相関行列 M^TM を固有値分解し，非ゼロ固有値に対応する固有ベクトルによってメタデータ空間を生成する.

これにより，語と語の関係を計量するメタデータ空間を構成が可能となる.

5. 数式データを対象とした複合連想検索

類似数式検索機能と数学用語等の言葉を適用した単語間関連連想検索機能を連結して，検索システムを実現することにより，言葉と数式からなる問い合わせに合致した統合された検索結果を得ることを考えた．数式と言葉に対して類似検索機能を用いることで，個々に検索機能を用いる場合よりも優れた結果が得られると考えられる.

5.1 実 現 方 式

数式を対象とした複合連想検索方式の全体概要図を図 5 に示す．本方式は次の流れで実現される.

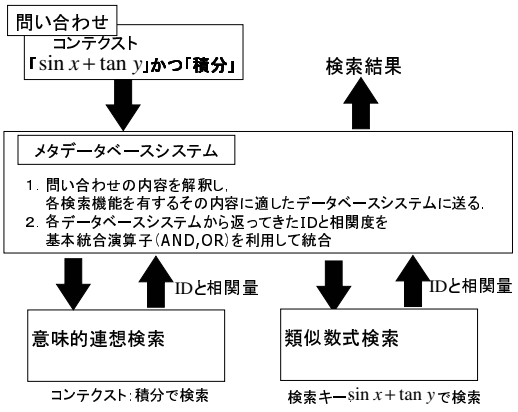


図 5　複合連想検索方式の全体図
Fig. 5　a picture of Composite Association Retrieval

Step1： 問い合わせ発行

ユーザに検索のための問い合わせを入力してもらう．本方式では，ユーザからの問い合わせは，数式と言葉 (数学用語) から与えられることを想定している．

Step2： 問い合わせの振り分け

ユーザからの問い合わせを数式は類似数式検索機構に，言葉は意味的連想検索機構に振り分ける．

Step3： 各検索機構による結果の統合

各検索機構の結果を基本統合演算子によって統合し，問い合わせに対する検索結果としてユーザに返す．

基本統合演算子「AND」,「OR」について以下に述べる．本システムで対象としている検索機構は，問い合わせに対して，検索対象データの相関量を返すものを想定している．ユーザに出力の際に，この相関量でソートをすることにより，問い合わせに近いものから順に出力することができる．ここでは，独立に実装されている検索機構 A と検索機構 B の検索結果の統合を考える．

検索機構 A で検索した結果を $\boldsymbol{A} = (a_1, a_2, \cdots, a_n)$，検索機構 B で検索した結果を $\boldsymbol{B} = (b_1, b_2, \cdots, b_n)$ とおく．なお，a_i は検索機構 A で検索したそれぞれの検索対象データの相関量の値，b_i は検索機構 B で検索したそれぞれの検索対象データの相関量の値，n は検索対象データの数である．ただし，$0 \leqq a_i \leqq 1$, $0 \leqq b_i \leqq 1$ とする．

このとき,「AND」統合演算子 $\otimes$ を以下のように定義する．

$$\boldsymbol{A} \otimes_{i=1}^{n} \boldsymbol{B} = (a_1 \times b_1, a_2 \times b_2, \cdots, a_n \times b_n) \qquad (3)$$

また,「OR」統合演算子 $\oplus$ を以下のように定義する．

$$\boldsymbol{A} \oplus_{i=1}^{n} \boldsymbol{B} = (\frac{a_1 + b_1}{2}, \frac{a_2 + b_2}{2}, \cdots, \frac{a_n + b_n}{2}) \qquad (4)$$

「OR」の演算子においてすべての値に除算を行っているが，これは「AND」の結果とスケールを合わせるための正規化である．

5.2 評価実験

5.2.1 実験方法

本方式に基づくシステムを構築し実験を行った．単語間関連連想検索機能を実現するためのデータには,「マグロウヒル大学演習　線形代数」[15] の索引を用いて作成したデータ行列を使った．具体的には，索引を用いて各ページを索引に出現する 376 語で特徴づけを行った．ただし，索引で参照されないページについては省略した．この操作により，149 行 376 列の初期行列となった．

検索対象の数式データとして，MathML で書かれた 36 個の数式とそれぞれの数式に対して付与された言葉を用いた．数式と言葉は「線形代数学の基礎」[16] より選んだ．数式データは，ID と数式と言葉のデータを 1 セットにしている．数式データの例を表 1 に示す．

実験環境を表 2 に示す．また，使用言語は Perl と Java である．

5.2.2 実験結果

類似数式検索機構と単語間関連連想検索機構のそれぞれの検索結果として問い合わせ「$\mathbf{a}\cdot\mathbf{b} = \|\mathbf{a}\|\,\|\mathbf{b}\|\cos\theta$」の場合，問い合わせ「内積」の場合をそれぞれ表 3，表 4 に示す．そして，複合連想検索の検索結果として問い合わせ“「$\mathbf{a}\cdot\mathbf{b} = \|\mathbf{a}\|\,\|\mathbf{b}\|\cos\theta$」and「内積」”の場合をそれぞれ表 5 に示す．これらは，検索結果の上位 5 位を示している．

表 1　実験用の数式データ例.

ID	式	言葉
1	$y = f(x)$	1 対 1 の写像
2	$\mathbb{R}^n$	n 次元，R^n
3	$\|\|a\|\|$	ノルム
4	$\cos\theta = \frac{a \cdot b}{\|\|a\|\|\|\|b\|\|}$	角，内積，ノルム
5	$Ker(f) \equiv \{x \in \mathbb{R}^n \| f(x) = 0\}$	核，Ker_f
⋮	⋮	⋮

表 2　実験環境.

(サーバ)		
OS：	Solaris8	
HTTP サーバ：	Apache	version1.3.17
言語：	Perl	version5.6.1
	Java	version1.4.1

(クライアント)		
OS：	Windows XP	Home Edition
Web ブラウザ：	Internet Explorer	version6.0
プラグイン：	MathPlayer	version1.0

表 3　実験結果 1(類似数式検索機構).

問い合わせ：「$\mathbf{a}\cdot\mathbf{b} = \|\mathbf{a}\|\,\|\mathbf{b}\|\cos\theta$」

順位	ID	式	相関量
1	(33)	$\mathbf{a}\cdot\mathbf{b} = \|\mathbf{a}\|\,\|\mathbf{b}\|\cos\theta$	1.000
2	(4)	$\cos\theta = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$	0.707
3	(17)	$Ax = \lambda x$	0.409
4	(20)	$B = P^{(-1)}AP$	0.286
5	(6)	$AA^{(-1)} = E$	0.273

表 4　実験結果 2(単語間関連連想検索機構).

問い合わせ：「内積」

順位	ID	言葉	相関量
1	(33)	内積	0.791
2	(3)	ノルム	0.780
3	(4)	角，内積，ノルム	0.721
4	(18)	三角不等式，ノルム	0.312
5	(19)	正規化，単位ベクトル	0.298

表 5　実験結果 3(複合連想検索).

問い合わせ：「$\mathbf{a}\cdot\mathbf{b} = \|\mathbf{a}\|\,\|\mathbf{b}\|\cos\theta$」and「内積」

順位	ID	式	言葉	相関量
1	(33)	$\mathbf{a}\cdot\mathbf{b} = \|\mathbf{a}\|\,\|\mathbf{b}\|\cos\theta$	内積	0.791
2	(4)	$\cos\theta = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$	角，内積，ノルム	0.510
3	(3)	$\|\mathbf{a}\|$	ノルム	0.053
4	(19)	$\|\mathbf{a}\| = 1$	正規化，単位ベクトル	0.039
5	(18)	$\|\mathbf{a}+\mathbf{b}\| \leqq \|\mathbf{a}\| + \|\mathbf{b}\|$	三角不等式，ノルム	0.021

5.2.3 考　　察

実験結果 1 において，類似している式が上位に上がっていることがわかる．また，3 番目以降の相関量は 2 番目の値に比べて小さい値となっている．これは数式のみの検索でも比較的よい結果を示している反面，大きく値の差となって現れないと考えることができる．

実験結果 2 において，最上位の「内積」の次に「ノルム」があがっている．これは意味的連想検索において,「内積」という言葉そのものを入れなくても「ノルム」という言葉によって「相似」が検索されたことを意味している．

実験結果 3 において，3 番目以降の相関量が 2 番目の相関量に比べてかなり小さくなっている．また実験結果 2 において 2 番目にあった ID(3) のデータは,「AND」の統合演算によって 3 番目に順位が下がり，相関量も小さくなっている．これは統合演算によって，言葉と数式の両方が適合している数式データが上位にあがることを表している．

したがって，本方式による検索結果は，類似数式検索と単語間関連連想検索を個別に適用した場合よりも適合率のよい結果が得られると考えられる．

6. おわりに

本稿では，独立した検索機構の統合方式として数式を対象とした複合連想検索の実現方法について示した．本方式を適用することにより，ユーザは言葉と数式との組み合わせにより，対象とする数式からなるコンテンツの検索が可能となり，ユーザの意図と合致した検索が可能となると考えられる．

今後の課題として，再現率や適合率による本方式の定量的な評価，それぞれの検索機能の更なる改善，数式入力インタフェースの実現，本方式を実現した，数式を含んだ文書を対象とした統合的なデータベースシステムの実現があげられる．また，数式の構造を考慮した検索手法の確立，統合方式の更なる検討や大規模なデータを対象としたシステムの実現があげられる．

本方式によって，言葉と数式からなる問い合わせによる数式を含む文書，および文書における数式の検索が実現できると考えられる．例えば，論文や公式集・数値計算マニュアルの検索への応用がある．また，本方式は XML と言葉からなるデータであれば，他の分野でも適応できると考えられる．

文　　献

[1] M.W. Berry, and S.T. Dumais, and G.W. O'Brien, "Using linear algebra for intelligent information retrieval," SIAM Review, vol.37, no.4, pp. 573–595, December 1995.

[2] "W3C Math Home," W3C. http://www.w3.org/Math/

[3] 岸本貞弥，中西崇文，櫻井鉄也，北川高嗣，栃木敏子，"MathML を用いた類似数式検索方式の実現，" 第 14 回データ工学ワークショップ (DEWS2003) 論文集，no.6-P-07，Mar 2003.

[4] 三枝義典，阿部昭博，佐々木建昭，増永良文，佐々木睦子 "数式処理システム GAL における数学公式データベースのインデキシング手法，" 信学論 (D–I)，vol.J74-D-I，pp.577–585，Aug 1991.

[5] "World Wide Web Consortium," W3C. http://www.w3.org/

[6] 中西崇文，岸本貞弥，櫻井鉄也，北川高嗣 "特定分野を対象とした連想検索のためのページベースのメタデータ空間生成方式，" データベースと Web 情報システムに関するシンポジウム (DBWeb2003) 論文集，pp.45–52，Nov 2003.

[7] "TtM, a TeX to MathML translator," Ian Hutchinson. http://hutchinson.belmont.ma.us/tth/mml/

[8] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. and Management, vol.24, no.5, pp.513–523, 1988.

[9] G. Salton, and C. Buckley, "Improving retrieval performance by relevance feedback," J. Am. Soc. Inf. Sci., vol.41, no.4, pp.288–297, June 1990.

[10] T.Kitagawa, Y.Kiyoki, "The Mathematical Model of Meaning and its Application to Multiodatabase Systems," *Proceedings of 3rd IEEE International Workshop on Research Issues on Data Engineering, Interoperability in Multidatabase Systems*, pp.130-135, April 1993.

[11] 清木康, 金子昌史, 北川高嗣: "意味の数学モデルによる画像データベース探索方式とその学習機構, " 信学論，D-II，vol.J79-D-II, no.4，pp.509–519，1996.

[12] Y.Kiyoki, T.Kitagawa, and T.Hayama, "A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning," Multimedia Data Management – using metadata to integrate and apply digital media –, McGrawHill, A. Sheth and W. Klas(editors), Chapter 7, 1998.

[13] 宮川祥子，清木康: "特定分野ドキュメントを対象とした意味的連想検索のためのメタデータ空間生成方式 ," 情処学論 : データベース, vol.40, no.SIG5(TOD2), pp.15-27，1999.

[14] 河本穣，清木康，吉田尚史，藤島清太郎，相磯貞和: "医療分野ドキュメント群を対象とした意味的連想検索空間の実現方式 ," 日本データベース学会 Letters，Vol.1，No.2，pp.12-15，2003.

[15] Seymour Lipschutz, 加藤明史訳，マグロウヒル大学演習　線形代数 (上)(下)，オーム社，東京，1995.

[16] 水本久夫，線形代数学の基礎，培風館，東京，2000.