Vol.2017-IS-142 No.2
2017/12/2

人工知能技術を用いた情報処理学会コンテンツのメタ解析システム

居駒幹夫[†1 †2]　大場みち子[†1]

概要：情報処理学会には，10 万本を超える論文や報告が蓄積されており，学会全体はもちろん，ある研究分野に限っても個人が全体を俯瞰したりトレンドを把握したりすることが困難になりつつある．本研究は，文書をベクトル化する人工知能技術 Doc2Vec を活用し，学会で電子的に蓄積されているコンテンツの概要や書誌情報から個々のコンテンツだけでなく，研究会，発行年等の複数コンテンツをベクトル化した．これらの情報をもとに学会全体や特定研究分野のトレンドを解析し，各年で発行される学会論文全体のベクトルが一定の方向で推移している等の知見を得た．また，これらの知見を新たに生み出すことが可能な情報システムを開発し一般研究者が学会情報をベクトル解析できるようにした．

キーワード：リサーチマイニング，自然言語解析, Doc2Vec，電子図書館

A Meta Analytical System of IPSJ Contents Using Artificial Intelligence Technologies

MIKIO IKOMA[†1 †2]　MICHIKO OBA[†1]

Keywords: Research mining, Natural language processing, Doc2Vec, Digital library

1. はじめに

「巨人の肩の上に立つ」の言葉通り，科学的な学術論文の価値は，それまでに積み重ねられた多くの論文の蓄積が前提になっている．一方，蓄積された学術論文の数は膨大なうえ，常に新しい論文が積み重ねられる状況にある．情報処理学会の場合，2017 年 10 月現在で 13 万本を超える論文や報告が蓄積されており，2016 年一年でも 5495 本が新たに登録された．個人が過去から最新の状況まで追随することは事実上困難となっているのである．

この問題の解決のために，特定分野の過去論文を調査，整理したレビュー論文，サーベイ論文と呼ばれる論文が重要である．しかし，レビュー論文，サーベイ論文の分野を選択する時点で主観的な判断が働かざるを得ず，学会レベルのマクロな動向を踏まえているのかの確認は事実上困難である．さらに，分野が特定されたとしても，その分野の論文数も多数にのぼりレビュー論文を執筆すること自体が大きな労力を要する状況となっている．結果として特定の問題領域の情報が整理されたレビュー，サーベイ論文が，極端に不足しているのが現状である[1].

本研究は，昨今各分野で実用化が目覚ましい人工知能技術 Doc2Vec[2]を活用し，個人では把握が困難になっている情報処理学会レベルでのマクロな動向をメタ分析した．この結果，学会全体で，各年で発行される論文全体のベクトルが一定の方向で推移していることや，ここ 10 年で各研究会のコンテンツが類似のベクトルを持つようになっていることが分かった．さらに，これらの結果を基に，簡易に検索や解析が可能な情報システム IPSJ2Vec を開発した．IPSJ2Vec を使うことにより，Web ブラウザ経由で学会や研究会レベルのトレンド解析をしたり，レビュー論文を企画したり，一般学会員が論文を投稿する際に適切な分野の特定したりすることが可能になった．

本研究での以降の構成は次の通りである．2 章では本研究が対象とする分野の概略，従来研究とその課題，着目点を明確にする．3 章では本研究のアプローチと使用する人工知能ライブラリ Doc2Vec を紹介する．4 章で，Doc2Vec を用いた学会コンテンツの解析事例および，情報システムを報告する．本研究の課題を 5 章で述べ，6 章でまとめおよび将来的な方向性を述べる．

2. 背景

2.1 対象分野

本研究は，研究者が専門分野および関連分野のレビュー論文，サーベイ論文等を企画したり，新しい研究を開始したり，論文を執筆投稿したりするときに，それらの分野の過去の研究の内容，トレンドの把握を容易にすることを対象とする．

ほぼすべての研究者は，専門分野の学会，研究会等に参加していると思われる．研究会や，小規模な学会，すなわち参加者が比較的少数の場合，その会の研究スコープやあ

†1 公立はこだて未来大学 FUTURE UNIVERSITY HAKODATE
†2 日立製作所　Hitachi, Ltd.

る時点でのトレンドは自明であろう．しかし，自分の参加していない関連研究会の内容や，大規模学会の全体としてのトレンドは必ずしも自明ではない. 情報処理学会の場合，学会誌，論文誌，全国大会という機会を用いて研究会間の交流や情報共有を目指している．しかし，客観的なテーマ設定がされている保証はなく，また，機会があっても自分の関連する研究分野をウォッチすることも容易でなくなっている．

研究分野を限定できる場合，過去の主要な論文を構造的にまとめ，その研究分野のトレンドや未解決課題などを明確にするレビュー論文，サーベイ論文が重要である．しかし，これらの論文を書くこと自体が多くの労力を要し，少なくても情報処理分野では例えば医療分野に比べるとレビュー論文，サーベイ論文はほとんど書かれていないのが現状である．

本研究では，情報処理学会を例にとり，10 年 20 年単位のマクロ解析により, 学会全体のメタな動向を知る. また，この動向を踏まえて，実際に研究者の実務で役に立つ情報システムを実現することを本研究の具体的な課題とする．最終的な目標としては，レビュー論文やサーベイ論文の負荷低減や，そもそも，レビュー論文が意図している課題をレビュー論文無しで解決することを目指す.

2.2 従来研究およびその課題

多くの学会は，その成果を書籍として発行してきたが，昨今, 電子的な形式により Web サイトで公開するようになってきた．情報処理学会の場合，電子図書館(情報学広場)1を設置し，学会員のみでなく，非学会員でも全文検索，著者，発行年等のキーワード検索等を可能にしている．現状の検索は，検索用語の意味にまでは踏み込んでおらず，同義語や類似概念の語での検索漏れが多く発生する．また，学会全体のトレンドや研究会の関連等も検索の対象外である．

レビュー論文，サーベイ論文の題材となる主要論文を機械的に把握する方法として，論文の参照関係を用いた方法が多く提案されている[1][3]. 一般の論文検索として全世界的にもっとも活用されているのは Google 社が提供しているサービス, Google Scholar2である．このサービスの検索アルゴリズムは公開されていないが，解析した論文[4]によると，通常の Google 検索のページランキングアルゴリズム[5]と同様に論文の参照関係を用いている．また，Google Scholar の場合, 入力した語の類似語による検索もサポートしている．

これら従来の検索方法は, 学会をまたがった主要な論文，参考にしたい論文の検出には有効であるが，ある学会のトレンドや，例えば，自分の思いついた論文をどの学会，研究会に投稿しなければならないかといった課題の解にはならない.

3. 本研究の方法

本研究では，単語の意味や文章の特徴を固定長ベクトルに変換可能な技術 Word2Vec[6]および Doc2Vec を用い，情報処理学会のコンテンツを解析し, 個々の論文のみならず，各研究会，発行年，著者や，これらを含めた学会全体の特徴をベクトル化し，それらの情報を用いてトレンドを解析する．本章では，基礎技術である Word2Vec, Doc2Vec の紹介し，これら技術の過去の活用事例と本研究での活用方法を説明する．

3.1 Word2Vec, Doc2Vec とは

自然言語処理の分野で Word2Vec は時代を画する技術である．Word2Vec は単語ごとに分解された文書を入力に，ある単語の特徴を他の単語との位置関連から固定長ベクトルとして求め，他の単語との類似度や関連を算出可能にする技術である．Word2Vec は英語，日本語といった自然言語の種類に依存せず，プログラム言語でも解析可能である．

Word2Vec により, 単語間の意味の演算が出来るようになる．例えば，英語において，King の女性形は以下のベクトル演算で Queen だと求められるようになる(図 1-(a)).

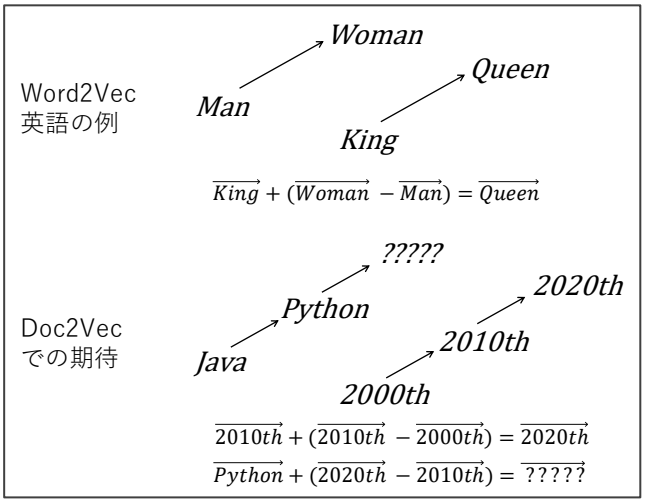

図 1　Word2Vec, Doc2Vec の意味演算
Figure 1　Semantic Relationship using Word2Vec and Doc2Vec.

Doc2Vec は，Word2Vec で算出した単語のベクトルから，可変個の単語からなる文章の固定長ベクトルを算出する技術で，Word2Vec 同様，文書間の類似度を算出できる．Doc2Vec の大きな特徴のひとつは，単語のベクトルと文章のベクトルが同じ空間に存在し，その相互での重ね合わせや類似度の算出が可能ということである. さらに, Doc2Vec の実装である Python のライブラリ Gensim[7]では，各文章

1 https://ipsj.ixsq.nii.ac.jp/

2 https://scholar.google.com

にタグを付与することが可能で，あるタグを付けた複数の文書に対応したベクトルも，単語や，１文書のベクトルと演算可能である．このため，類似度のみでなく，文書のトピック等を抽出したり，文書間の意味演算をしたりすることも期待されている．

情報処理学会の論文の例で言うと，ある（複数単語で構成される）論文と同じような特徴を持つ単語，すなわちトピックを算出することも可能であるし，ある著者の複数論文と特徴を同じにする他の著者や，著者と特徴を同じくする研究会を検索するということも可能である．さらに，情報処理学会の特定の研究会や，ある年発行されたすべての論文に対するベクトルも算出可能になる．これらのベクトルを使って，図 1-(b)に示したような「学会活動に意味のある演算」が可能になることが期待できる．

3.2 Doc2Vec の実用事例

Doc2Vec の有効性はすでの多くの先行研究により実用的な効果があることが示されている．例えば，インタネット上のオープンなコンテンツ Wikipedia，コーパスとして提供されている各種データ，AP 通信などのニュースを入力に各記事のベクトル化によって検索，トピック抽出，意味解析が可能になっている[8]．さらに，同様な事例が特許の解析[9]，プログラムのソースコード解析[10]，歌詞の解析[11]等多く存在する．本研究と類似の分野でも，学会の特定分野のトピックについて Doc2Vec を用いた研究[12]がある．

3.3 本研究のアプローチ

本研究では，Doc2Vec を活用して情報処理学会で蓄積しているコンテンツをベクトル化し，その関係やトレンドを解析する．さらに，自然言語や人工知能の研究者だけでなく，一般の研究者が対象となる研究分野での特徴を把握するために活用可能な情報システムとして実装した．

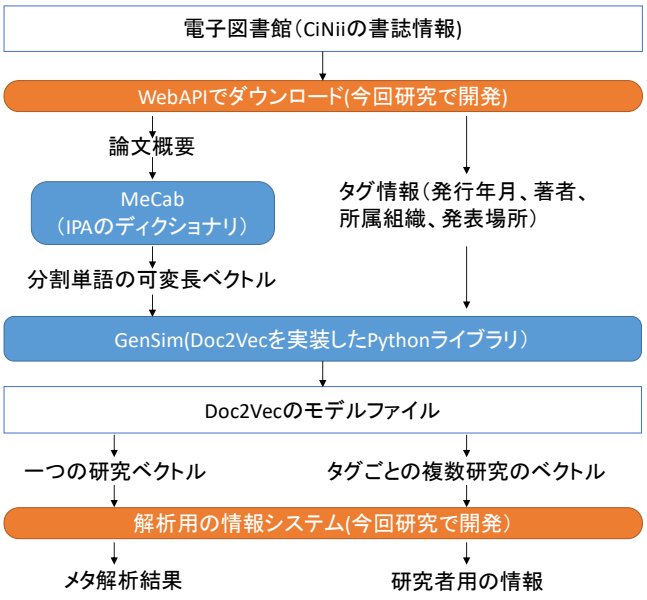

図 1　本研究のデータフロー

Figure 1　Dataflow of the research

図 2 に，本研究全体のデータフロー図を示す．まず，CiNii 上の情報処理学会電子図書館のデータを WebAPI 経由でダウンロードする．論文概要部分は MeCab[13]および IPA のディクショナリ[14]を使い単語列に分割する．分割された単語列に，タグとして，書誌情報の著者名，所属組織名，発表研究会，発行年，発行年月を付与する．Doc2Vec の実装である Python ライブラリ Gensim にこれらのデータを与え Doc2Vec 用のモデルを生成した．本論文で使用したデータは，2017 年 11 月 4 日現在のデータである．モデル生成時に指定した Gensim のパラメタを表 1 に示す．

表 1　モデル生成時の Gensim パラメタ

Table 1　Gensim parameters for generating model

パラメタ	指定値
エポック数	20
ベクトル次元	200
Window	20
最小出現回数	10
最大出現頻度	0.1%
アルゴリズム	DBOW/ DM-PV の両方で作成

生成されたモデルの概要を表 2 に示す．

表 2　IPSJ コンテンツモデルの概要

Table 2　Outline of IPSJ contents model

パラメタ	値/備考
コンテンツ数	135,764(単語数 51 以上を対象)
単語数	39,396
タグ種類(出現数)	発行年(45)，著者(95634)，所属組織(40244)，発表場所(5630)
モデルのサイズ	DBOW,DMPV ともに 285MB

今回の研究で開発した情報システムのソフトウェア環境を表 3 に示す．

表 3　本研究のソフトウェア環境

Table 3　Software environments of the study

種類	ソフトウェア名	バージョン/備考
OS	Ubuntu	16.04.2 LTS x86_64
言語	Python	3.5.2
Doc2Vec	Gensim	2.3.0
単語分割	Mecab	MeCab-python3 0.7 + MeCab 0.996
ビジュアライズ	Tensorboard	1.0.0a6

4. 実験，評価

本研究では，情報処理学会全体のトレンドを把握するとともに，その結果に基づき，各研究者が自分の実務に役に立つ情報システム化を行った．本章では，この研究課題それぞれについて，実験とその結果を述べる．

4.1 学会全体の発行年によるトレンド

学会全体で，年という単位でトレンドの推移があることが明確になった．図 3 は，Gensim で，各暦年に発行された全論文のベクトルを Doc2Vec の DM-PV アルゴリズムを使って算出し，その推移を，TensorBoard[15]が実装している PCA アルゴリズムで 3 次元に圧縮した結果である．

論文数も少なかった 1970 年, 80 年代を除き，それ以降，学会全体での Doc2Vec ベクトルには，一定の方向性がある．例えば，ある年（n 年）に発行された全研究のベクトルがあったとき，その翌年の全研究のベクトルは以下の式で求めることができる．今回の研究で取得したデータの場合，Doc2Vec のアルゴリズムを DM-PV にした場合，以式(1)が成立することを確認した．

$$\overrightarrow{n+1} = \vec{n} + \vec{n} - \overrightarrow{n-1} \qquad (1)$$

この関係が，分野や暦年に関係なく成り立つことが明確な場合，図に示したような関係を過去のトレンドの確認だけでなく将来のトレンドを予測することもできる可能性がある．残念ながら，現状，（Gemsim の）Doc2Vec 関連のパラメタによっては，上記のような明確な関係が得られない場合もある．

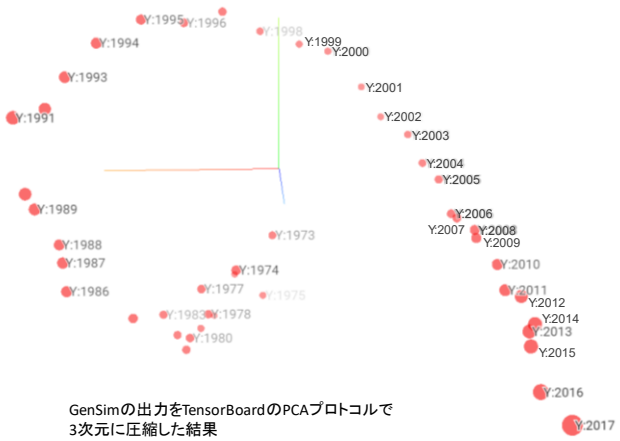

図 3　情報処理学会コンテンツベクトルの経年変化
Figure 3　Trend of IPSJ contents vectors in these 40 years

4.2 学会の研究会のトレンド，アイデンティティ

2017 年 10 月現在で，電子図書館に論文を登録している研究会の数は 39 である．各研究会のある年のベクトルを Doc2Vec で算出し，それを比較することにより，時代よるトレンド, 他の研究会からのアイデンティティを評価した．

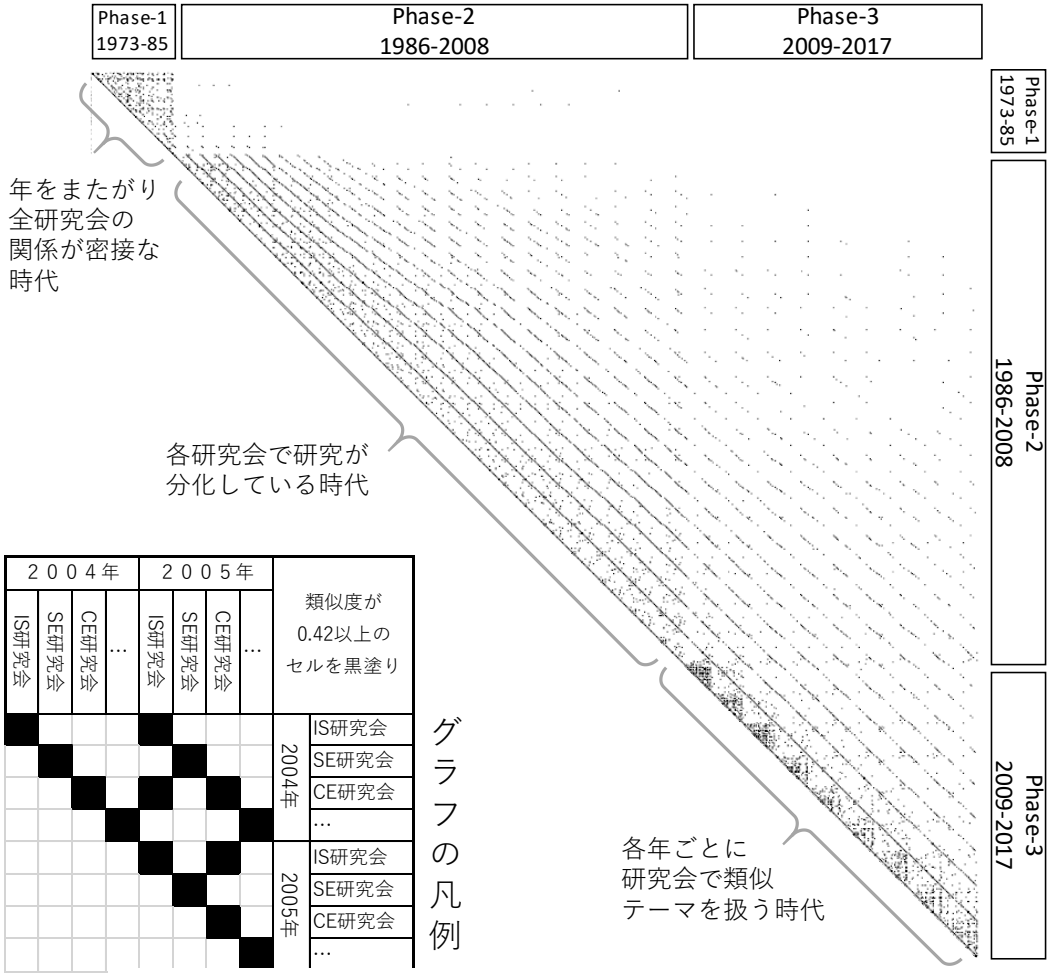

図 4　学会研究会のトレンドの一例
Figure 4　A trend of SIGs of IPSJ

図 5　IPSJ2Vec のコンテンツ検索システム画面
Figure 5　Contents search screen of IPSJ2Vec

図 4 は，1973 年以降の一年ごとの全研究会のベクトルの類似度をクロス分析し，類似度の比較的高い（Microsoft Excel の条件付き書式での初期値の 0.462）相関を黒塗りした結果である．左上から右下への対角線は，同一年の同一研究会で，類似度は 1 固定である．その他の部分で，斜めに黒塗りが続いている部分は，年が異なっても，同じ研究会でベクトルに類似度が高いことを示している．

学会全体の研究会の関係をマクロにみてみると，年代により，大きく 3 個のフェーズがあるように見える．第一のフェーズは 1973 年から 1985 年で，年をまたがって各研究会のベクトルの類似度が高いフェーズである．このフェーズは研究会数もあまり多くなく，学会全体での交流が強かったことが想定できる．第二のフェーズは 1986 年から 2008 年で，各研究会の独立性が高いフェーズである．すなわち，研究会ごとの類似は，各年度をまたがって強く観測（対角線より右側の斜線）できるが，研究会をまたがった類似度は他のフェーズに比べて低い．第三のフェーズは 2009 年以降で，研究会をまたがり，各年で強い類似度が観測できる．ただ，第一のフェーズのように，年をまたがって強い類似を示すことはなく，各年でテーマが固まっているように見える．具体的なトピックについては未解析であるが，ビッグデータ，人工知能等の研究会横断的なトピックが影響している可能性がある．

4.3 情報システムの実装

今回作成した情報処理学会コンテンツの Doc2Vec モデルデータベースを使用し，研究者が実務で役に立つ Web ベースの情報システム IPSJ2Vec を開発した．IPSJ2Vec は，各研究者が自分のマシンで環境構築することなく容易に学会レベルの解析や，単なる既存研究検索や，新しい論文の新規度，各研究会との類似度なども検索可能である．IPSJ2Vec の検索の画面例を図 5 に示す．

本情報システムの特長は 2 点ある．最初の特長は，類似情報とともに，相違情報の入力が可能ということである．Doc2Vec のサポートしている意味の演算を簡単に検証可能になる．例えば，類似情報に 2000 年当時のトピック（例えば「オブジェクト指向」）と，Y:2017 を入れ，相違情報に Y:2000 と入れると，2017 年現在の同種のトレンドが表示される（正確にいうと表示されることが期待できる）．もう一点の特長は，モデルデータベースでは訓練していない新しい論文の概要も入力可能にしている点である．研究者が新たな論文を執筆した場合，その論文の新規性のチェック，類似論文の検索や，その論文を投稿すべき研究会の推薦等が可能になる．

5. 本研究の信頼性への懸念

本研究は目標としてレビュー論文やサーベイ論文の負荷低減や，そもそも，レビュー論文が意図している課題をレビュー論文無しで解決することを挙げた．残念ながら 4 章で述べた解析結果や開発した情報システム IPSJ2Vec だけでは，現状この目標は達成できていない．本章では，現時点で見えている本研究の課題を列挙する．

主要でない論文によって，信頼にかける結果がでてくる危険性がある．情報処理学会の場合，論文誌のように厳密な査読を経ている論文だけでなく，研究会の報告では信頼性に欠けるものがある可能性は否定できない．この場合，人間による判断や，従来研究である，サイテーションを加味した論文の選択や重みづけ等を加える必要がある．

本研究での統計解析方法は将来的に改善すべき懸念がある．Doc2Vec は類似度があるということは分かっても，その理由については分からない. また, 今回使用した Python ライブラリの Gensim や，ビジュアライズで使用した TensorBoard は，パラメタによって，結果が大きく異なることがある．使用したライブラリの信頼性にも懸念があり，またツールの使用方法が適切でない危険性もある．全体的に，2017 年現在のライブラリの技術レベルでは，レビュー論文，サーベイ論文を置き換えるような情報システムは困難であると筆者らは考えている．

6. おわりに（結論と今後の展望）

本研究では，情報処理学会という特定の学会を対象に学会全体でのトレンド，各研究会の位置づけ等を解析するとともに，その知見を活かして，一般学会員が有効活用可能な情報システムを開発した．

今回提案した情報システムの実装は情報処理学会電子図書館が使用している NII 学術情報ナビゲータ(CiNii)のフォーマットを使用している．しかし，同じ考え方は，電子的に概要レベルのコンテンツを取得可能な他のフォーマットでも実装可能である．例えば，科学技術情報発信・流通総合システム（J-STAGE）3に対応したフォーマットに対応すれば, その応用範囲は多くの学会もカバー可能と考える．Doc2Vec は，言語を限定しないため，日本語以外の他言語や，複数言語によるハイブリッドなコンテンツに対しても有効に解析可能である．例えば，情報処理学会と欧米の同種学会のトレンド比較というようなことも可能だと考える．

本研究により，興味深い学会全体のトレンドや方向性が明らかになったが，所期の目標である，レビュー論文，サーベイの論文の置き換えは, 現状の技術レベルでは難しい．しかし，人工知能関連のライブラリは，日進月歩の世界であり，今後も今回使用した Doc2Vec の改善や置き換わるようなライブラリが登場すると思われる．これらの動向をウォッチすることにより，新たなブレークスルーが起きる可能性は決して低くないと考える．

参考文献

[1] 鷲崎弘宜, 深澤良彰「パターン: 引用からの品質」,第７ 回パターン研究会,(2001).
[2] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pages 1188–1196
[3] 難波英嗣, 奥村学, 論文間の参照情報を考慮したサーベイ論文作成支援システムの開発", 自然言語処理, Vol.6, No.5, 1999
[4] J. Beel and B. Gipp, "Google Scholar's ranking algorithm: The impact of citation counts (An empirical study)," 2009 Third International Conference on Research Challenges in Information Science, Fez, 2009, pp. 439-446.
[5] Brin, S.; Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine" (PDF). Computer Networks and ISDN Systems. 30: 107–117. ISSN 0169-7552. doi:10.1016/S0169-7552(98)00110-X.
[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111–3119
[7] “Gensim”.https:/radimrehurek.com/gensim/, (参照 2017-11-05).
[8] Jey Han Lau, Timothy Baldwin, “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation”
[9] 稲垣祐一郎,深層学習の最近の進展,」, みずほ情報総研 技報, Vol.7, No.1（2015/10）
[10] David C, “Doc2Vec to Assess Semantic Similarity in Source Code”, https://gab41.lab41.org/doc2vec-to-assess-semantic-similarity-in-source-code-667acb3e62d7, (参照 2017-11-05)
[11] Ö. Çoban and I. Karabey, "Music genre classification with word and document vectors," 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, 2017, pp. 1-4.
[12] 中村雄太, 浅野泰仁, 吉川正俊, 「分散表現空間解析モデルに基づく研究トレンドに関する考察」, DEIM Forum 2017, http://db-event.jpn.org/deim2017/papers/305.pdf, (参照 2017-11-05)
[13] “Mecab”, http://taku910.github.io/mecab/, (参照 2017-11-05)
[14] “mecab-ipadic-neologd”, https://github.com/neologd/mecab-ipadic-neologd/wiki/Home.ja, (参照 2017-11-05)
[15] “TensorBoard”, https://www.tensorflow.org/get_started/summaries_and_tensorboard, (参照 2017-11-05)

3 https://www.jstage.jst.go.jp/