様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 ２０ 日現在

機関番号：６２６１５
研究種目：基盤研究(B)（一般）
研究期間：2012～2015
課題番号：２４３０００６２

研究課題名（和文）言語的アプローチによる数学的知識の理解と利用に関する研究

研究課題名（英文）Natural Language Processing Approach to Understand and Utilize Mathematical Formulae

研究代表者
相澤　彰子（Aizawa, Akiko）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：９０２２２４４７

交付決定額（研究期間全体）：（直接経費）　　13,600,000円

研究成果の概要（和文）：本研究課題では、数式を独自の構造を持つ科学技術文書の構成要素として捉え、言語処理アプローチによって数式の意味を扱うための手法の研究開発に取り組んだ。研究期間内では、数学記号を含む数式の自然言語による説明記述の自動抽出、数式間の依存関係の抽出、数式の大規模複雑な木構造に対する高速な類似度計算アルゴリズムの手法を新たに提案して有効性を示した。また、数式検索システム評価型ワークショップの企画・運営を通して、数式検索の評価用テストコレクションを構築して、関連研究者に広く公開した。

研究成果の概要（英文）：A mathematical formula is an important component of scientific documents with a specific semantic structure. In this research, we aim at developing a framework for semantic enrichment of mathematical formulae using natural language processing techniques. We conducted researches on following techniques for mathematical information access and showed the usefulness of the proposed methods: automatic extraction of natural language description of mathematical symbols and formulae, extraction of dependencies between mathematical formulae in a document, and a fast algorithm for similarity search of a large-scale, complicated tree-structures. We also organized a shared task for mathematical formula search and constructed several evaluation datasets which can be shared by researchers in the related field.

研究分野：知能情報学

キーワード：数式検索　数式理解　自然言語処理　数学知識基盤　MathML

１．研究開始当初の背景

数式は、科学・教育のさまざまな場面で使われる数学の記述法であり、多くの科学分野で重要な役割を果たす。しかし数式は、非言語的な表現であることから、計算機によるテキスト処理では、不要な情報として読み飛ばされ、検索エンジンで検索することが困難であった。数式を中心とした情報アクセス技術が実現されれば、論文間で共通して使われている定理や方程式を提示したり、関連が強い数式を検索して他分野の成果を利用したりするといった、幅広い応用が可能になることが期待される。

２．研究の目的

本研究では、数式を画像や記号列の一種ではなく、独自の構造と解釈を持つ文書の構成要素として捉え、数式とその説明文を対応付けて解析することで、数式の意味を扱うための言語処理アプローチを研究する。

数式の電子化と利用に関する従来の研究は、数学電子図書館で数式を扱うためのものが主流であり、情報検索や自然言語理解のタスクとして定式化されたものはなかった。そこで本研究では、数式を含む文書コーパスを整備した上で、数式検索のためのタスクを設計して運営することで、評価用データコレクションを作成する。これに基づき、数式検索で必要となる要素技術を開発するとともに、大規模な数式検索にも対応できるシステムを実装・評価して手法の有効性を評価する。

３．　研究の方法

(a) 数式検索のためのデータセットの設計と構築

平成 24 年度では、情報検索の評価型ワークショップである NTCIR-10 の新たなパイロットタスクとして、数式検索に特化した「NTCIR-10 Math」を立ち上げ、タスクの設計およびデータセット作成を進めた。続く平成 25-26 年度は、前年度で得られた知見に基づき、本格的な数式検索タスクである「NTCIR-11 Math-2」を立ち上げ、海外の２名のオーガナイザと協力して数式検索のための大規模なデータセットの構築に取り組んだ。さらに平成 26-27 年度では、これまでの経験を踏まえ、海外３名の共同オーガナイザと協力して 3 回目となる数式検索タスク「NTCIR-12 MathIR」を企画・運営し、クエリの設計と再利用可能なデータセットの構築に取り組んだ。

これらのタスクでは、MathML で表現された大量の数式を含む検索用の文書集合を用いて、数式検索用のクエリを準備し、参加システムのランキング結果をプーリングして人手で適合性判定を行った。

(b) 数式検索手法の研究開発

数式の意味解釈のために、統計的機械翻訳に基づく数式の意味構造解析手法、および自然言語文の解析に基づく数式の説明記述の自動抽出手法に関する手法の研究を進めた。また、数式検索システムの機能向上のために、数式間の依存関係の自動抽出とそれを利用した説明記述の補完、および数式の大規模複雑な木構造を扱うための高速な類似度計算アルゴリズムの開発に取り組んだ。

４．　研究成果

(a) 数式検索のためのデータセットの設計と構築

NTCIR のもとで開催した３回の数式検索タスクを通して以下のデータセットを作成した[学会発表 3,6,8,14,17,20]。

- 数式に対する自然言語による説明文を人手でアノテーションしたデータセット
- 科学技術論文を用いた評価用テストコレクション(合計 143 個のクエリと人手による適合性判定結果)
- Wikipedia を用いた評価用テストコレクション（合計 70 個のクエリと人手による適合性判定結果）

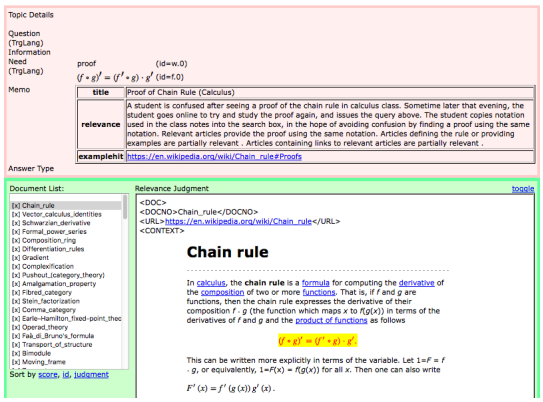

図１：人手判定用のインタフェース（SEPIA）

本研究で構築したデータセットは、現在、数式検索に関する唯一の評価用データセットとなっている。運営したタスクにおいては、独、米、チェコ、オーストリア、中国、インドなどのチームからの参加を得た。また、科学技術論文だけではなく Wikipedia についても文書中の数式を対象とする検索タスクの設計とデータセットの構築を行い、幅広い応用の可能性を示した。本タスクを通して構築したクエリや判定結果は、数式検索システムの開発や評価に有用な研究リソースとして、NTCIR の枠組みのもとで関連研究者に広く公開している。

(b) 数式検索手法の研究開発

まず、言語的アプローチによる数式の意味構造の解析について、統計的機械翻訳に基づく数式の意味構造解析手法を提案し、新たに構築したデータセットを用いて評価して有効性を示した[雑誌論文 2，学会発表 12.15,16,22]。

また、数式説明記述の抽出について、人手アノテーションを用いて機械学習の適用と評価を行い、自動抽出の有効性を検証するとともに[学会発表 18,19,21,23]、NTCIR-Math のデータセットを用いて、数式検索における数式検索における説明記述の有効性を定量的に示した[学会発表 9]。

さらに、数千万個におよぶ大規模な数式木構造の検索を高速に行うために、変数を考慮したハッシュ関数に基づく木構造の高速類似度計算アルゴリズムを提案し、数式検索における有用性を示した[雑誌論文 1，学会発表 7,13]。

加えて、数式間の構造の依存関係を論文から自動抽出して情報を補完する手法の開発に取り組み、アノテーションしたデータを用いて数式依存関係グラフの抽出性能を評価した[学会発表 10,11]。情報検索タスクへの参加にあたっては、数式の部分構造やキーワードなど性質が異なる複数個の索引を最適に組み合わせる手法を適用するとともに[学会発表 5]、変数を含む数式に対応するため、単一化を利用した再ランキングの仕組みを実現した。

これらの手法を実装した数式検索システムを用いて NTCIR-12 MathIR タスクに参

加し、すべてのタスクにおいて本研究で開発した数式検索システムが優れた性能を持つことを示した[学会発表 4]。

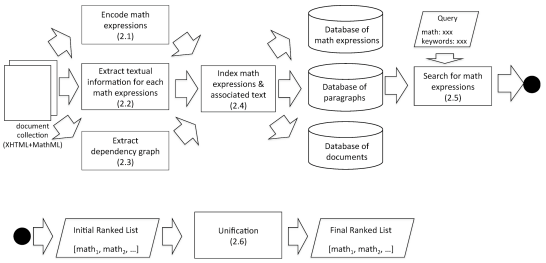

図２：提案数式検索システムの全体図

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

(1) Shunsuke Ohashi, Giovanni Yoko Kristianto, Goran Topić, Akiko Aizawa: “Efficient Algorithm for Math Formula Semantic Search.” IEICE TRANSACTIONS on Information and Systems. Vol.E99-D, No.4, pp.979-988 (2016).査読有
DOI：10.1587/transinf.2015DAP0023

(2) Minh-Quoc Nghiem, Giovanni Yoko Kristianto, Akiko Aizawa: “Using MathML Parallel Markup Corpora for Semantic Enrichment of Mathematical Expressions.” IEICE TRANSACTIONS on Information and Systems, E96-D (8), 1707-1715 (2013). 査読有
DOI：10.1587/transinf.E96.D.1707

〔学会発表〕（計 21 件）

(1) Richard Zanibbi, Akiko Aizawa, Michael Kohlhase, Iadh Ounis, Goran Topić, Kenny Davila: “MathIR Task Overview.” Proceedings of the 12th NTCIR Conference (20160610) 国立情報学研究所（東京）

(2) Giovanni Yoko Kristianto, Goran Topić, Akiko Aizawa: “MCAT Math Retrieval System for NTCIR-12 MathIR Task.” Proceedings of the 12th NTCIR Conference (20160610) 国立情報学研究所（東京）

(3) Giovanni Yoko Kristianto, Goran Topić, Akiko Aizawa: “Combining Effectively Math Expressions and Textual Keywords in Math IR.” In the 3rd International Workshop on “Digitization and E-Inclusion in Mathematics and Science 2016” (DEIMS2016) (20160204) 湘南国際村センター（葉山）

(4) 相澤彰子: “コンピュータによる数式理解と数式検索システム”. 第 120 回情報基礎とアクセス技術研究会･第 47 回ディジタル図書館ワークショップ合同研究会. （20160125）東京工業大学（東京）招待講演

(5) 大橋駿介, 相澤彰子: “SIGURE Hash: 数式検索のための高速な類似検索アルゴリズム”. 第 7 回データ工学と情報マネジメントに関するフォーラム（第 13 回日本データベース学会年次大会）(20150303) 磐梯熱海ホテル華の湯（郡山）

(6) Akiko Aizawa, Michael Kohlhase, Iadh Ounis, Moritz Schubotz: “NTCIR-11 Math-2 Task Overview.” Proceedings of the 11th NTCIR Conference (20141210)国立情報学研究所（東京）

(7) Giovanni Yoko Kristianto, Goran Topic, Florence Ho, Akiko Aizawa: “The MCAT Math Retrieval System for NTCIR-11 Math Track.” Proceedings of the 11th NTCIR Conference (20141210) 国立情報学研究所（東京）

(8) Giovanni Yoko Kristianto, Goran Topic, Akiko Aizawa: “Exploiting Textual Descriptions and Dependency Graph for Searching Mathematical Expressions in Scientific Papers”, The 9th International Conference on Digital Information Management (ICDIM 2014) (20140930) Bangkok, Thailand 査読有

(9) Giovanni Yoko Kristianto, Goran Topić, Akiko Aizawa: “Extracting Textual Descriptions of Mathematical Expressions in Scientific Papers”, The 3rd International Workshop on Mining Scientific Publications, held in conjunction with Digital Libraries 2014 (20140912) London,

United Kingdom 査読有

(10) Minh-Quoc Nghiem, Giovanni Yoko Kristianto, Goran Topić, Akiko Aizawa: “Which one is better: presentation-based or content-based math search?” The Conference on Intelligent Computer Mathematics (CICM 2014) (20140707-11) Coimbra, Portugal 査読有

(11) 大橋駿介, 高須淳宏, 相澤彰子: “表記が異なる同義の数式の高速な検索法”. 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(第 12 回日本データベース学会年次大会) (20140305) 淡路夢舞台＆ウェスティン淡路(淡路)

(12) 相澤 彰子, Michael Kohlhase, Iadh Ounis: “数式検索タスク NTCIR-11 Math-2.” 情報アクセス技術の評価ワークショップ特別セッション, NTCIR-11, インタラクティブ情報アクセスと可視化マイニング(SIG-AM)第 5 回研究会 (20131025) 慶応義塾大学（東京）

(13) Minh-Quoc Nghiem, Giovanni Yoko Kristianto, Goran Topic, Akiko Aizawa: ”Sense disambiguation: from natural language words to mathematical terms” The 6th International Joint Conference on Natural Language Processing (IJCNLP 2013) (20131015) Nagoya, Japan 査読有

(14) Minh-Quoc Nghiem, Giovanni Yoko Kristianto, Goran Topić, Akiko Aizawa: “A Hybrid Approach for Semantic Enrichment of MathML Mathematical Expressions.” The Conference on Intelligent Computer Mathematics (CICM 2013) (20130710) Bath, UK. 査読有

(15) Akiko Aizawa, Michael Kohlhase, Iadh Ounis: “NTCIR-10 Math Pilot Task Overview.” Proceedings of the 10th NTCIR Conference. (20130621) 国立情報学研究所（東京）

(16) Goran Topić, Giovanni Yoko Kristianto, Minh-Quoc Nghiem, Akiko Aizawa: “The MCAT Math Retrieval System for NTCIR-10 Math Track” Proceedings of the 10th NTCIR Conference. (20130621) 国立情報学研究所（東京）

(17) Giovanni Yoko Kristianto, Goran Topić, Minh-Quoc Nghiem, Akiko Aizawa: “Annotating Scientific Papers for Mathematical Formulae Search.” Proceedings of the 5th workshop on Exploiting semantic annotations in information retrieval (ESAIR 2012) of The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), pp.17-18, (20121102) Hawaii, USA 査読有

(18) Akiko Aizawa, Michael Kohlhase, Iadh Ounis: “An Overview of NTCIR-10 Math Pilot Task,” MIR 2012 Workshop — Mathematics Information Retrieval at CICM 2012. (20120708) Bremen, Germany

(19) Giovanni Yoko Kristianto, Minh-Quoc Nghiem, Nobuo Inui, Goran Topić, Akiko Aizawa: “Annotating Mathematical Expression Definitions for Automatic Detection.” MIR 2012 Workshop — Mathematics Information Retrieval at CICM 2012. (20120708) Bremen, Germany

(20) Minh-Quoc Nghiem, Giovanni Yoko Kristianto, Yuichiroh Matsubayashi Akiko Aizawa: “Automatic Approach to Understanding Mathematical Expressions Using MathML Parallel Markup Corpora.” 第 26 回人工知能学会全国大会(JSAI 2012) International Organized Session (20120612) 山口県教育会館ほか（山口）

(21) Giovanni Yoko Kristianto, Minh-Quoc Nghiem, Yuichiroh Matsubayashi, Akiko Aizawa: “Extracting Definitions of Mathematical Expressions in Scientific Papers.” 第 26 回人工知能学会全国大会 (JSAI 2012) International Organized Session (20120614) 山口県教育会館（山口）

〔その他〕
ホームページ等
http://ntcir-math.nii.ac.jp/

６．研究組織

(1) 研究代表者
相澤　彰子（AIZAWA AKIKO）
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：90222447