3S-2

数式構造と周辺テキストの両面を考慮した数式情報抽出

横井 啓介[1]　Minh-Quoc NGHIEM[2]　松林 優一郎[3]　相澤 彰子[4]

東京大学[1]　総合研究大学院大学[2]　国立情報学研究所[3]

1　はじめに

科学論文において，数式は読み手に主張を明確に伝えるための重要な媒体である．そのため数式に関する情報を読み手にわかりやすく伝える必要があるが，現時点では数式に関する実用的な検索・理解支援サービスは存在しない．

本稿では，(1) 数式構造を用いた類似数式検索手法と (2) 数式周辺のテキストを利用した数式中の変数や関数，および数式全体に対する説明記述の抽出手法について説明し，その後それらを統合した (3) 科学論文の読解支援を行うサービスである“MathDA”について説明する．

2　数式構造を考慮した類似数式検索

科学論文中の数式に対し，用途や形が類似した数式や，その数式が出現した論文中での文脈が理解に役立つことは多い．そこで我々は数式の持つ独自の構造を利用した数式間類似度を提案した[1]．ここでは数式が Web 上の標準的な記法である MathML (Content Markup 方式) で表現されていることを想定する．図 1 に例を示す．

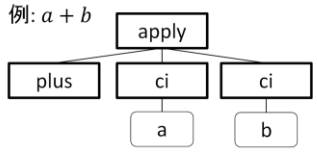

図 1 MathML(Content Markup) の例

我々はこのように木構造として表現された数式の類似度を測るために Subpath Set を構築する．Subpath Set とは文献[2]においてテキスト構文構造の類似度を測定するために提案された尺度であり，我々はそれを数式構造に応用した．具体的には，根から葉までの全部分経路の集合として Subpath Set を定義し(図 2)，その集合間の類似度をスコアとして類似する数式を抽出する．

我々は Content Markup で表現された数式に対し，Subpath Set が構造的な類似度をより深く考慮できるように変換を施した上で，Jaccard 係数を用いて類似度を算出した．

提案手法に対し，類似数式検索の評価・比較

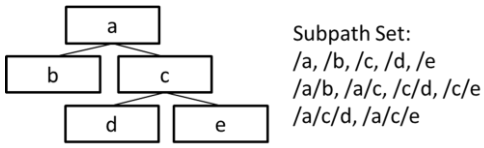

図 2 Subpath Set の例

を行った．実験データとして 1 論文あたり 3~5 式，合計 29 の数式を選出し，判定者に対し対象の数式・システムが出力した類似数式・それぞれの式の出典論文を共に提示した．ユーザは，論文を読む際に類似数式が対象の数式と「参考になる(3)」「部分的に参考になる(2)」「参考にならない(1)」の 3 段階で判定する．比較対象として橋本らの XPath を用いた数式検索手法[3]を用いた．その結果を以下に示す(表 1).

表 1 類似数式検索の評価結果

		評価 1	評価 2
提案手法	1	0.483	0.828
	2	0.397	0.672
	3	0.356	0.609
	4	0.310	0.526
	5	0.262	0.469
比較手法		0.250	0.577

評価 1 は判定(3)を得た数式の割合，評価 2 は判定(2,3)を得た割合を精度とした．提案手法は類似度の上位n件までを判定した結果を$n = 1 \sim 5$について示した．比較手法はランキングを行っておらず，全く結果を返さない数式も存在する．今回用いた実験セットでは 1 数式あたり 1.79 式の出力があったことを考えると，我々の提案手法は適合性，特に判定(3)の数式の割合で有意であると言える．また，結果から我々の手法は上位の数式ほど精度が高く，類似度の判定が有効であると言える．

3　数式の周辺テキストからの情報抽出

数式は抽象化された表現であり，その意味や使い方などの様々な説明記述があって初めて正確な理解が可能になる．そこで我々は，数式の周辺テキストを解析し，機械学習を用いて数式中の変数や関数に関する説明記述を抽出する手法を提案した[3]．はじめに学習のため，情報処理学会論文誌より 100 の論文を選択し，データセットを手作業で作成した．例を以下に示す(表 2)．簡単のために数式の説明記述はすべて名詞，および複合名詞とし，またその対応する数式と

Mathematical Information Retrieval Using Mathematical Expressions and Surrounding Texts

[1] University of Tokyo

[2] The Graduate University for Advanced Studies

[3] National Institute of Informatics

同一文中に存在するという制約を設け，それを踏まえてデータセットを作成している．

表 2 データセット例

ID	単語	タグ	
0	ただし	O	O
1	Exp	Pred	O
2	は	O	O
3	語	O	B
4	Exp	O	Pred
5	の	O	O
6	出現	B	O
7	頻度	I	O
8	と	O	O
9	する	O	O

データセット中では数式は Exp と変換し，数式の数に応じて，それぞれの単語が数式の説明記述であるかどうかを示す BIO タグを振り分けている．

以上のデータセットを用いて機械学習を行う．素性には，周辺の単語の品詞や名称などの MeCab による形態素情報の他に，Cabocha により得られる係り受け関係などを用いた．そしてそれぞれの単語がそれぞれの数式の説明記述であるかどうかをサポートベクターマシンの二値分類モデルにより判定を行った．その結果を以下に示す(表 3)．本研究に関しては比較対象研究が見つからないため，今回はベースラインとして「名詞と数式が連続して現れる場合にその名詞を数式の説明記述とする」手法を用いた．論文中の数式の説明記述は多くがこのようなパターンに従うため，シンプルではあるが有効な手法と言える．

表 3 情報抽出の評価結果

手法	Precision	Recall	F1
提案手法	0.8732	0.8139	0.8425
ベース	0.8990	0.5817	0.7064

実験において，今回は連続した名詞は全て複合名詞として扱っており，その区切りによって機械学習の精度に 6%程度の影響があることを考えると，本手法は高い精度で抽出できていると言うことができる．

4 科学論文読解支援サービス“MathDA”

これまで述べてきた類似数式検索，意味情報抽出の結果を利用して，我々は科学論文を閲覧する際の理解補助を行うシステム“MathDA (Mathematical Document Analysis)”を制作した．本システムの外観を以下に示す(図 3)．

本システムはブラウザ上で論文を読むことができるシステムであり，大きく 2 つのページ構成からなる．左部が論文ページ，右部が数式ページである．論文ページは論文を閲覧するためのページであり，数式に照準を合わせると，その数式およびその数式に含まれる変数や関数などの説明記述がポップアップされる．また，数式をクリックするとその数式に対応した数式ページが表示される．数式ページには先にも述べた説明記述の他に，論文中のその数式に関する記述，また論文中，他論文，Wolfram Functions Site より集めた公式集から類似数式を検索して提示している．本システムにおいて実際に利用者実験を行い，良い評価を得るとともに，ユーザビリティに関する課題も明らかになった．

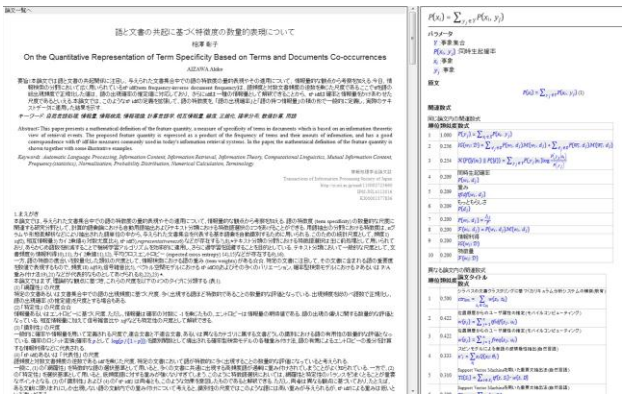

図 3 MathDA 画面

5 まとめ

我々はコンピュータ上で科学論文を読む際に，数式に関する情報を読み手にわかりやすく掲示するために，類似数式検索・周辺テキストからの情報抽出の 2 つの手法，および論文読解支援サービスを提案し，有効性を確認した．

参考文献

[1] Keisuke Yokoi and Akiko Aizawa: An Approach to Similarity Search for Mathematical Expressions using MathML, DML 2009, pp. 27-35, 2009.

[2] 市川宙，橋本泰一，徳永健伸，田中穂積: テキスト構文構造類似度を用いた類似文検索手法，情報処理学会研究報告，2005-DBS-136, pp.39-46, 2005.

[3] 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

[4] Keisuke Yokoi, Minh-Quoc Nghiem, Yuichiroh Matsubayashi and Akiko Aizawa: Contextual Analysis of Mathematical Expressions for Advanced Mathematical Search, CICLing 2011, 2011 (accepted).