人工知能学会研究会資料
SIG-KBS-B501-01

部分構造を考慮した数式理解支援システムの提案

A Subexpression-based System to Support Mathematical Expression Understanding

澁谷 海渡[1*]　白川 真一[2]　大原 剛三[3]　豊田 哲也[3]
Kaito Shibuya[1]　Shinichi Shirakawa[2]　Kouzou Ohara[3]　Tetsuya Toyota[3]

[1] 青山学院大学大学院理工学研究科
[1] Graduate School of Science and Engineering, Aoyama Gakuin University
[2] 筑波大学システム情報系
[2] Faculty of Engineering, Information and Systems, University of Tsukuba
[3] 青山学院大学理工学部
[3] College of Science and Engineering, Aoyama Gakuin University

Abstract:
This work aims at helping students understand complex mathematical expressions that appear in academic documents such as research papers. To this end, we focus on subexpressions of an expression given as a query, and try to annotate them using information about known similar expressions, while most existing approaches tend to find out an expression similar to the whole of a query. Assuming that an expression is given in the form of Content Markup of MathML, we construct its DOM tree, and extract subtrees each of which corresponds to one of its subexpressions from the whole tree. Then, we search a database storing known expressions associated with their meta-data for ones similar to the subexpressions. To evaluate the usefulness of our approach, we conducted experiments using actual mathematical expressions collected from a web site and a mathematical textbook, and confirmed that considering subexpressions can bring on more information helpful for better understanding of a given query expression compared to annotating itself.

1　はじめに

現在，日本の教育現場において，「数学嫌い」「数学離れ」は大きな社会問題となってきている [1, 2]．この問題は大学の理系学部に進学した学生も例外ではなく，数学に対して苦手意識を抱いている学生は数多くいる．特に理系の学生にとっては，学術論文や研究資料の内容を知る上で数式を正確に理解する事は非常に重要である一方，こういった数学に対して苦手意識を抱いている多くの学生にとって，学術論文や研究資料などにでてくる複雑な数式を短時間で理解することは必ずしも容易ではない．この問題を解決する手段としては，対象となる数式を理解する上で必要な情報を提示するシステムなどが考えられる．実際にそのようなシステムを構築するためには，その数式が数学的にどのような意味を持った数式なのかを示す必要があり，その実現には，対象の数式と構造的・意味的に類似した数式を検索する技術が必要になる．しかし，現在のテキストベースの検索エンジンでは分数や指数をはじめとする数式独自の構造的特徴を扱うことができないため，複雑な数式を適切に検索することは難しい．

このような類似数式検索に関しては，これまで数式画像を対象とした検索手法はいくつか提案されているが [3, 4]，画像データでは数式の意味的な情報が失われるため数式の内容を考慮した検索は困難である．一方，1998 年 4 月に World Wide Web Consortium（W3C）より数式の構造情報を表現可能なマークアップ言語である Mathematical Markup Language（MathML）[1]が公表されて以降，MathML を利用した類似数式検索の研究が盛んに行われてきた [5, 6, 7, 8]．MathML には Web 上で数式を表現する事に特化した Presentation Markup，数式の意味情報を表現することに特化した Content Markup という 2 種類の表記方法が存在する．岸本らは，Content Markup のタグの出現頻度から抽出したメタデータに対して LSI（Latent Semantic Index）

*連絡先：青山学院大学大学院理工学研究科
〒 252-5258 神奈川県相模原市中央区淵野辺 5-10-1
E-mail: c5615137@aoyama.jp

[1] http://www.w3.org/Math

を適用することにより数式検索を行うシステムを実現している [5]．橋本らは，Presentation Markup で表記された数式の DOM（Document Object Model）構造の Xpath（XML Path Language）[2]を用いた転置インデックスを作成し，このインデックスを利用して数式を検索するシステムを実現している [6]．横井らは，科学論文中にでてくる数式を対象に，MathML を用いた木構造マッチングと数式の周りに存在するテキスト情報を利用する事で，数式の構造的特徴だけでなく数式の意味的特徴も考慮した検索システムを実現している [7]．また，Nghiem らは，Presentation Markup と Content Markup の両方のクエリに対して類似した数式を提示するシステムを実現している [8]．

このように，問合せ対象の数式全体の特徴に基づいた類似数式検索に関する研究はこれまで盛んに行われてきた．一方，従来手法がいずれも数式単位の検索であるのに対し，学術論文などに現れる数式は，基本的な数式の組み合わせになっていることが多い．言い換えると，数式全体が何かしらの数学的概念定義に直接対応する式になってはいないが，その数式のある部分構造は何かしらの数学的概念に対応する場合が多分にある．そこで本研究では，数式全体の理解を支援するために，数式中の部分構造に対してもその解釈を提示できる数式アノテーションシステムの実現を目的とする．そのために，MathML 形式の数式が容易に木構造に変換可能であり，その部分木が数式の部分構造に相当することに着目し，系統的に抽出した部分木ごとに類似数式検索をかけることで，元の数式の部分構造に対するアノテーションを実現することを考える．本稿では，提案手法によって問合せ数式の部分構造に対してどの程度のアノテーションが可能であるかという点について実験的に評価したので，その結果について報告する．

2　提案手法

本研究では，図 1 に示すような，数式全体では特定の数学的概念に対応しないが，その部分構造は何かしらの数学的概念に一致するような数式を対象に，数式の理解を支援するための数式アノテーションシステムの実現を目指す．そのため，図 2 に示すように，クエリとなる問合せ数式から部分構造を抽出し，それらに対する特徴ベクトルを生成し，それを検索対象となる数式データベース中の各数式の特徴ベクトルと比較することで，類似度の高い数式に対する情報を対応する部分構造に対するアノテーションとして利用する．以下，数式データベースの作成方法の概要，および問合せ数式から部分構造の抽出，数式間の類似度計算について述べる．なお，本研究では，類似数式検索の基本的な枠組みは岸本らの手法 [5] を利用している．

[2]http://www.w3.org/TR/xpath/

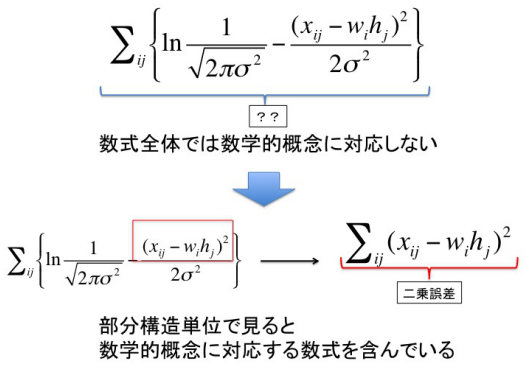

図 1: 部分構造単位で意味のある数式を含んでいる例

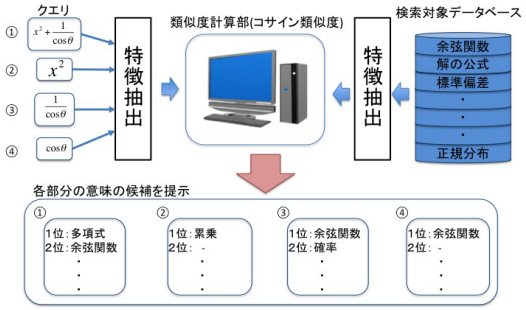

図 2: 提案システムの概要

2.1　検索対象数式データベースの作成

本研究で用いる検索対象数式データベースは，その定義が明確に与えられている数式のみから構成するものとし，各数式に対しては事前に固定長の属性ベクトルを生成しておく．その結果，データベース中の数式の総数を N，属性ベクトルの次元数を M としたとき，データベース全体は $M \times N$ の行列 $\boldsymbol{D}$ として表現できる．具体的には，各数式を表す M 次元の特徴ベクトルを $\boldsymbol{p}_i (i = 1, \cdots, N)$ としたとき，検索対象数式データベースは次式のように表すことができる．

$$\boldsymbol{D} = (\boldsymbol{p}_1^T,\ \boldsymbol{p}_2^T, \cdots,\ \boldsymbol{p}_N^T) \qquad (1)$$

数式に対する特徴量としては，その数式を Content Markup で表現した場合の MathML タグの出現頻度を用いる．例として，$\sin x$ に対する Content Markup を図 3 に示し，それに対するタグの数え上げ例を表 1 に示す．本

```
<math>
  <apply>
 <sin/>
 <ci>x</ci>
  </apply>
</math>
```

図 3: $\sin x$ の Content Markup による表記

表 1: $\sin x$ のタグの種類とその頻度のカウント

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

研究では，$M = 27$ 個のタグを数式を表す特徴として用いた．以下，特徴ベクトルに対応するタグのベクトルを $(t_1, \cdots, t_M)$ とする．ただし，実際には，各数式データをより明確に特徴づけるために，各特徴量 p_{ij} をテキストマイニングで用いられる TF-IDF に基づいて重み付けたものを利用する．具体的には，特徴ベクトルの j 番目の要素に対応するタグ t_j が数式データベース中の g_j 個の数式に現れた場合，p_{ij} の代わりに次式で与えれる p'_{ij} を用いる．

$$p'_{ij} = p_{ij} \times \log(\frac{N}{g_j}) \tag{2}$$

したがって，$\boldsymbol{p}'_i = (p'_{i1}, \cdots, p'_{iM})$ としたとき，数式データベース $\boldsymbol{D}$ は次式のように定義されるものとする．

$$\boldsymbol{D}' = (\boldsymbol{p}'^T_1, \cdots, \boldsymbol{p}'^T_N) \tag{3}$$

実際には，各数式に含まれるタグの種類は限定的であるため，各特徴ベクトルは疎なベクトルとなり，結果として数式データベース $\boldsymbol{D}'$ 自体が疎な行列になる．このような疎な特徴ベクトルを用いて類似度計算をした場合，適切な類似度を求めることが困難となることが予想されるため，実際には，各数式に対する特徴ベクトル $\boldsymbol{p}'_i$ を k $(\leq M)$ 次元の特徴ベクトル $\boldsymbol{p}'^{(k)}_i$ に次元圧縮した $k \times N$ 行列 $\boldsymbol{D}'^{(k)}$ を利用する．ここでの次元圧縮としては，文献 [5] と同様に LSI を用いるものとする．すなわち，$\boldsymbol{D}$ を特異値分解し，降順の特異値の上位 k 個に対する座標系に $\boldsymbol{D}'$ を射影することにより，$\boldsymbol{D}'^{(k)} = ((\boldsymbol{p}'^{(k)}_1)^T, \cdots, (\boldsymbol{p}'^{(k)}_N)^T)$ を求める．

2.2 問合せ数式から部分構造式の抽出

問合せ数式は，検索対象となる数式と同様に Content Markup 形式で与えられることを前提とし，そのタグ構造を DOM ツリーとして表現し，そこから部分木を抽出することで，元の数式の部分構造を抽出する．ここで，DOM ツリーの中間ノードは演算子などを表す MathML のタグに相当し，葉ノードは具体的な定数項や変数を表す．なお，本研究における部分構造の定義は，「単項演算子・二項演算子・三項演算子・その他の演算子のいずれかを 1 つ以上含む部分木」とする．実際には，深さ優先探索で DOM ツリーを根からたどり，この部分構造の定義を満たす部分木を抽出する．部分構造の抽出例を図 4 に示す．

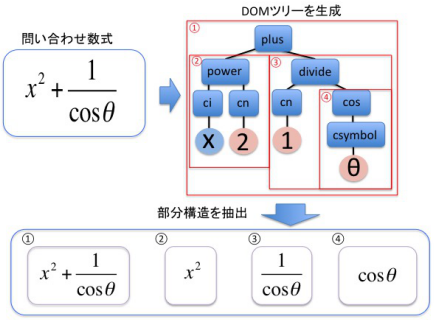

図 4: 問合せ数式からの部分構造の抽出例

問合せ数式から L 個の部分構造が抽出された場合，各部分構造式は検索対象データベース中の数式と同様に，次式で定義される M 次元の特徴ベクトル $\boldsymbol{q}'_l$ $(l = 1, \cdots, L)$ で表現する．

$$\boldsymbol{q}'_l = (q'_{l1}, \cdots, q'_{lM}) \tag{4}$$

ここで，$\boldsymbol{q}'_l$ の各成分 q'_{lj} は，部分構造 $\boldsymbol{q}'_l$ におけるタグ t_j の出現頻度 q_{lj} に，式 (2) で用いた IDF 値 $\log \frac{N}{g_i}$ を重みとして乗じた値とする．

2.3 数式間の類似度計算

前節で述べたように問合せ数式から各部分構造式 $\boldsymbol{q}'_l$ を抽出した後，検索対象数式データベース $\boldsymbol{D}'^{(k)}$ から類似度の高い数式を探し，その数式に対する情報を元の数式に対するアノテーションとして利用する．実際には，検索対象となる数式は前述のように k 次元に圧縮された特徴ベクトルで表現されているため，各部分構造式 $\boldsymbol{q}'_l$ も同様の座標変換により k 次元のベクトル $\boldsymbol{q}'^{(k)}_l$ に圧縮した上で，類似度を計算する．ここでは，類似度としてコサイン類似度を用い，検索対象数式 $\boldsymbol{p}'^{(k)}_i$ と問合せ数式の部分構造式 $\boldsymbol{q}'^{(k)}_l$ の類似度 $sim(\boldsymbol{p}'^{(k)}_i, \boldsymbol{q}'^{(k)}_l)$

を次式により求める．

$$sim(\boldsymbol{p'}_i^{(k)}, \boldsymbol{q'}_l^{(k)}) = \frac{\boldsymbol{p'}_i^{(k)} \cdot \boldsymbol{q'}_l^{(k)}}{|\boldsymbol{p'}_i^{(k)}||\boldsymbol{q'}_l^{(k)}|} \tag{5}$$

具体的には，各部分構造に対する検索結果は，ここで求めた類似度の降順にソートした検索対象数式の上位 x 件とする．

3 評価実験

本節では，提案システムの基本的な性能を評価することを目的とした，高校レベルの数式に対するアノテーション付与実験について述べる．ここで，高校レベルの数式を用いた理由は，適度な複雑さをもち，かつ正解ラベルを付与することのできる数式を多数用意することが容易であるからである．本実験では，問合せ数式自体に対する類似数式を検索した場合（以下，数式単位の検索と呼ぶ）と，問合せ数式の部分構造に対しても類似数式を検索対象とした場合（以下，部分構造を考慮した検索と呼ぶ）のアノテーション結果に関する再現率を比較することで，提案アプローチの有用性を評価した．また，部分構造を考慮した検索における次元圧縮の効果も検証した．なお，Content Markup の DOM ツリーへの変換には PHP5 の DOM 拡張モジュールを利用し，行列の特異値分解には redSVD[3] を用いた．

3.1 実験設定

検索対象数式として，Web 上の高校数学公式まとめサイト[4]，および高校数学問題集 [11] から何かしらの数学的概念に対応した数式 100 個を選択し，その Content Markup 形式を手作業で入力した．本実験では，各数式に対応する概念ラベル（たとえば，余弦関数など）をアノテーション対象として用いた．一方，これら 100 個の数式からなる検索対象数式データベース D に対して，本実験における問合せ数式としては，同じ高校数学問題集の数式のうち D に含まれない 100 個をランダムに選択し，数式全体，および部分構造ごとに手動で正解ラベルを割り当てた．ただし，問合せ数式 100 個のうち，実際に意味のある正解ラベルを数式全体に割り当てることが可能であったのは 59 個であった．正解ラベルの例を表 2 に示す．

本実験におけるアノテーション結果に対する正解，不正解は，検索結果の上位 x 件中（$1 \leq x \leq 5$）の数式に対するラベルに正解ラベルが含まれていれば正解，それ以外を不正解と定義した．このとき，評価指標として，以下のように定義する再現率を用いた．

表 2: 正解ラベルの例

数式	正解ラベル
$\sin x$	余弦関数
$\int x^2 dx$	不定積分
x^2	冪乗
$\sin^2 x + \cos^2 x$	三角関数の性質
$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$	2 点間の距離

表 3: 各閾値に対する次元数

閾値（τ）	未設定	0.9	0.8	0.7
次元数（k）	27	19	15	12

$$\text{再現率} = \frac{TP}{P} \tag{6}$$

ここで，P は数式全体に正解ラベルが割り当てられた問合せ数式数を表し，TP はそのうちアノテーション結果が正解となった問合せ数式数を表す．なお，TP に関しては，部分構造を考慮した検索では，数式全体に正解ラベルが割り当てられた問合せ数式のうち，アノテーション結果が正解となった部分構造が 1 つ以上あったものの数とした．提案法では，元の問合せ数式自体も 1 つの部分構造とみなして利用するため，必然的に部分構造を考慮した検索の再現率は，数式単位の検索の再現率以上の値となることに注意されたい．したがって，ここでの評価の目的は，部分構造を考慮することで，この再現率がどの程度改善されるかを調べることとなる．なお，前述のとおり，本実験においては $P = 59$ である．

一方，LSI による次元圧縮の効果を評価するために，次元数 k を閾値 τ に対して次式を満たす最小の整数とし，τ の値を 0.9，0.8，0.7 とした場合について部分構造を考慮した検索に対する再現率の変化を調べた．

$$\frac{\sum_{i=1}^{k} \sigma_i}{\sum_{i=1}^{M} \sigma_i} \geq \tau \tag{7}$$

ここで，σ_i は上位 i 番目の特異値を表す．閾値 τ の各値に対する次元数 k を表 3 にまとめる．ここで，閾値未設定とは次元圧縮をしない場合を意味する．

3.2 実験結果と考察

まず，次元数 $k = 19$（$\tau = 0.9$）の場合の数式単位の検索，および部分構造を考慮した検索それぞれに対

[3] https://code.google.com/p/redsvd/
[4] http://www.nakamura-sanyo.ed.jp/sanyo/yanase/kousiki/new/

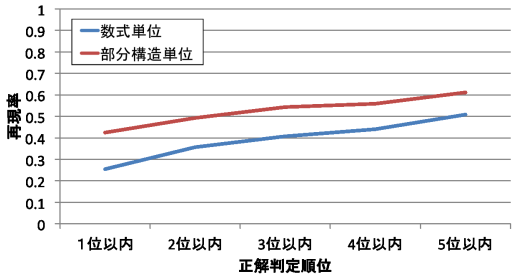

図 5: 数式単位と部分構造考慮の再現率（次元数 19）

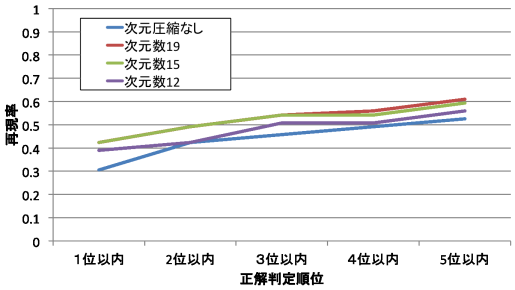

図 6: 部分構造を考慮した検索における次元数の変化と再現率の関係

する再現率を図 5 に示す．横軸は x の値に相当する．この結果から，類似度上位 1 位までを正解判定順位とした場合（$x = 1$）の再現率が，数式単位の検索では約 0.25 であるのに対し，部分構造を考慮した検索では約 0.4 と約 1.6 倍になっていることがわかる．また，類似度上位 5 位までを正解判定順位とした場合（$x = 5$）は，数式単位の検索の再現率が約 0.5 であるのに対し，部分構造を考慮した検索では約 0.6 であり，約 1.2 倍となっている．また，ここでの再現率の計算では考慮されていないが，数式全体に対して正解ラベルを割り当てることのできなかった 41 個の数式に対しても，部分構造を考慮した検索では，類似度上位 1 位までを正解判定順位とした場合には 9 個，上位 5 位までとした場合には 12 個の数式に対してアノテーション結果が正解となった．これらの結果から，問合せ数式の部分構造を考慮することで，その式を理解する上で有用な情報をより多くユーザに提示することが実際に可能であり，その点において，提案システムは数式単位の検索より有用であると言える．

次に，部分構造を考慮した検索において，次元圧縮後の次元数 k を変化させた場合の再現率を図 6 に示す．この結果から，次元圧縮をしない場合と比べて，次元数 k を 19，もしくは 15 として次元圧縮した場合の方が再現率が約 1.5〜1.2 倍ほど高くなっていることが分かる．$k = 19$ と $k = 15$ を比べると，より下位の検索結果を考慮した場合にわずかに $k = 19$ の方が再現率が高くなるものの，両者の間には大きな差は認められない．また，次元数を $k = 12$ とした場合においても，$k = 19$，$k = 15$ と比べると約 1 割ほど低い値をとっているものの，その再現率は次元圧縮をしなかった場合の値を上回っている．これらの結果から，問合せ数式の部分構造を対象とした類似数式検索においても特徴ベクトルの次元圧縮は有効に働くことがわかる．

表 4: 不正解となった検索結果の例

問合せ数式	検索結果
$x^2 + y^2 + 6x - 2y + 7$	$(a+b)(a^2 - ab + b^2)$
$\int_0^{\frac{\pi}{2}} sin^n x dx$	$\frac{4}{3}\pi r^3$
$\int_0^{\frac{3}{4}\pi} \sqrt{1 - cos4x} dx$	$\frac{4}{3}\pi r^3$

最後に，アノテーション結果が不正解となる理由について考察する．まず，問合せ数式に対して間違って得られた検索結果の例を表 4 に示す．この表の 1 つ目の例では，問合せ数式の一部が検索結果の 2 つ目のカッコ内の数式と形式が一致していることがわかる．ただし，カッコによる優先順位が考慮されていないため，ここでの一致は実際には妥当であるとは言えない．単純な部分構造のマッチングを取った場合，このような不適切な符合が生じ得るため，カッコに基づく部分構造の探索範囲の制限を設ける必要があると言える．一方，表 4 における 2 つ目と 3 つ目の例は，いずれも π を含んでいることがわかる．今回用いた検索対象数式データベース中では，記号 π を含む数式が少なかったため，対応する IDF 値が高くなり，その結果，π を含む数式間の類似度が過大評価されたと考えられる．この例にあるような誤検出を回避するためには，一致した部分構造を含むより大きな部分構造に対する類似度の変化を利用するか，文献 [7] のように，数式周辺の文脈情報を利用することなどが考えられる．

4 結論

本研究では，学術論文などに現れる複雑な数式の理解を支援するために，数式の部分構造を考慮した数式アノテーションシステムを提案した．提案システムは，従来の類似数式検索システムを基礎としつつ，数式単位での類似数式検索ではなく，数式の部分構造に対して類似数式検索を行う事で，対象数式を理解する上で有用な数式のアノテーションを可能にする．評価実験では，部分構造を考慮することで，数式単位でのアノテーションと比較して約 1.2〜1.5 倍の数式に対して何らかのアノテーションが可能であることを確認した．ま

た，従来手法でも用いられていた特徴ベクトルの次元圧縮が，部分構造式を対象とした数式検索でも有効に機能することを確認した．ただし，得られた再現率は類似度上位5位までの検索結果を用いた場合でも約0.6程度であり，今後，さらなる再現率の向上が必要と言える．また，今回の実験で利用した数式は高校レベルのものであったため，より高度で複雑な数式に対する有用性の評価も今後必要である．さらに，実用面では，携帯情報端末などでの利用を考慮すると，画像として保存した数式からMathML形式の情報を生成する手法などの実現が必要と言える．

参考文献

[1] 国立教育政策研究所：TIMSS2011 国際比較結果の概要・問題例, http://www.nier.go.jp/timss/2011/T11_gaiyou.pdf (2011).

[2] 内田昭利，守一雄：中学生の「数学嫌い」「理科嫌い」は本当か－潜在意識調査から得られた教育実践への提言－，教育実践学論集，第13号, pp.221-227，兵庫教育大学大学院連合学校教育学研究科 (2012).

[3] 山田奉子，上田洋，村上晴美，岡育生：数式理解を支援するWeb上の数式画像の検索，第28回人工知能学会全国大会，2l1-3 (2014).

[4] M. Shirmenbaatar，古賀久志, 渡辺俊典：数式画像をクエリとする類似数式検索システム，第4回データ工学と情報マネジメントに関するフォーラム, E6-2 (2012).

[5] 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（DEW2003）論文集，6-P-07 (2003).

[6] 橋本英樹，土方嘉徳，西田正吾：MathMLを対象とした数式検索のためのインデックスに関する調査，情報処理学会研究報告，2007-DBS-142，pp.55-59 (2007).

[7] 横井啓介，M. Q. Nghiem，松林優一郎，相澤彰子：意味と構造を考慮した数式検索手法の提案，第3回データ工学と情報マネジメントに関するフォーラム，B2-2 (2011).

[8] M. Q. Nghiem，G. Y. Kristianto，G. Topic，A. Aizawa：Which one is better: Presentation-based or content-based math search?，Proceedings of Conference on Intelligent Computer Mathematics, pp.200-212 (2014).

[9] G. Salton and C. Buckley：Term-weighting approaches in automatic text retrieval, Inf. Process. and Management, Vol.24, No.5, pp.513-523 (1988).

[10] G. Salton and C. Buckley：Improving retrieval performance by relevance feedback, J.Am. Inf. Sic, Vol.41, No.4,pp.288-297 (1990).

[11] 数研出版編集部：2015　実戦　数学重要問題集－数学I・II・III・A・B（理系）, 数研出版株式会社 (2014).