Letter

定理類似度を計測する情報理論的枠組みの提案

Proposal of Information-theoretical Framework for Measuring Theorem Similarity

中正 和久 [1,*]
Kazuhisa Nakasho[1,*]

1 山口大学大学院創成科学研究科, 山口県宇部市常盤台 2-16-1
1 Graduated School of Science and Technology for Innovation, Yamaguchi University, 2-16-1 Tokiwa-dai, Ube, Japan

* nakasho@yamaguchi-u.ac.jp

Received: August 7, 2020. Accepted: October 9, 2020.

Abstract

With the recent expansion of formalized mathematical libraries, the establishment of theorem search technology in these libraries has become a crucial and urgent issue. Since a theorem changes its expression aggressively by equivalent transformations, the search algorithms based on pattern matching tend to miss their search targets. To overcome this drawback, we introduce a similarity index between theorems and develop search algorithms based on it. In this paper, we propose to formulate "Logical similarity of theorem B to theorem A" as "The quantity of information of the shortest proof that leads $A \to B$." We aim to introduce a quasi-metric function between theorems through this formulation and apply them to automated theorem proving, theorem retrieval, theorem classification and clustering, and theorem library quality check.

1 緒言

1.1 形式化ライブラリの巨大化

近年，定理証明支援系のライブラリの大規模化が進み，検索技術の確立が喫緊の課題となっている．2017 年において，Mizar 数学ライブラリ（MML）は，8,852 のシンボル，12,114 の定義，59,075 の定理を含む約 300 万行のソースコードから構成されている [1]．これほどの大きさのライブラリにおいて，利用したい定理がすでに収録されているか，収

録されているならば何処にあるのか，などの情報全てを一個人が掌握することはほぼ不可能に近い．

1.2 パターンマッチングの限界

現在，定理証明支援系で利用される検索システムの多くは，構文解析と同等のパターンマッチングによって実装されている．しかし，定理や閉論理式は，些細な同値変形によって大きく表現を変えてしまうため，パターンマッチングによる検索とは本質的に相性が悪い．例えば，以下の 3 つの閉論理式は全て同等の意味を持つが，論理式の木構造は大きく異なる．

1. $\forall x, y(A[x] \to B[y])$
2. $\forall x, y(\neg B[y] \to \neg A[y])$
3. $\exists x.A[x] \to \forall y.B[y]$

このため，これら全てを包含するようなパターンマッチを構成することは骨の折れる仕事であり，定理検索システムの利用者は誰もこのような作業を歓迎しないだろう．

1.3 本稿の目的

前述の 2 つの課題を克服するためには，論理式から意味的な特徴量を抽出して，特徴量に基づいた類似度判定を行うタイプの検索システムを構築することが求められる．しかし，現時点では「定理同士が類似している」ということの尺度が定式化されておらず，これが課題の本質的解決を阻んでいると筆者は考える．本稿の目的は，定理間の類似度を定式化するための情報論的フレームワークを提案し，その応用について様々な角度から考察することである．

2 先行研究

2.1 数式検索

数式検索については，2013 年から 2016 年にかけて国立情報学研究所によって主催された国際会議 NTCIR-10, 11, 12 [1] において重点的に研究されている．本会議では，提案されたアルゴリズムの精度やパフォーマンスをベンチマークするために，参加者に共通のデータセットが提供されている．これらのデータは人力によってタグ付けされたものと思われる．今までに提案されている検索アルゴリズムとしては，構文木の部分パス集合（Path indexing）を特徴量とする方法 [2]，シンボルの位置・構文木内での深さを特徴量とする方法 [3]，構文木での 2 つのシンボル間の相対位置を特徴量とする方法 [4] などが挙げられる．多くの数式検索システムは研究用途に実装されたもので一般には利用されていないものの，Springer LaTeXSearch[2]，Wolfram Alpha[3] など商用システムも存在し，数式検索技術は実用レベルに達しているものと考えられる．

[1] http://ntcir-math.nii.ac.jp/
[2] https://link.springer.com/
[3] https://www.wolframalpha.com/

2.2 定理検索

定理証明支援系向けの定理検索の研究については，数式検索に比べると十分に進んでいない状況にあり，2019 年現在においてもベンチマーク用のデータセットが存在していない．これは，そもそも定理の類似度が定式化されておらず，データセットをうまく構築できないからではないかと筆者は推察している．定理検索に利用されているアルゴリズムは，文法構造に着目したパターンマッチングが主であるが，単語の出現回数を特異値分解して特徴量を得る方法 [5] のように機械学習を利用した研究も少数ながら存在する．多くの定理証明支援系は定理検索コマンドを有するが，ライブラリ全体を一挙に検索できるシステムは MML Query [6] の他には存在しない．

2.3 自動定理証明

自動定理証明において，証明中で引用される定理をライブラリから絞り込むプロセスは，自動証明の成功可否に大きな影響を及ぼすことから盛んに研究が進められてきた．自動定理証明は古くから研究されている分野のため，ベンチマークに利用されるデータセットも豊富で，TPTP ライブラリや HolStep などの大規模なライブラリも存在する．自動定理証明に用いられるアルゴリズムには，Term Indexing に利用される Path Indexing や Discrimination Tree などのパターンマッチングによる方法，Premise Selection で利用されるカーネル法, n-gram, k-NN, Random Forest, PageRank, DNN などの機械学習による方法などが挙げられる．また，代表的な自動定理証明器には，E，Otter，SPASS, Vampire どがある．

3 定理検索の課題

定理検索の研究は発展途上にあり，研究に必要不可欠なデータセットが存在していない．では，どのようにデータセットを構築すれば良いだろうか？通常，検索アルゴリズムの性能評価や機械学習に用いられるデータセットは人力で作られることが多い．しかし，定理は些細な同値変形によって構文的な木構造を大きく変化させるため，人間が感覚的に定理同士の類似度を判定すると，少なからず誤りが混入する危険性がある．また，人の脳による認識が絶対的な判断基準となる音声認識や画像認識と同じように，定理のような論理的な対象を人力で分類する必然性があるだろうか？このため，筆者はデータセットを人力で構築することは望ましくないと考える．

以上の議論から,「定理の類似性」を定義することが定理検索の研究の根幹をなす課題であるということには同意いただけると思う．また，定理類似度の定義が万人に受け入れられるためには，定義が論理に深く根ざしたものとなっていなければならないだろう．勿論,「定理」の論理性を担保するものとは「証明」である．

4 定理類似度の定義

4.1 定理類似度が満たすべき性質

定理類似度が満たしてほしい性質について以下にまとめる．

1. 自然さ：万人が納得できる基準であること．理にかなった定義であること．
2. 計算容易性：ベンチマーク用のデータセットを構築できる程度には計算が容易であること．検索などパフォーマンスが要求される状況下では，高速な近似アルゴリズムが利用できること．
3. 有用性：定理検索・定理分類とクラスタリング・定理自動証明等に効果的に適用できること．
4. 距離公理：d を距離関数，A, B, C を定理としたとき，距離公理は以下のように表現される．

   (a) $d(A, B) \geq 0$

   (b) $d(A, B) = 0$ iff $A = B$

   (c) $d(A, B) = d(B, A)$

   (d) $d(A, C) + d(C, B) \geq d(A, B)$

   ただし，非負性 (a) は他の条件から導くことができるため必須ではない．また実際には，対称性 (c) を満たさずとも，$d'(A, B) = (d(A, B) + d(B, A))/2$ と置くことにより d' が距離公理を満たすため，応用上は十分である．距離公理を満たせば，検索やクラスタリング等で使われる様々な既存アルゴリズムを利用することができるため都合が良い．

4.2 定理類似度を定義するためのアイデア

本稿で提案する定理類似度の定義は，定理を点，証明を経路とする空間を考え，証明が持つ情報量を定理間の距離とみなすアイデアに根ざしている．

定理 A, B に対して命題 $A \to B$ とその証明について考えよう（図 1）．定理 B は単独でも証明可能であることから，命題 $A \to B$ も定理となり必ず証明を持つ．このとき，もし命題 $A \to B$ が容易に証明できるのであれば，定理 A と定理 B は距離が近いと考えるのは自然であろう．ただし，命題 $A \to B$ の証明（=経路）は無数に存在すること，命題 $B \to A$ の証明は 命題 $A \to B$ とは非対称であることには注意が必要である．

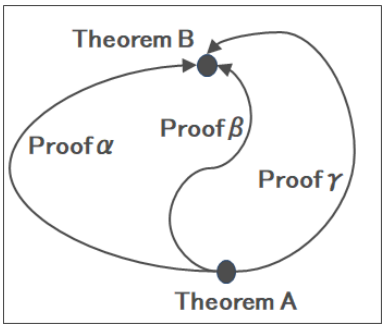

図 1. Theorem Space

次に定理の証明難易度であるが，これは図 2 のような情報源モデルにより計測される情報量と考えてみてはどうか．定理の証明に要する情報量は，情報源に定理を与えたときに，正しい証明を出力する確率から定義される．例えば図 2 において，情報源に定理を与えると確率 p で証明に成功するので，この証明に要する情報量は $-\log p$ となる．このモデルでは情報源にどのようなバイアスを設定するかに応じて証明に要する情報量が定まる．このバイアスは，証明に要する情報量が自然さ，計算容易性，有用性，距離公理などの性質を持つように設定したい．

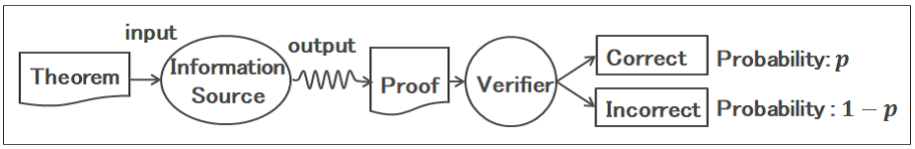

図 2. Information source to evaluate information content of proof

4.3 定理類似度を計測する枠組み

「定理の証明に要する情報量」を「情報源モデルに定理を与えたときに正しい証明を出力する確率」から解釈するアイデアについて考えた．このアイデアは理論的にも合理的に思えるが，計算容易性の面で難がある．一つの定理に対して証明は無数に存在するため，これらの証明に要する情報量を全て合算することは非常に難しい．

そこで，定理の最短証明，つまり情報量が最小の証明を考えてみてはどうだろうか？証明の最短性を示すことは決して容易ではないが，人や自動定理証明器によって導き出された証明がほぼ最短であると仮定してしまえばデータセットを構築できる程度には計算が容易である．また，後述の通り，情報源に与えるバイアスをうまく選択すれば，このようにして定義される定理類似度は準距離公理（距離公理から対称性を除いたもの）を満たすことがわかる．この枠組みは，定理を点，証明を経路とみたときに，定理類似度を最短経路の距離として解釈することに相当する．以下では，このような情報量を「定理の最短証明の情報量」と呼ぶことにする．

4.4 情報源バイアスの候補

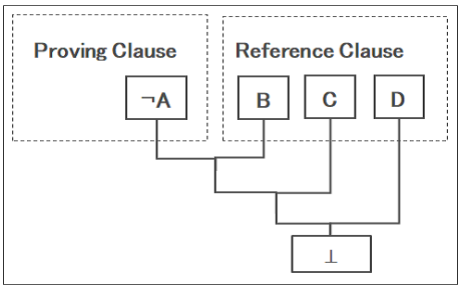

図 3. Proof structure of linear deduction

ここでは，簡単のために一階述語論理体系での線形導出を対象としよう．また，証明において前提とするライブラリ（=定理群）は固定されているものとする．このとき，情報源が証明を生成するときにユニフィケーションするリテラルの選択を間違えないとするバイアスを設定してしまえば，情報源は単にライブラリから証明中で引用する定理（正確には節）を順序良く選択するだけの機械とみなせる（図 3）．実際，ユニフィケーションでリテラルの選択が問題となることは（特定の命題を除けば）多くはないため，このようなバイアスを設定することは十分に合理的であると考えられる．

さらに，情報源が定理を出力する確率にバイアスを仮定することにより，最短証明の情報量は準距離公理を満たす．例えば，ライブラリ中の定理を等確率で発生させるような情報源バイアスを仮定すると，最短証明の情報量は準距離公理を満たす．このバイアスによって導き出される情報量は，証明中で定理を引用する総回数に比例するため，証明の長さとほぼ同義といえよう．また，定理毎に一定の発生確率を仮定するバイアスも準距離公理を満たす．各定理の発生確率は，例えば既存の証明でその定理が引用される確率に応じて定めるとよい．このような情報源は，定理を等確率で発生させる情報源よりも高確率で正しい証明を出力することができるため，より望ましいバイアスであると考えられる．さ

らに，定理の発生確率にマルコフ性を仮定する，証明する定理に応じて定理の発生確率を変動させるなどのバイアスが考えられるが，(現時点では）これらは準距離公理を満たす設定が見つかっていない．

5 定理類似度の応用

5.1 定理の検索

定理検索アルゴリズムの開発には，機械学習や性能評価用のデータセットが必要不可欠である．定理類似度の定義から，既存ライブラリや自動定理証明器による証明を解析することによって，データセットを半自動的に構築することが可能となるであろう．

今回定義した定理類似度を求めるためには，定理の最短証明を発見する必要があるため，そのままの形で検索アルゴリズムに適用するのは負荷が高すぎる．このため，定理類似度を近似する軽量アルゴリズムの開発が必須となる．

5.2 定理の分類とクラスタリング

定理の分類やクラスタリングは，ライブラリの自動分類やドキュメントの自動編纂などに利用することができる．定理類似度が距離公理を満たせば，既存のアルゴリズムが適用できる見込みである．

5.3 自動定理証明

5.3.1 評価関数

現在，機械学習を利用して自動定理証明を強化する研究が Josef Urban 氏，Cezary Kaliszyk 氏らを中心に精力的に進められており，特に Premise Selection（証明中で引用される定理をライブラリから絞り込む処理）において大きな成果を上げている．しかし，Strategy Selection（証明中で定理を引用する順序や組み合わせを決定する処理）については，証明が完了にどの程度近づいたのかを判定するための良い評価関数が見つかっていないため，機械学習における疎な報酬の問題から脱却できていない．しかし，定理類似度が十分精度良く近似できれば，それを評価関数とすることによって証明完了までの距離を推定できるようになる．

5.3.2 サブゴール生成

自動定理証明のアルゴリズムは導出原理の他にもタブロー法や Inverse Method などがあるが，原理的にはすべてゴール（=証明完了）を目指してツリー状に探索を行う．このため，証明が長くなると，探索空間が指数関数的に増大するという問題がある．これに対して人間が証明を行うときには，証明途中で満たされるべき条件を埋め込み，複雑な問題を複数の簡単な問題に分解してゆく分割統治法の原理が随所に用いられる．この操作を一般化すると，命題 $A \to B$ を証明するために $A \to C$ かつ $C \to B$ を満たす C を探す

ことに相当する．(ここで，A，B，C は閉論理式）クレイグの補間定理により，古典的な一階述語論理体系においては，このような補間が必ず存在することが保証されている．

閉論理式 A, B の類似度を $d(A,B)$ とする．定理 $A \to B$ の証明において，$|d(A,C) + d(B,C)|$ と $|d(A,C) - d(B,C)|$ ができるだけ小さく，かつ $A \to C$ および $C \to B$ が双方とも定理となるように閉論理式 C を選ぶことができたとする．このとき，$A \to C$ および $C \to B$ それぞれの証明情報量は（理想的には）$A \to B$ の証明情報量の半分程度になるはずで，$A \to B$ を直接証明するよりも遥かに効率的に証明探索が可能となる．この種の自動定理証明を実現するためには，今回定義した定理類似度を閉論理式の場合に拡張し，さらに閉論理式 C を効率良く生成するアルゴリズムを開発する必要があるため，実用化に向けてのハードルは高い．しかし，もし実現すれば，自動定理証明器は今よりも格段に難しい証明問題を解くことができるようになるであろう．

5.4 定理ライブラリの品質測定

定理類似度ならびに証明情報量を応用することにより，定理ライブラリの指標として以下のようなものを考えることができるようになる．

1. 大きさ：ライブラリ内の全定理の最短証明の情報量の総和と定義する．
2. 網羅性：定理ライブラリ X，Y に対して，X を仮定したときに Y の全定理の最短証明の情報量の総和が小さければ小さいほど X は Y に対する網羅性が高いと定義する．
3. 直交性：定理 A，B に対して，$A \to B$ と B の最短証明の情報量が近ければ近いほど，定理 B は定理 A に対してより直交していると定義する．

一般論として，大きさ・網羅性・直交性が高い定理ライブラリほど品質が高いものとみて良いだろう．

6 結言

本稿では，定理類似度を証明に要する情報量として定義する方法について考察し，これを定理分類，定理検索，自動定理証明などへ応用する構想について論じた．今後の研究計画について以下に述べる．

6.1 理論面での研究

1. 情報源モデルの精密化：証明したい定理に応じて，証明中で引用する定理の発生確率を変化させるようなバイアスで性質が良いものを探す．
2. 閉論理式への拡張：自動定理証明のサブゴール生成で見た通り，類似度は定理だけではなく閉論理式にも拡張されるべきである．
3. 導出原理以外への拡張：本稿では線形導出に限定して定理類似度を提案したが，さらに自然演繹，シーケンス計算，ラムダ計算などへの拡張を行う．

6.2 実践面での研究

1. データセットの構築：まずは Mizar 数学ライブラリ（MML）に対してデータセットの構築を行う．また，自動定理証明器を利用して，2 つの定理を入力すると定理類似度を出力するプログラムを作成し，データセットを自動構築する．

2. 近似アルゴリズムの開発：定理の同値変形に対して頑強な特徴量を土台としてアルゴリズムを設計することにより，高速かつ高精度な近似アルゴリズムを構築できるのではないかと筆者は睨んでいる．

参考文献

[1] Bancerek G, Byliński C, Grabowski A, Korniłowicz A, Matuszewski R, Naumowicz A, et al. The Role of the Mizar Mathematical Library for Interactive Proof Development in Mizar. Journal of Automated Reasoning. 2018 Jun;61(1):9–32. Available from: https://doi.org/10.1007/s10817-017-9440-6.

[2] Yokoi K, Nghiem M, Matsubayashi Y, Aizawa A. Contextual Analysis of Mathematical Expressions for Advanced Mathematical Search. Polibits. 2011;43:81–86. Available from: http://polibits.cidetec.ipn.mx/ojs/index.php/polibits/article/view/43-11/1774.

[3] Asperti A, Guidi F, Coen CS, Tassi E, Zacchiroli S. A Content Based Mathematical Search Engine: Whelp. In: Types for Proofs and Programs, International Workshop, TYPES 2004, Jouy-en-Josas, France, December 15-18, 2004, Revised Selected Papers; 2004. p. 17–32. Available from: https://doi.org/10.1007/11617990_2.

[4] Pattaniyil N, Zanibbi R. Combining TF-IDF Text Retrieval with an Inverted Index over Symbol Pairs in Math Expressions: The Tangent Math Search Engine at NTCIR 2014. In: Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo, Japan, December 9-12, 2014; 2014. Available from: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/Math-2/08-NTCIR11-MATH-PattaniyilN.pdf.

[5] Cairns PA. Informalising Formal Mathematics: Searching the Mizar Library with Latent Semantics. In: Mathematical Knowledge Management, Third International Conference, MKM 2004, Bialowieza, Poland, September 19-21, 2004, Proceedings; 2004. p. 58–72. Available from: https://doi.org/10.1007/978-3-540-27818-4_5.

[6] Bancerek G, Rudnicki P. Information Retrieval in MML. In: Mathematical Knowledge Management, Second International Conference, MKM 2003, Bertinoro, Italy, February 16-18, 2003, Proceedings; 2003. p. 119–132. Available from: https://doi.org/10.1007/3-540-36469-2_10.