数式データを対象とした複合連想検索システムの実現

An Implementation Method of Composite Association Retrieval System for Data of Mathematical Formulas

中西 崇文♥ 岸本 貞弥♥
村方 衛♦ 大塚 透†
櫻井 鉄也♠ 北川 高嗣♠

Takafumi NAKANISHI Sadaya KISHIMOTO
Mamoru MURAKATA Toru OTSUKA
Tetsuya SAKURAI Takashi KITAGAWA

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．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を実現している．また，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専門用語に対する意味的連想検索を実現している．本稿では，この類似数式検索と，用語に対する意味的連想検索を連結した複合連想検索について示す．また，この検索に適した GUI を提案する．

Mathematical Markup Language (MathML) was released by World Wide Web Consortium (W3C). We can use mathematical contents on the Web. We implement a function of similarity-based retrieval for mathematical formulas with Latent Semantic Indexing (LSI) utilizing formulas encoded by MathML as queries. In addition, we have implemented the semantic associative search applied to mathematical terms. In this paper, we present composite association retrieval system for data of mathematical formulas and propose a GUI system which is suitable for this retrieval system.

1. はじめに

現在，コンピュータネットワーク上に科学技術分野を対象とした多種多様な情報群が広域に遍在しつつある．また，情報群は増加を続けており，それらのデータ群は，知識・情報の源として重要な存在となっている．このような環境下で，これらのドキュメント群を対象とした，高度な検索方式と知識の発掘方式が重要となっている．

しかしながら，科学論文等，科学技術分野の情報の多くには数式が含まれており，それらの数式の持つ意味が重要となる場合が多い．このような科学論文等の数式を含んだドキュメントについて，意味的な内容を反映した検索を行うためには，数式を対象とした類似検索方式の実現が重要であると考えられる．

これまで，数式や公式を対象とした検索方式として，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[1]にて実現されている．数式は，どの演算子が含まれているか，どのような構造になっているか，どのような分野で使われるかなど，見方によって数式の意味合いが変わってくることが多い．例えば，$F = ma$ という式は，構造から見ると単なるmとaの掛け算を表す式である．しかし，応用範囲を考えれば，物理学で言えば「運動の法則」を表す公式であり，もしくは，買い物をしている人にとっては，単なる単価a のものをm 個購入したときの価格Fという式でもある．つまり，数式を対象とした類似検索方式を実現するためにはこのような見方によって意味合いが変化する，数式の多角性を導入することが重要となると考えられる．

本稿は，数式を対象とした複合連想検索方式の実現について示す．本方式は，ユーザが発行した複数の問い合わせに対して，それぞれの問い合わせに合致した複数の検索方式，つまり複数の計量系で計量し，それらの結果を統合する．このことによって，数式と問い合わせとの関連性を様々な見方から計量を行い，かつ，その結果を統合することにより，ユーザの見方に合致した検索結果を得ることが可能であると考えられる．本稿では，MathML を用いた関数や演算子，数学記号の出現による類似数式検索機構と，数式を表す言葉による意味的連想検索機構とを統合した複合数式検索システムを実装する．またこれらのシステムを用いて，本方式の有効性を検証する．ここで複合連想検索とは，様々な計量系から出てくる検索結果をAND やOR などの演算子を用いて結合し，検索結果のリストを得る検索のことである．

2. 各検索機構の実現

2.1 類似数式検索機構の実現方式

本節では，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．

2.1.1 類似数式検索機構の概要

本方式の流れとしては以下の通りである．

(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 2．1．2 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

♥ 学生会員 筑波大学大学院システム情報工学研究科
{takafumi, kishimoto}@mma.cs.tsukuba.ac.jp
♦非会員 筑波大学大学院システム情報工学研究科
murakata@mma.cs.tsukuba.ac.jp
†非会員 富士ゼロックス
♠ 非会員 筑波大学大学院システム情報工学研究科
{sakurai,takashi}@cs.tsukuba.ac.jp

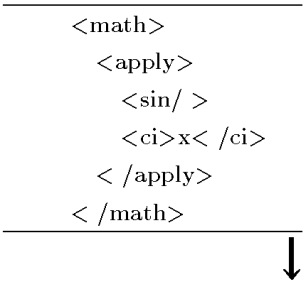
メタデータを抽出する．
(3) 類似度を計量
上記項目(1),(2) により抽出されたメタデータから，類似度を計量し，その値の大きい順にソートする．これにより，問い合わせの数式とタグの構成が類似した数式が検索される．本方式では，類似度の尺度としてコサイン尺度を用いている．

2.1.2 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$\mathbf{d}_i$ (i=1, 2,…, 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．

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

t_{1i}, t_{2i}, …,t_{mi} は対応するMathML のタグの出現頻度を表す．例として図１ 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 [2], [3] を用いて重み付けを行う．



$\sin x$ 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１ sin x の例
Fig. 1 Example of sin x.

2.2 数学用語等の言葉を適用した意味的連想検索機構の実現方式

本節では，数学用語等の言葉を適用した意味的連想検索機構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[4][5][6]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．

2.2.1 意味の数学モデルの概要

(1) メタデータ空間 *MDS* の設定
検索対象となるメディアデータをベクトルで表現したデータを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 *MDS*)を設定する．本稿では，このメタデータ空間 *MDS* を複数の書籍の索引を用いることによって生成する方式について提案している．
(2) メタデータをメタデータ空間 *MDS* へ写像
設定されたメタデータ空間 *MDS* へメディア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，同じ空間に検索対象データのメタデータが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ノルムとして計算することが可能となる．
(3) メタデータ空間 *MDS* の部分空間(意味空間)の選択
検索者は与える文脈を複数の単語を用いて表現する．検索者が与える単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 *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 *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(以下，意味空間)が選択される．
(4) メタデータ空間 *MDS* の意味空間における相関の定量化
選択されたメタデータ空間 *MDS* の部分空間(意味空間)において，メディアデータベクトルのノルムを検索語列との相関として計量する．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．この意味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．

2.2.2 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[7]を示す．本方式では，検索対象を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） 初期データ行列の設定
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．索引に出現する語を特徴語とみなし，索引情報から各ページ番号を用いて特徴付ける．

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

ここで i はページ番号，f_{ik} は特徴語に対応したページ番号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番号を参照している場合：”1”
- 索引中で特徴語がそのページ番号を参照してない場合：”0”

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， 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_1を生成する．
まず，章，節の番号を特徴語として初期データ行列 M_0 を修正，追加する．章，節番号について該当ページを全て”1”，それ以外のページを”0”と特徴付ける．例えば,23 ページが2章3節に該当する場合,「2」,「2-3」を特徴語として，23 ページの「2」,「2-3」に”1”と特徴付ける．
以上により, m 行$n + R$ 列のデータ行列M_1 を生成できる．ここで，R は章，節番号を特徴として付け加えた分である．

3. 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意味的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．

3.1 数式データを対象とした複合検索実現方式

本方式は次の流れで実現される．

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 $A=(a_1,a_2,\cdots,a_n)$，検索機構 B で検索した結果を $B=(b_1,b_2,\cdots,b_n)$ 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，n は検索対象データの数である．ただし，$0\le a_i\le 1$，$0\le b_i\le 1$ とする．このとき，「AND」統合演算子を以下のように定義する．

$$A\otimes_{i=1}^{n} B=(\sqrt{a_1b_1},\sqrt{a_2b_2},\cdots,\sqrt{a_nb_n})\cdot$$

また，「OR」統合演算子を以下のように定義する．

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

3.2 入力 GUI

本方式では数式の問い合わせに MathML を用いている．しかしながら，数式を MathML で記述するには MathML タグとその文法を知っておく必要があり，検索する際ユーザに入力させるのは現実的ではない．そこで複合連想検索システムでは，数式の入力をより簡単にするために GUI による入力を考案した．この GUI は「拡張可能な GUI システム“exGUIde”」をもとに作っている．拡張可能な GUI システムとは，数式の入力メニュー・出力形式をユーザが自由にカスタマイズできるシステムであり，様々な数理ソフトウェアの利用支援が可能である．カスタマイズは XML 定義ファイルと XSLT スタイルシートにより行う．拡張可能な GUI システムを実装した Java アプリケーションをここでは“exGUIde”と呼ぶ．図 2 にその概観を示す．

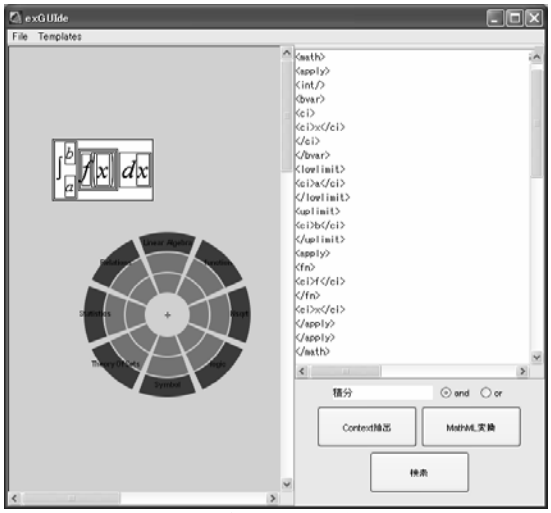


図 2 exGUIde
Fig.2 exGUIde.

4. 実験例

本方式に基づくシステムを構築し実験を行った．本実験では，意味的連想検索機能を実現するための空間生成のためのメタデータとして「基礎物理学第 2 版」[8] の索引を用いて作成したデータ行列を作成し，メタデータ空間を生成した．

検索対象の数式データとして，MathML で書かれた 325 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「Essential 物理学」[9] 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．

ここで，実験例として「*F=mg*」,「運動方程式」に注目する．類似数式検索機構と意味的連想検索機構のそれぞれの検索結果として問合せ「*F=mg*」の場合をケース 1，問合せ「運動方程式」の場合をケース 2 として，それぞれ表 1，表 2 に示す．ただし，表 1 では 1 番目の順位のデータが多いので，5 件を超えて示した．そして，複合連想検索の検索結果として問合せ“「*F=mg*」and「運動方程式」”の場合をケース 3 として表 3 に示す．これらは，検索結果の上位 5 件を示している．

ケース 1 において，いずれも積の形をした数式であり，類似している式が上位に上がっていることがわかる．上位 6 件は類似した数式であり値の差が全くない. ケース 2 において，上位 5 件中 3 件には問合せにある「運動方程式」という言葉が入っているが，2 番目と 3 番目のデータには入っていない．しかしながら，「質量」「重力加速度」「万有引力定数」は「運動方程式」と関わりの深い言葉である．ケース 3 において，最上位に現れている ID(24) のデータはケース 1 とケース 2 の上位にも現れていた．しかしながら，ケース 1 に現れていた式は ID(24)を除いてどれも上位 5 件には入っていない．代わりに，ケース 2 の上位に現れていた ID(25)の式が 2 番目に出力されている．したがって，これらの実験例から一方の検索機構からの出力結果により，他方の検索結果がフィルタリングをかけられたような結果が得られたことがわかる．他のコンテクストを与えた場合でも同様の結果が得られた．

表 1 実験結果 1(ケース 1)
Table.1 Experimental results1 (case.1).

問合せ：「$F = mg$」			
順位	ID	式	相関量
1	(24)	$F = mg$	1.000
1	(30)	$v = gt$	1.000
1	(120)	$f = ce$	1.000
1	(123)	$f = n\theta$	1.000
1	(303)	$E = \hbar\omega$	1.000
1	(305)	$p = \hbar k$	1.000

表 2 実験結果 2(ケース 2)
Table.2 Experimental results2 (case.2).

問合せ：「運動方程式」			
順位	ID	言葉	相関量
1	(48)	運動方程式	0.866
1	(293)	運動方程式	0.866
3	(24)	質量，重力加速度	0.646
3	(25)	万有引力定数，質量	0.646
5	(60)	運動方程式，運動量	0.629
5	(118)	減衰運動，運動方程式	0.629
5	(119)	減衰運動，運動方程式	0.629

表 3 実験結果 3(ケース 3)
Table.3 Experimental results3 (case.3).

問合せ：「$F = mg$」and「運動方程式」				
順位	ID	式	言葉	相関量
1	(24)	$F = mg$	質量，重力加速度	0.804
2	(25)	$F = G\frac{Mm}{r^2}$	万有引力定数，質量	0.563
3	(118)	$\ddot{x} + 2\gamma\dot{x} + \omega^2 x = 0$	減衰運動，運動方程式	0.441
4	(293)	$p = \frac{m}{\sqrt{1-\beta^2}}v$	運動方程式	0.410
5	(119)	$x = Ae^{-\gamma t}\cdots$ (b)*	減衰運動，運動方程式	0.399

*表中に収まらないため，数式 (b) は以下に別記した．

$$x = Ae^{-\gamma t}\sin(\sqrt{\omega^2 - \gamma^2}t + a)$$

5. まとめと今後の課題

本稿では, 数式データを対象とした複合連想検索について示し，この検索に適した GUI を提案した．また，実験例を示し考察を行った．GUI を用いることで数式の問い合わせが容易に作成でき，本システムの有用性が高まると期待できる．また，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，より大きな数式データに対する本システムの適用，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，数式の構造を考慮した検索手法の検討が挙げられる．

[文献]

[1] 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.

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

[3] 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.

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

[5] 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.

[6] 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 –, Mc-GrawHill, A. Sheth and W. Klas(editors), Chapter 7, 1998.

[7] 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書籍の索引部を用いたメタデータ空間生成方式,” 電子情報通信学会論文誌，VOL.J88-D1 No.4 , pp.840–851, 2005.

[8] 後藤憲一, 小野廣明, 小島彬, 土井勝, 基礎物理学第 2 版，共立出版，東京，2004.

[9] 阿部龍蔵, Essential 物理学, サイエンス社, 東京, 2003.

中西 崇文 Takafumi NAKANISHI
筑波大学大学院システム情報工学研究科在学中．2001 年筑波大学第三学群情報学類を卒業．マルチメディアシステムに関する研究に興味を持つ．情報処理学会学生会員．電子情報通信学会学生会員，日本データベース学会学生会員．

岸本 貞弥 Sadaya KISHIMOTO
筑波大学大学院システム情報工学研究科在学中．数理ソフトウェア利用支援の研究に興味を持つ．日本データベース学会学生会員．

村方 衛 Mamoru MURAKATA
筑波大学大学院システム情報工学研究科在学中. GUI の研究に興味を持つ．

大塚 透 Toru OTSUKA
2005 年筑波大学大学院博士課程システム情報工学研究科修士取得中退．現在，富士ゼロックス勤務．

櫻井 鉄也 Tetsuya SAKURAI
1986 年名古屋大学院工学研究科博士課程前期課程情報工学専攻修了．同年同大学助手．筑波大学講師を経て，現在，筑波大学大学院システム情報工学研究科助教授．工学博士．大規模固有値問題の並列解法，非線形方程式の反復解法，および数理ソフトウェアの利用支援の研究に従事．1996 年日本応用数理学会論文賞受賞．日本応用数理学会，情報処理学会会員．

北川 高嗣 Takashi KITAGAWA
筑波大学大学院システム情報工学研究科教授．1978 年名古屋大学工学部卒業．1983 年同大学院工学研究科博士過程修了. 工学博士. スタンフォード大学計算機科学科客員研究員，愛媛大学理学部数学科講師，筑波大学電子・情報工学系助教授を経て現在に至る．数値解析，逆問題，マルチメディア情報システムの研究に従事．日本応用数理学会会員．