人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第5回)
SIG-AM-05-11

数式検索タスク NTCIR-11 Math-2

Math Retrieval Task : NTCIR-11 Math-2

Akiko Aizawa[1*]　Michael Kohlhase [2]　Iadh Ounis [3]

[1] National Institute of Informatics
[2] Jacobs University Bremen
[3] University of Glasgow

Abstract: NTCIR-11 Math-2 aims at promoting researches in mathematical content access. Based on the past experience in NTCIR-10 Math Pilot Task, we settled two major goals in our task: construction of a reusable test collection, and establishment of a research community in the field. In this paper, we introduce an outline of the task and briefly discuss possible challenges in mathematical content access.

1　はじめに

「数式」は、数や記号、およびそれらの関係を表すための記法であり、科学に不可欠な知識表現法である。数式は、演算や推論を必要とするあらゆる分野で用いられるが、その普遍性にもかかわらず、現在の検索システムでは、数式の構造や意味を適切に扱うことは困難である。数式の検索を可能にすることは、単に検索システムで扱える対象を拡大するというだけではなく、数学の知識基盤構築に向けた第一歩ともなる。

NTCIR-11 Math-2 は、数式の検索手法を研究・開発するための評価基盤の構築を通して、数式検索の研究に貢献し、研究コミュニティを活性化することを目的としている。数式検索は従来から、数学電子図書館や計算機による数学知識処理などの分野で検討されてきた [2]。また、木構造を扱うことから、データベース分野からのアプローチもされている。しかしながら、情報検索の観点からのタスク設計やの評価の試みはこれまでほとんどなかった。Math-2 では、評価基盤を共有することで、情報検索や自然言語解析の先端的な手法の適用を可能にし、数学知識アクセスの新たな展開を目指す。

本稿では以下、NTCIR-11 における Math-2 タスクの概要を紹介し、数式検索の研究課題を概観する。

2　Math-2 タスクの概要

数式検索とは、クエリまたは対象文書に数式を含む検索である。NTCIR-11 Math-2 では特に、クエリと検索対象の両方に数式を含む場合を想定してタスクを設計している。以下、Math-2 のタスクの概要について紹介する。

(1) タスクの目標

NTCIR-10 で行ったパイロットタスクの経験を踏まえ、Math-2 では、２つの大きな目標をタスク設計の指針としている。第一は、再利用可能なテストコレクションの構築である。具体的には、なるべく多様な検索手法を導入するとともに、適合性判定を複数の評価者により行うことで、タスク終了後も活用できる資源の構築を目指す。第二は、数式検索の研究コミュニティの形成と支援である。具体的には、数式検索で検討するべきタスク設計や評価手法について、数学の電子図書館や数学知識処理を専門とする研究者と、情報検索や自然言語処理を専門とする研究者が意見交換をする場を提供する。

(2) 検索対象文書

Math-2 では、MathML 形式の数式を含む英語文書を検索対象とする。MathML は、数式を計算機で処理するための表記法として、World Wide Web コンソーシアムが勧告する XML アプリケーションで [3]、数式の標準記法として広く普及しつつある。既存の数式検索システムの多くは、MathML 形式の数式を想定して設計されている。

Math-2 の検索対象文書は、LaTeX ソースの形で公開されている ArXiv.org [4] の論文から選んだ 100,000 文書である。これらの latex ソースは、arXMLiv プロジェクト [5] において MathML を含む XML 形式に機械的に変換されており、NTCIR Math の参加者は arXMLiv プロジェクトから変換済の文書を入手可能である。これらの文書中には、約 35M（１論文あたり平均 350）

*連絡先：国立情報学研究所コンテンツ科学研究系
〒 101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: aizawa@nii.ac.jp

個の数式が含まれている。ただし、この場合の「数式」は、LaTeX ソース中の数式モードの文字列を MathML 形式に機械的に変換して ID を付与したもので、変換に伴うノイズを含んでいる。オリジナルの LaTeX 表記も併記されているので、LaTeX ベースの検索システムでの参加も可能である。

なお、Math-2 の検索対象文書は、NTCIR-10 の Math パイロットタスクで用いたものと同じである。ただし、NTCIR-10 では検索の単位を「数式」に限定していたのに対して、今回のタスクでは、通常の情報検索システムによる参加を容易にするため、論文をセクションなど最小の検索単位に分割して、テキストを含む検索単位のランキング結果を評価する枠組みを検討予定である。

(3) トピック

数式検索は情報検索分野では新しい課題であることから、Math-2 では伝統的な Ad hoc 検索タスクのスタイルを踏襲している。すなわち、文書集合の中から与えられたトピックに適合する文書を検索し、スコア順にソートした検索結果をプーリングして人手による適合性判定を行う。

Math-2 のトピックは、高度な検索や信頼性の高い適合性判定を可能にするための詳細説明を含み、以下の形式をとる予定である。トピックの数は 50 を予定している。また、すべてのトピックは、文書コレクション中に複数の適合文書を持つものとする。

表 1: トピックの例.

Topic ID	トピックの識別子
Query	数式とキーワードで表されるクエリ
Description	ユーザの検索要求の記述
Narrative	適合性判定の手がかりとするべき、ユーザの検索状況や検索意図に関する詳細な説明

クエリは数式とキーワードの両方を含み、数式については、MathML および LaTeX 形式で表現されている[1]。NTCIR-10 の Math Pilot Task におけるクエリの例を以下に示す。

表 2: クエリの例.

Math	$\sum_{n=1}^{\infty} \frac{sin(n)}{n}$
Text	infinite series conditionally convergent

[1]NTCIR-10 では、数式中で任意の記号とマッチングできる「ワイルドカード」変数を記述できるよう、タスク固有の拡張を行っている

(4) 評価

Math-2 では、Query フィールドによる検索のランキング結果の提出が必須である。また、オプションとして、人手により生成したクエリに基づく結果の提出なども歓迎する。提出結果には、根拠となる数式 ID などの情報を追加可能である。

オーガナイザによる評価では、参加システムからのランキング結果をプーリングし、上位 100 件について、2 名の判定者が適合性判定を行う。判定は、*Relevant*（適合）、*Partially relevant*（部分適合）、 *Non relevant*（不適合）、*Cannot be assessed*（判定不能）の 4 通りで行う。

3 数式検索における研究課題

数式検索には大きく分けて次の 2 つの課題がある。1 つは数式が持つ構造情報の扱いであり、もう 1 つは数式の意味の考慮である。まず前者について、数式は明示的な木構造を持つため、類似数式の検索に適した木構造の索引付けや、クエリが変数を含む場合への対応などの検討が必要となる。また後者について、数式の意味を考慮するために、文書のドメインや数式周辺のテキストなどの検索コンテクストを、いかに獲得して利用するかの検討が必要となる。

(1) 数式の構造情報の扱い

文書中の「数式」は、改行を伴う関係式だけではなく、行の中に埋め込まれた変数記号なども含む。MathML において数式は木構造の形で表現されることから、数式検索における検索対象は、大規模かつ不均一な木構造データ集合となる。また、数式の木構造表現と数式の持つ意味は 1 対 1 対応とはならないので、木構造データを検索する際には、あいまいさを許すマッチングが必要となる。

上記を踏まえると、数式検索システムの設計では、以下の課題を考慮することが必要である。第一は、数千万規模の木構造データを検索するための索引づけ手法の検討である。効率的な検索のために、ルートからのパスや部分木などの部分構造を索引として用いる実装が考えられるが、これらの部分構造に対する重みづけや、ランキングのためのスコア関数、Top-k 検索などの効率的な検索手法を検討する必要がある。

第二は、数式の特性を考慮した木構造データの扱いである。たとえば、子ノード間の順番を考慮するかどうかなどを、検索効率や性能を踏まえて考える必要がある。また、数式中の変数は他の任意の記号と置き換え可能な場合が多い。このような変数の扱いは、現状では検索システムによって様々であり、評価タスクを通して必要性や有効性を確認することが重要である。一方で、

クエリの中では一般に、どの変数が書き換え可能であるかは明示されないので、変数と解釈するべき記号をどのようにして特定するかという検索意図の推測も課題として残る。さらに、より一般的に、数式は意味を変えずに書き換えることが可能であるが（「$(-1) \times x$」と「$-x$」など)、どの範囲の書き換えを許すのかは検索システムの判断に委ねられているため、タスクを通した検証が必要である。

数式検索の第三の課題は、テキストと数式の統合である。クエリや文書には数式以外の通常のテキストも含まれるので、木構造検索の結果と通常テキストの検索結果をどのように統合してスコアを計算するか、数式とテキストの意味的な対応づけをどのように獲得して検索に利用するかなどを検討する必要がある。

(2) 数式の意味の考慮

数式を使えば、ものごとの厳密で形式的な記述が可能であると思われがちであるが、実際には文書中に出現する数式の解釈には、多くのあいまい性が存在する。たとえば「w^2」は「w の 2 乗」と読むのが通常であるが、実は「w^2」が表現しているのは、2つの文字「x」と「2」の間に存在する特定の位置とサイズの関係に過ぎない[2]。また、「$f(x)$」の「f」は関数であるが「$1/f$ ゆらぎ」の「f」は周波数である、「$y = ax^2$」は x の二次方程式であるが「$E = mc^2$」を二次方程式と呼ぶ人は少ないなど、解釈に必要な知識も多様である。

このように、数式それ自体は抽象的な構造の表現に過ぎず、文脈により与えられる解釈なしには、検索において必要となる同一性や類似度を定義することが困難である。具体的な数式解釈の手法としては、(1) 数式のレイアウト構造の意味構造への変換、(2) 関数や変数の型、束縛変数のスコープの決定、(3) 物理的な単位系や概念クラスへの対応づけ、などの意味処理が考えられるが、多様なアプローチが想定され、現時点では検討がはじまったばかりである。

4 おわりに

本稿では NTCIR-11 における Math-2 タスクの概要について紹介した。数式検索は、数式という独特の構造を扱うものであるが、検索システムの中心になるのは汎用的な木構造検索およびテキスト言語解析の技術である。実装上は、汎用の検索エンジンにテキストと数式の両方を索引付して読み込むことで、手軽にベースライン・システムを構築することが可能である。このようなベースライン・システムを出発点として、数学知識の抽出・体系化と利用、大規模 XML 木構造の検索、深い言語解析に基づく意味解析、テキストと非テキスト要素の統合検索、などの幅広い課題を検討することができる。

タスクの詳細設計やオプショナルなサブタスクについては現在検討中であり、参加者からのコメントやフィードバックを歓迎している。また、スケジュール等を含めた詳細については、順次、タスクのメーリングリストやウェブサイト等でアナウンス予定である。参加問合せや最新情報については以下を参考にして頂ければ幸いである。

表 3: NTCIR-11 Math-2 関連情報.

Organizers ML:	ntcadm-math@nii.ac.jp
Community ML:	ntcir-math@nii.ac.jp
Task Webpage:	http://ntcir-math.nii.ac.jp/
Community Site:	http://ntcir.mathweb.org/

謝辞

本研究の一部は、科学研究費補助金基盤研究 (B) 課題番号 24300062 の助成を受けた。

参考文献

[1] A. Aizawa, M. Kohlhase, and I. Ounis: “ NTCIR-10 Math Pilot Task Overview, ” Proceedings of the 10th NTCIR Conference (2013).

[2] J. Carette, et. al. (Eds.): Intelligent Computer Mathematics - MKM, Calculemus, DML, and Systems and Projects 2013, Part of CICM 2013 Proceedings, Lecture Notes in Computer Science, Springer (2013).

[3] W3C: Mathematical Markup Language (MathML) Version 3.0, Word Wide Web Consortium Recommendation 21 October 2010, http://www.w3.org/TR/MathML3/ .

[4] ArXiv.org e-Print archive: http://arxiv.org/ .

[5] H. Stamerjohanns, M. Kohlhase, D. Ginev, C. David and B. Miller: ”Transforming large collections of scientific publications to XML,” Mathematics in Computer Science. 3(3):299-307 (2010) (Also, see https://trac.kwarc.info/arXMLiv/).

[2]「w^i」とあれば、「i」を添え字だと考える人も多いのではないだろうか