D-021

類似したタイトルを持つ論文同士の内容的差異に対する特徴分析

Analyzing Characteristics of Content Differences between Papers with Similar Titles

玄道俊 † 松下 光範 ††
Shun Gendo † Mitsunori Matsushita ††

1 はじめに

研究を遂行するうえで，自らの研究の位置づけを明確にするには，関連する論文を体系的かつ効率的に収集・把握し，各論文の主張点を適確に掴むことが重要である．Web 上に多くの論文が公開されるようになってきている現在，このような目的のために論文を収集する際には検索エンジン (e.g.,Google Scholar，CiNii) を用いることが一般的になっている [2]．検索エンジンに，自身の研究に関係するクエリを入力することで，論文のタイトルが検索結果として表示され，その中から自身に適した論文を選択する．研究は過去の成果を土台として拡張されていくものであるため，同じような目的や手法のもとで行われる研究の場合，タイトルが類似している論文が散見される．

大量の論文の中から論文読者が論文を収集する際，タイトルから自身の研究に適した論文のみを収集することは困難であるため，内容を理解した上で選択することが必要になる．しかし，類似したタイトルの論文が多く存在する場合にそれらすべてを読んで選択するには多大な労力と時間がかかるため，効率的に論文を選択することができれば，研究遂行の効率化を図ることが期待できる．

こうした背景の下，本研究では類似したタイトルの論文の差異を可視化することでユーザの論文選択を支援するシステムの実現を目指す．その端緒として本稿では，類似したタイトルの論文を対象に内容の差異に対する計算機の判断と人間の判断を比較し，その一致度を調査する．

2 関連研究

程岡らは引用論文に着目し，研究動向の把握を支援するインタフェースの開発を行った [6]．提案インタフェースではノードとエッジを用いることで，引用論文の一致度を直感的に表すことを可能とした．また，起点となる論文以外の論文をクリックすると色が変化し，引用論文の差異を把握できる．提案インタフェースの有用性を確認するために課題を作成し，コミック工学分野の論文の執筆経験があるユーザを対象に実験を行った．提案インタフェースの使用後インタビューを行い，起点論文と比較論文が引用している論文の一致度や，差異の側面で有用性が確認された．

また，南浦らの研究では論文集合のクラスタリングをすることにより類似論文を抽出し，その結果から関連用語の提示を行う論文検索が可能なインタフェースを提案している [7]．提案インタフェースではユーザがキーワードを入力することで関連用語が提示され，CiNii での結果と類似論文の結果も表示される．関連用語の算出方法として，論文テキストデータからキーワードを抽出し，タームベクトルを作成し，クラスタリングを行っている．クラスタリングの粒度を関連度と定義し，提案システムでは，同じクラスタに所属する論文を類似論文と定義し，推薦に用いている．

3 提案手法

本稿では，論文のタイトルの中に同一単語が 2 つ以上，かつ同一著者が 1 人以上含まれている 2 つの論文を類似論文と定義する．同じ論文構成要素の章には同一主旨の内容が記載されているため，2 つの論文を章ごとに比較することでその差異の明瞭化が期待される．また，類似論文を特徴づける単語の違いは論文間の内容の差異を表していると考えられるため，それらの単語に着目し，内容的差異を判断する．本稿では，これらの単語を識別特徴語と定義する．

3.1 論文の収集

本稿ではコミック工学分野の論文を用いる．コミック工学分野の研究領域が広いため，タイトルから論文の差異が把握しづらいと考え，本稿では 2003 年から 2018 年に上梓されたコミック工学の論文 40 本をデータとして用いることとする．

また，類似論文を章で比較するため本稿で使用する論文は同一章で構成される類似論文を使用する．枚数の少ない論文は，本文量が少ないため，記載する情報も少ない．よって比較する類似論文は 4 枚以上の論文を用いた．

3.2 章の統一

同一章の論文を比較するため，章の統一を行う．論文執筆者は自身の研究に適した独自の章題をつける場合がある．例えば，「従来の手法と問題点」と「関連研究」はどちらの章も過去の研究を述べた上で自身の

関西大学大学院総合情報学研究科, Graduate School of Informatics, Kansai University (†)
関西大学総合情報学部, Faculty of Informatics, Kansai University (††)
〒 569-1095 大阪府高槻市霊仙寺町 2 丁目 1-1

研究の立ち位置を明確化するための内容が記載されている．章題の異なるものでも比較を可能にするため，章の統一を行う必要がある．統一の定義は科学論文で主に使用される構造の IMRAD 形式を参考にした [4]．IMRAD 形式は「序論」(Introduction),「方法」(Methods),「結果」(Results),「討論 」(Discussion) が IMRAD 形式である．Ling ら [3] は 39 分野 433 論文を調査した結果，最も頻繁に使用されていた形式である「序論」「方法」「結果」「討論」の他に「文献レビュー」(Literature review) と「結論」(Conclusion) が存在していた．また，石田らの研究 [8] では「方法」と「実験」に同じ「方法」のラベルが付与されていたが，本稿で扱うデータセットであるコミック工学分野の論文は提案手法と実験が別の章で執筆されていることを考慮し，「実験」(Experiment) と分離させた．また，本稿で対象とする分野の論文では「討論」と「結果」が「実験」としてまとめて書かれている場合が多いため，これらの 2 つの章をまとめて「実験」として扱う．したがって，「序論」「方法」「文献レビュー」「結論」「実験」の 5 種類のラベルで章を統一した（表 1）．例えば，「はじめに」，「関連研究」，「提案手法」，「実験」，「おわりに」で構成される章の場合は「序論，文献レビュー，方法，実験，結論」で統一した．また，「実験と考察」の章の場合は「実験」のラベルを付与した．ただし，論文に 5 つ全ての構成要素が含まれない場合は，存在する構成要素のみを扱う．

3.3 識別特徴語の算出

識別特徴語を算出するために，論文の本文のデータセットから単語ごとに分ける必要がある．その前処理として，単語を抽出するため形態素解析器 MeCab 0.996[†]を用いて論文のテキストデータを品詞ごとに分解する．論文中で用いられる手法は名詞が多いため，本研究では名詞のみを取り出す．論文の内容に関係ない語が含まれることを防ぐため，SlothLib[9] にストップワードとして設定されている名詞群をストップワードとする．

特徴語の算出において，TF-IDF 法を用いて特徴語の算出を行う [10]．TF-IDF 値は単語の出現頻度を表す TF 値と逆文書頻度数を表す IDF 値の積から算出する．TF-IDF 値の算出方法は式 (1) の通りである．

$$TFIDF_{i,j} = \frac{n_{i,j}}{X} \cdot \log_e \frac{N}{df_i} + 1 \qquad (1)$$

$n_{i,j}$ は文章 j における単語 i を示しており，X は文章 j に出現する単語の総頻出度である．また N は総文章を示しており，df_i は単語 i を含む文書数である．

† https://taku910.github.io/mecab/

表 1: 章の統一ラベル

ラベル	例
「序論」	はじめに，序論
「方法」	提案手法
「実験」	実験，考察，討論，結果
「文献レビュー」	関連研究
「結論」	おわりに

TF-IDF 値を算出し，類似論文を比較した．TF-IDF 値が高い単語を比較した結果，コミック工学分野で頻繁に使用される単語が上位に算出された．つまり，各論文の特徴語ではなく，コミック工学分野における特徴語が算出されている．そこで上位に算出された 20 単語をストップワードとして追加した．

類似論文の差異を算出するために TF-IDF 値の差を算出する．TF-IDF 値の差の値が負になることを考慮し，絶対差を算出した．算出した絶対差の数値を降順に並べ，上位 10 単語を類似論文の比較する語は論文の差異を判断する為の語であると考え識別特徴語として使用する．識別特徴語算出の流れを図 1 に示す．

4 実験

本章では，識別特徴語の分布と人が差異を判断する際に着目する章の関係性について調査するため実験を行った．

4.1 実験内容

実験参加者は論文を読んだ経験のある情報学部の大学生男女 18 名とした．実験では当該分野の論文 40 本から，3 ペアの類似論文を選び，1 ペアにつき 6 人をランダムに割りあてた．実験参加者には印刷した紙媒体の類似論文を読んでもらった上で，「2 つの類似論文の違いを報告するために各論文を 500 字で要約してもらいます．要約に使う文をマーカーで線を引いてください」といった内容の課題に取り組んでもらった．マーカーで線を引いた後，その理由を自由記述で記載してもらった．

4.2 実験結果の前処理

実験参加者がマーカーで線を引いた文を，人手でテキストデータにした．前処理として，3.3 節で述べた SlothLib に設定されているストップワードとコミック工学論文の頻出語上位 20 件をストップワードとした．テキストデータから，名詞のみを抽出した単語（以下：選択語）を使用し分析する．これらの選択語中に識別特徴語がどの程度存在するか算出する．また，選択語がどこの章に頻出しているかを測る．

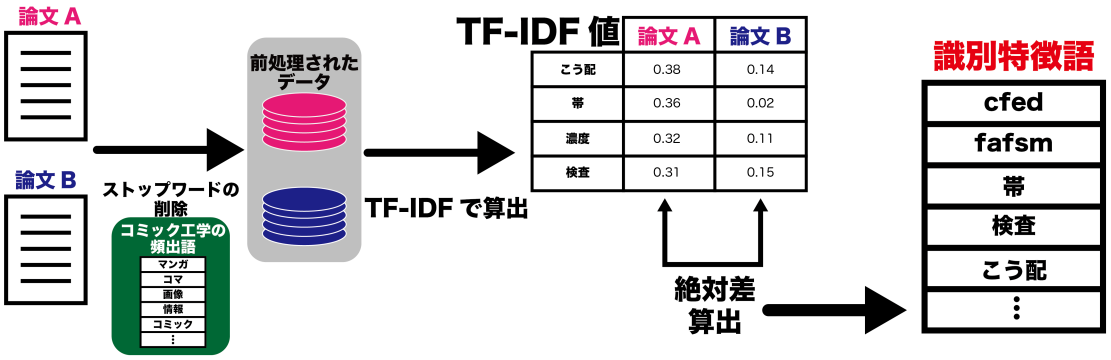

図 1: 識別特徴語算出までの流れ

4.3 実験結果

人が論文の差異を判断する際に識別特徴語に着目していたか確認するため，各論文の識別特徴語の割合における平均値と，選択語中に含まれる識別特徴語の全体における平均値を比較した．その結果，対象とした類似論文において，論文中に含まれる識別特徴語の割合が平均 4.7% (最大: 6.7%，最小: 3.0%) に対し，マーカーが引かれた文中に含まれる識別特徴語の割合が 平均 16.7% (最大: 24.7%，最小: 9.2%) であった（表 2 参照）．全ての論文において，マーカーで引かれた文中に含まれる識別特徴語の割合の方が高かった．

また，全ての類似論文で識別特徴語の異なる語数の平均値は 5.0 単語 (最大: 6.0 単語，最小: 3.3 単語) であった（表 3 参照）．このことから，人が論文の差異を判断する際に，識別特徴語は手がかりとなることが示唆された．

5 考察

本章では，実験結果をもとに得られた結果から平均を求めそれを踏まえて考察を述べる．

識別特徴語が出現する章の割合と人が着目していた章の割合が同一であるか確認するため，各章ごとの割合の差を算出した．提案手法により識別特徴語が分布している章と選択語の章の分布の結果を図 2 に示す．人と提案手法を比べた結果，3 つの類似論文全てにおいて人の方が提案手法より「序論」と「結論」の割合が高く算出された．そのため，人はこれらの章に重きをおいて論文間の差異を判断していると考えられる．

次に，度数分布表の類似性を表す事が可能である Bhattacharyya 距離を算出する [1][5]．計算式はコンピュータビジョン向けライブラリの Open CV2.2 にある calcHist 関数を使用した††（式 2 参照）．提案手法で算出した識別特徴語の章における分布と，選択語の章における分布から Bhattacharyya 距離を算出

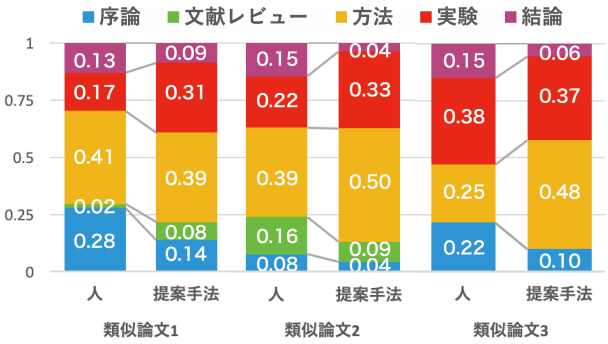

図 2: 各章における識別特徴語の割合と選択語の割合

した．

$$d(H_1, H_2) = \sqrt{1 - \frac{1}{\sqrt{\overline{H}_1 \overline{H}_2 N^2}} \sum_I \sqrt{H_1(I) \cdot H_2(I)}} \tag{2}$$

N は区間の数を表し，H_1, H_2 は各区間の値を，$\overline{H}_1$, $\overline{H}_2$ は平均値を表している．単一の Bhattacharyya 距離の値では相対的な評価が行えないため，章における識別特徴語の分布が全て同一だった場合（以下，チャンスレベルと記す）の分布と選択語の分布の Bhattacharyya 距離を算出し，算出した提案手法の Bhattacharyya 距離と比較した（表 5 参照）．類似論文 2，類似論文 3 ではチャンスレベルの Bhattacharyya 距離の値が低くなっている．これは，チャンスレベルの方が，人が着目していた章の分布に近いことを示している．このことから，章において識別特徴語の分布と人が着目する章の分布は必ずしも一致しておらず，論文間の差異を識別特徴語の分布から判断するには不十分であることが示唆された．その理由として，人と提案手法を比べた結果，3 つの類似論文全てにおいて提案手法より人の方が「序論」と「結論」の割合が高いことが挙げられた．そのため，人はこれらの章に重きをおいて論文間の差異を判断していると考えられる．

†† http://opencv.jp/opencv-2svn/cpp/histograms.html

表 2: 論文中に含まれる識別特徴語の割合と選択語中に含まれる識別特徴語の割合

類似論文	論文	論文中に含まれる識別特徴語の割合	選択語中に含まれる識別特徴語の割合
1	A	3.6%	17.1%
	B	3.9%	12.0%
2	C	3.0%	10.5%
	D	4.9%	14.3%
3	E	6.7%	25.5%
	F	5.9%	21.8%
	平均	4.7%	16.7%

表 3: 類似論文における識別特徴語の異なり語数

類似論文	論文	異なり語数
1	A	3.3 語
	B	5.3 語
2	C	4.0 語
	D	5.7 語
3	E	6.0 語
	F	5.7 語
	平均	5.0 語

表 4: Bhattacharyya 距離の比較

	提案手法	チャンスレベル
類似論文 1	0.193	0.280
類似論文 2	0.185	0.183
類似論文 3	0.199	0.117

6 おわりに

本稿では，類似したタイトルの論文の内容における本文を構成する要素である章と単語に着目し，差異に対する計算機の判断と人間の判断を比較し，その一致度を調査した．差異の算出方法として TF-IDF 法を用い，その値における差の絶対値を出し，値が高い上位 10 件を識別特徴語と定義した．人が論文の差異を判断する際に識別特徴語に着目していたか，また識別特徴語が出現する章の割合と人が着目していた章の割合が同一であるか確認するため，実験を行った．実験の結果，識別特徴語は人が論文の差異を判断する一助となることが示唆された．また，人は識別特徴語の分布と人が着目していた章の割合は必ずしも一致していなかった．この理由として，人は「序論」と「結論」に重きをおいて論文間の差異を判断していると示唆された．今後は「序論」と「結論」の中から識別特徴語の他に判断する単語に着目し，比較方法について検討する．

参考文献

[1] Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions, *Bulletin of the Calcutta Mathematical Society*, Vol. 35, pp. 99 – 109 (1943).

[2] Nakano, Y., Shimizu, T. and Yoshikawa, M.: A Visualization of Relationships Among Papers Using Citation and Co-citation Information, *International Conference on Asia-Pacific Digital Libraries* , pp. 1-6 (2016).

[3] Lin, L. and Evans, S.: Structural patterns in empirical research articles: A cross-disciplinary study, *English for Specific Purposes* , Vol. 31, No. 3, pp. 150 – 160 (2012).

[4] Gastel, B. and A., R.: 世界に通じる科学英語論文の書き方 : 執筆・投稿・査読・発表, 丸善 (2010).

[5] 川端隼矢, 長名優子: タッチの類似性を考慮したイ ラスト検索の精度向上 ー特徴量と類似度の算出方 法の変更 ー, 第 79 回全国大会講演論文集, No. 1, pp. 31 – 32 (2017).

[6] 程岡晃一, 大杉隆文, 松下光範: 引用論文に着目した研究動向把握支援インタフェースの基礎検討, 第 21 回インタラクティブ 情報アクセスと可視化マイニング研究会, Vol. 17, pp. 102-107 (2019).

[7] 南浦 佑介, 新美礼彦: 類似論文からの関連用語抽出による論文検索支援システムの提案, 言語処理学会　第 17 回年次大会　発表論文集, pp. 69-72 (2011).

[8] 石田 栄美, 安形 輝, 宮田 洋輔, 池内 淳, 上田 修一: 構造と構成要素に基づく学術論文の自動判定, 構造と構成要素に基づく学術論文の自動判定 Vol.60, No.1, pp. 18-34 (2014).

[9] 大島 裕明, 中村聡史, 田中 克己: SlothLib- Web サーチ研究のためのプログラミングライブラリ, 日本データベース学会, Vol.6, pp. 113-116. (2007).

[10] 鈴木 啓, 大内 紀知: テキストマイニングによる学会の特徴分析, 経営情報学会 2016 年秋季全国研究発表大会, pp. 79-82 (2016).