マルチメディア通信と分散処理ワークショップの研究テーマの変遷

木原 民雄[1,a)]　加藤 由花[2,b)]

概要：本稿は，第 25 回を迎える「情報処理学会マルチメディア通信と分散処理（DPS）ワークショップ」におけるこれまでの研究活動を総括し，今後の研究の方向性を見据えることを目的に，発表論文のテキストマイニングにより研究動向の分析を行った結果を報告するものである．分析方法としては，まず，学術データベースから入手した発表論文の書誌情報を利用し，研究動向と関連の深いキーワードを抽出する．次に，これらのキーワードの出現頻度を時系列データとして整理することで，1993 年から 2016 年までの研究テーマの推移を示す．その結果を分析し，本研究分野の今後の展望を議論する．

1. はじめに

情報処理学会マルチメディア通信と分散処理（DPS）研究会では，分散コンピューティング，マルチメディア情報処理，プロトコルなどの研究分野について，活発な研究発表が行われている．これらの研究について，通常の研究会ではできない深い議論を行うため，1993 年より合宿形式のマルチメディア通信と分散処理ワークショップ（DPS ワークショップ）を継続的に開催している [1]．2017 年度は，本ワークショップも第 25 回という節目の年を迎えるため，四半世紀にわたる研究活動を総括し，今後の研究の方向性を見据える未来指向の企画が検討されている．このような背景の下，本稿では，この議論のベースとなることを目的に，25 回分のワークショップ発表論文のテキストマイニングを行い，そこからの研究動向の抽出を試みる．

本稿の構想は，我々のこれまでの DPS ワークショップとの深い関わりから生まれた．第一著者の木原は，2002 年函館で開催された第 10 回ワークショップのプログラム委員長，および 2009 年層雲峡で開催された第 17 回ワークショップのワークショップ委員長を務めている．ちなみに初参加は 1994 年（飯坂開催）であり，これが初発表，初座長であった．第二著者の加藤は，木原がワークショップ委員長を務めた 2009 年にプログラム委員長，そして今年度 2017 年に温根湯で開催予定の第 25 回ワークショップのワークショップ委員長を務める．ワークショップ初参加は 2003 年（阿蘇開催）である．我々は，このように，長期にわたり DPS ワークショップの運営に携わり，当該分野の動向を見据えてきた知見を，本稿における分析に活かせるのではないかと考えた．

DPS ワークショップでは，過去にもいくつかの節目で，将来の研究を見据えたパネル討論等が行われてきた．1999 年（別府開催）には「パネルディスカッション：21 世紀の DPS について」，2011 年（十和田湖開催）には「20 周年記念パネル：過去から未来へ」が開催されている．本稿では，これらを踏まえ，より定量的な議論を可能とするために，発表論文の分析という手法を採用することにした．これには 2 つの理由がある．1 つは，学術データベースが整備され，25 回分の発表論文データの入手が容易になり，時系列データとして分析が可能になったことである．そしてもう 1 つは，自然言語処理のための様々なツールが整備され，容易に利用可能になったことである．

学術情報の分析に関しては，主に自然言語処理の分野で，これまでも様々な研究が行われてきた．例えば，大量の自然言語データから有用な情報を抽出する技術（テキストマイニング）を用いることで文献書誌情報などを分析し，研究動向を調査する研究などがある [2][3][4]．ここでは，時系列データなどで単語の出現頻度を数え，どういった単語が時系列的に増加したか減少したかを分析し，興味がどのように移り変わったかを調べることなどが行われている．近年では，陽に表現されない各文書のトピックを，確率分布として推定するトピックモデルも多く用いられている．トピックモデルは，BoW（Bag of Words: 出現した単語の

[1] 昭和女子大学
Showa Women's University, Setagaya, Tokyo 154–8533, Japan
[2] 東京女子大学
Tokyo Woman's Christian University, Suginami, Tokyo 167–8585, Japan
[a)] kiharatamio@swu.ac.jp
[b)] yuka@lab.twcu.ac.jp

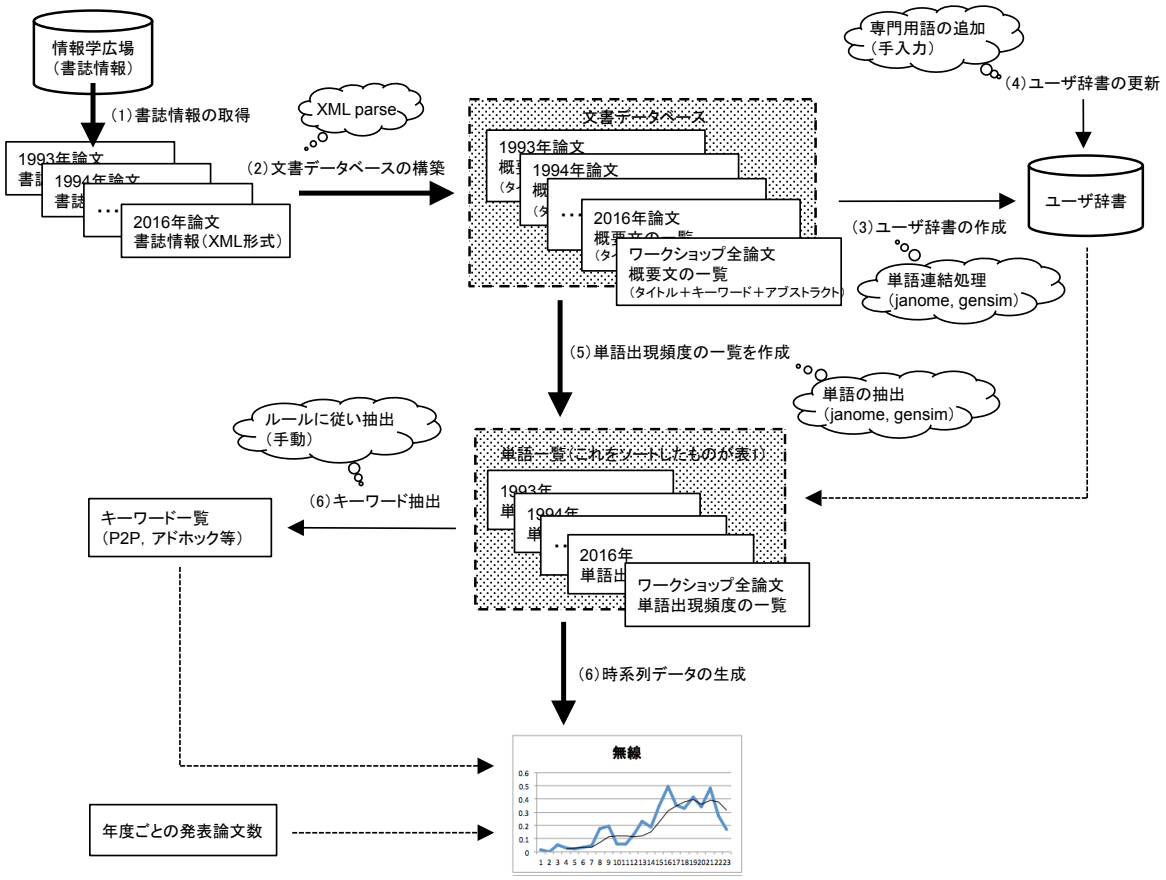

図 1　分析の方法

頻度を並べたベクトル）表現されたある文書の生成過程を確率的にモデル化したものであり，ある文書に含まれる各単語は，文書固有のトピック比率に従ってあるトピックを選択した後，そのトピック固有の単語出現確率分布に従って生成されると仮定するものである．代表例として Latent Dirichlet Allocation（LDA） [5] などがあり，様々な分野に応用されている．

本稿では，対象とする発表論文数が 1145 件と，本格的なテキストマイニングの対象としては小規模であること，ワークショップという性質上，対象となるトピックはそれほど多岐に渡るわけではないことを鑑み，統計的手法を用いることはせず，発表論文中の単語の出現頻度とその推移を分析することとする．

2.　分析の方法

分析の手順を **図 1** および以下に示す．

(1) 学術論文データベースから文献書誌情報を取得する．
(2) 取得した書誌情報から，文書データベースを構築する．
(3) 単語連結処理をした用語をユーザ辞書に追加する．
(4) 専門用語等を手動でユーザ辞書に追加する．
(5) ツールを用いて年度ごとの単語の出現頻度を出力する．
(6) キーワードを抽出し，その時系列での推移を分析する．

次に，それぞれの手順について，詳細を説明する．

2.1　文献書誌情報の取得と文書データベースの構築

まず，分析の手順（1）と（2）に相当する処理について説明する．本稿では，文書書誌情報は，情報処理学会の学術論文データベースである情報学広場 [6] から取得する．具体的には，1993 年から 2016 年までの全ワークショップ発表論文について，その書誌情報を XML 形式で取得する．その後，各論文ごとに，タイトル，キーワード（取得できる年度のみ，具体的には 2012 年から 2016 年のデータ），アブストラクトを連結して 1 行としたデータを，年度ごとに 1 ファイルとし，これを 25 ファイル分まとめて文書データベースとする．この文書データベースが，単語抽出の対象となる文書を格納したデータベースとなる．

つまり，本稿では，各発表論文のタイトル，キーワード，アブストラクトに含まれる単語を利用し，その単語を研究テーマを示すキーワードととらえる．

2.2　ユーザ辞書の作成

次に，ユーザ辞書を作成する．これは分析の手順（3）に相当する処理である．単語の出現頻度を調べるためには，まず対象となる文書を形態素解析し，分かち書きを行う必要がある．このとき，辞書登録されていない複合語は分割され，分割された単語としてカウントされてしまう．例えば，「情報処理」という単語が辞書に登録されていないと，

この単語は「情報」と「処理」として別々にカウントされ，「情報処理」という単語は存在しないことになる．そのため，対象となる文書を読み込み，単語の出現頻度および単語の対（バイグラム）の出現頻度を合わせて調べることにより，出現頻度の高いバイグラムを一単語に置き換える処理を数回（本稿の場合は 5 回）繰り返す．そうして生成された単語が辞書に含まれていない単語であれば，その単語をユーザ辞書に追加する処理を行う．

自然言語処理やテキストマイニングのためのツールには様々なものが存在するが，本稿では，形態素解析用に Python 用ライブラリである `janome` [7] を，頻出単語ペアの検出用に Python のトピックモデル用ライブラリである `gensim` パッケージ [8] の `Phrases` クラスを用いる．

2.3 専門用語等の追加

抽出した複合語をユーザ辞書に追加しても，そもそも辞書に含まれていない専門用語の抽出はできない．そのため，重要な用語については手動で辞書に追加する必要がある．これは分析の手順（4）に相当する処理である．専門の辞書を利用する方法等も考えられるが，本稿では，発表論文のタイトルから，年度ごとに重要と思われる用語を人手で抽出し，ユーザ辞書に追加する．具体的には，今回，「クラウド」「P2P」「フラッディング」などの用語がユーザ辞書に追加された．

2.4 単語出現頻度とその時系列データの生成

最後に，分析の手順（5）と（6）に相当する処理について説明する．ここでは，前項で生成したユーザ辞書を用いて，各単語ごとの出現頻度一覧を作成する．この処理には，前述した `gensim` を利用する．その結果，年度ごとの単語出現頻度一覧が取得できる．ここでは，全発表論文の書誌データを 1 ファイルに格納した文書データベースを利用し，25 回分の発表論文全体に対する単語出現頻度一覧も合わせて取得しておく．このとき，研究テーマとしてふさわしくない単語（「研究」「考察」など）は人手で取り除く．

次に，出現頻度の推移を時系列データとして分析する対象となる「キーワード」を抽出する．これは，本研究分野の動向を示す重要な単語であり，以下の条件を満たす単語を選択する．

- 発表論文全体において出現頻度が高いこと
- それ以外の単語で，ある一時期の出現頻度が高いこと
- それ以外の単語で，本分野において重要な単語

抽出されたキーワードについて，年度ごとの出現回数を出力とする．このとき，年度ごとに発表論文数が異なるので，この影響を排除するために，キーワードの出現回数を発表論文数で割ることにより，値の正規化を行う．

3. 分析の結果

2 章で述べた分析の方法で，まず，単語の出現頻度を調べた．抽出された単語の総数は 5201 個であり，分析に適さない単語を除いた上で，意味の近い単語をグルーピングすることにより（映像，動画，ビデオを「ビデオ」にまとめる等），84 個の単語を分析対象とすることにした．代表的な年度ごとに，上位 20 位までの分析対象の単語とその出現回数をまとめた結果を **表 1** に示す．この結果から，年度ごとに頻出単語の傾向は変化していることがわかる．ちなみに，年度ごとの発表論文数（ポスター・デモ発表を含む）は **表 2** に示すとおりである．年度ごとに差はあるものの，40〜70 件程度で推移していることがわかる．なお，1993 年はワークショップは 2 回開催されているが，ここでは 2 回の合計値を示している．1997 年はワークショップは開催されていない．

次に，前章で示した条件に従い，ワークショップの研究動向を表現するキーワードを抽出する．今回は，「分散処理」「位置情報」「災害情報」「無線通信」「交通システム・車」「インターネット」「P2P・オーバレイ」「プロトコル」「ビデオ」「センサー」「マルチメディア」「Web」「エージェント」「教育支援」「QoS」「電力」「アドホック」「暗号・セキュリティ」「スマート」「クラウド」「ユビキタス・IoT」の 21 の単語をキーワードとして抽出した．

これらのキーワードについて，年度ごとの出現回数（発表論文数で正規化した値）を時系列でグラフ化した結果を **図 2** に示す．例えば， 表 2 において「分散処理」というキーワードは，常に 20 位以内に入ってはいるものの，1993 年から 2016 年にかけて減少し続けていることがわかる．このような変化を年度ごとにプロットしたものが 図 2 である．年度ごとの発表論文数は数十件程度と少ないため，年によって発表テーマにはばらつきが生じる．そのため，年度ごとの頻度を青線（実線）で示すとともに，時間推移の傾向を見るために，3 区間の移動平均を黒線（破線）で示した．

4. 考察

分析においては，単語の出現回数を併記したが，これは単語のグルーピングの仕方により変動する値であり，絶対数にそれほどの意味はない．あくまでも目安として用いるべきである．例えば，「分散処理」のように一般的な単語の方が，そのグループ内に多くの単語を含み（分散システム，分散オブジェクト，分散アルゴリズムなど），総数が多くなる．また，一般的な用語は，専門用語に比べ，1 つの論文のアブストラクト中で何度も言及される傾向があり，出現頻度が高くなる．年度による推移の仕方を分析するのが原則である．

表 1 年度ごとの上位 20 位までの出現単語とその出現回数（抜粋）

1993 年	2000 年	2008 年	2016 年	全体
ネットワーク（27）	通信（19）	災害情報（21）	交通システム（35）	通信（399）
分散処理（26）	インターネット（14）	無線通信（19）	位置情報（23）	分散処理（254）
マルチメディア（25）	分散処理（10）	交通システム（17）	災害情報（22）	位置情報（244）
通信（25）	災害情報（9）	センサー（15）	通信（19）	災害情報（235）
プロトコル（19）	パケット（8）	シミュレーション（15）	シミュレーション（12）	無線通信（227）
データ処理（13）	IP（7）	分散処理（11）	配送（11）	交通システム（215）
オブジェクト（9）	QoS（7）	暗号（11）	移動（11）	シミュレーション（158）
IP（8）	ネットワーク（7）	救急救命（10）	ビデオ（10）	インターネット（151）
メッセージ（8）	教育支援（7）	位置情報（10）	センサー（10）	P2P（147）
負荷分散（7）	エージェント（6）	行動認識（8）	クラウド（9）	プロトコル（143）
OSI（6）	コンテンツ（5）	P2P（8）	行動認識（8）	ビデオ（132）
LAN（5）	ビデオ（5）	携帯システム（7）	エージェント（7）	ネットワーク（128）
メディア（5）	メッセージ（5）	プロトコル（7）	インターネット（7）	センサー（123）
リアルタイム（5）	Web（4）	インターネット（5）	スマート（7）	マルチメディア（120）
画像（5）	マルチメディア（4）	Web（4）	ネットワーク（7）	情報検索（110）
暗号（5）	メディア（4）	配信（4）	無線通信（7）	データ処理（102）
インターネット（4）	位置情報（4）	コンテンツ（4）	IoT（6）	Web（95）
エージェント（4）	帯域制御（4）	マルチホップ（4）	電力（6）	エージェント（92）
ビデオ（4）	資源管理（4）	ビデオ（4）	分散処理（6）	IP（90）
経路制御（4）	オブジェクト（3）	IP（3）	負荷分散（5）	教育支援（88）

この前提の下，図 2 の推移を見ると，研究テーマには変遷があり，上昇傾向にあるもの，継続して長く出現しているもの，最近ではあまり見られなくなったものなど様々である．以下，それぞれの傾向を分析した後，今後の展望と分析の限界について述べる．

4.1 研究テーマの変遷

ワークショップが始まった 1993 年は，IIJ が日本初の商用インターネットサービスを開始した年である．「インターネット」は初期の頃からの研究テーマであり，現在まで長期間継続して出現しているキーワードである．当初はインターネット自身の研究として，様々なプロトコルや方式の提案が行われ，その後インターネットを利用した研究へと推移してきたと考えられる．「Web」は 1994 年に最初に出現し，これも現在まで継続して出現するキーワードとなっている．その他，「プロトコル」「ビデオ」「セキュリティ」などが継続して長く出現し続けているキーワードである．これらは，新しいトピックが出現し，研究テーマが変遷する中でも，DPS ワークショップのベースとして存在し続けているキーワードと考えられる．

表 2 年度ごとの発表論文数の推移

年度	発表数	年度	発表数
1993	69	2006	35
1994	34	2007	54
1995	38	2008	58
1996	69	2009	53
1998	40	2010	45
1999	30	2011	49
2000	45	2012	46
2001	40	2013	47
2002	52	2014	48
2003	51	2015	51
2004	66	2016	42
2005	83	2017	??

ワークショップの名称になっている「分散処理」と「マルチメディア」については，当初は中心的なトピックであったが，時代とともにこれらの技術の利用が当たり前になり，現在では，マルチメディアという単語を前面に出した研究はほとんど行われていない．インターネット上でマルチメディア情報を扱うために必要になる「QoS」に関する研究や，分散処理を実現するための「エージェント」に関する研究なども，比較的長期にわたり出現していたキーワードであったが，現在ではほとんど見られなくなった．当分野のアプリケーション例の一つに「教育支援」があり，インターネットを利用した遠隔授業等の研究が活発に行われた．しかし，これも技術的なものから教育内容等に課題が推移していき，出現頻度が低くなったキーワードである．

2000 年代に入ると，「無線通信」に関する研究が活発化し，現在までこの傾向は続いている．最近やや減少しているが，キーワードとして「アドホック」が登場したのも 1999 年である．また，少し遅れて「P2P・オーバレイ」が登場し，中心的な研究テーマとなっていった．

社会的な出来事が研究テーマに影響を与える事例も見られる．「災害情報」は，何度か浮き沈みがあるが，2007 年新潟県中越沖地震，2008 年岩手・宮城内陸地震の後に急増

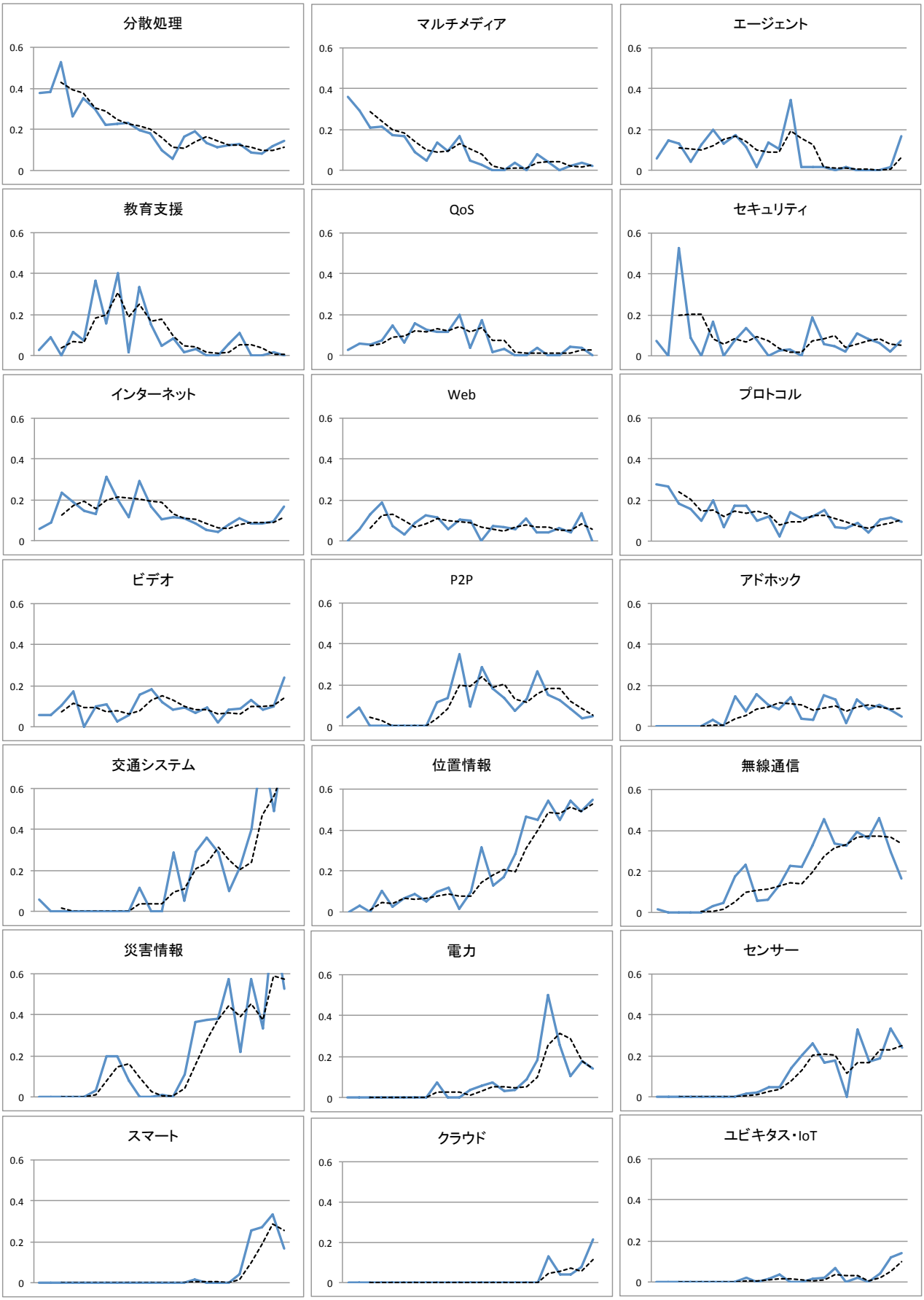

図 2　キーワードの出現頻度の推移

し，一つの大きな研究テーマになっている．2011 年の東日本大震災後は，「電力」の出現回数が急増し，省電力に関する研究が活発化した．また，これをきっかけにクラウド環境の利用が急速に普及したこともあり，「クラウド」の出現頻度も増加している．「電力」については，2011 年以降に急増しているが，それ以前から，「無線通信」の出現頻度と合わせて増加していた．電力消費の問題が顕在化したためと考えられる．

最近では，「位置情報」「交通システム・車」の出現頻度が伸び続け，現在のワークショップでの中心テーマの一つとなっている．これらのキーワードは，初期のワークショップにおける「分散処理」や「マルチメディア」に相当する役割を担っているとも考えられる．ここ数年，出現頻度が上昇しているキーワードは，「スマート」と「ユビキタス・IoT」である．特にスマートの伸びは大きい．また，ユビキタス・IoT という単語を明示しているものは少ないが，IoT，センサー，位置情報，行動認識など，広い意味でこの分野に含まれる研究テーマが増えている．これは，世の中の趨勢でもあるが，DPS ワークショップにおいても勢いのあるテーマである．

ちなみに，登場する単語の種類は徐々に増えてきており，最近では，MAC レイヤからアプリケーションレイヤまで，非常に多様なテーマが対象になっていることがわかる．出現頻度が低い，または出現しないキーワードの中で重要だと思われるものには，ビッグデータ，人工知能（AI），インタラクション関連，ソーシャル関連の用語などがある．また，SDN や OpenFlow は，DPS ワークショップの関連テーマであるにも関わらず，予想外に出現頻度が低く，伸びも少ない．

4.2 今後の展望

今回の分析で，情報処理技術が現在のように普及したのは，これまでの地道な研究の成果であることを改めて認識した．その上で，情報処理技術の活用範囲が広がっており，これをどう使っていくか，新しい領域を開拓していく必要性も感じた．

一方，学生や次世代の研究者・技術者の継続的な育成のためには，研究テーマの持続性も考慮すべき事項である．例えば，災害情報システムなど，社会的なトータルシステムは，課題が階層的で多岐に渡り，複数のサブシステムの研究が必要であるため持続性が高い．アドホックルーティングやグループ通信などの方式研究は，大きな手法は変わらないものの，段階的な改善方法や組み合わせ方法が必要であるため，持続性が高い．安心や感性など，人間の心理的な尺度を研究テーマとするものは，尺度の設定によりゴールが異なるなど，多面的な取り組みがあり得るため持続性が高い．DPS ワークショップが，これらの研究テーマに継続的に取り組んできた意義は大きく，今後もひとつの方向性として研究が推進されることを期待したい．

4.3 分析の限界

テキストマイニングの難しさは，言語表現の曖昧性，多義性にある．ある程度限定されたトピックを扱うワークショップ論文であっても，ある単語の解釈が一通りに定まらない場合，時代によって異なる意味で用いられる場合などが散見された．例えば，「ブロードキャスト」という単語は，パケットのブロードキャストを意味する場合もあれば，放送を意味する場合もある．「インターネット」という単語は，1990 年代と現在とでは，意味するところが異なっていると考えられる．さらに今回，単語のグルーピングにより分析対象の単語を決定したが，ここではグルーピングが恣意的になりがちであり，これが分析結果に影響を与えている可能性がある．そもそも，書誌情報に明示的に含まれる単語により研究テーマや研究内容を表現できるという立場を取っているが，この方法では潜在的なトピックを的確にとらえられない可能性もある．

また，ワークショップ論文の場合，情報処理という大きな分野の研究動向（またはもう少し限定して分散処理やネットワークの分野の研究動向）以外に，ある組織や研究グループの貢献が分析結果に影響を及ぼす効果も無視できない．今回，これらの影響に対する考察は行っていない．

本稿における分析結果は，これらの問題点を認識した上で解釈する必要がある．

5. おわりに

本稿では，DPS ワークショップにおける研究テーマの変遷を，発表論文のテキストマイニングを行うことにより分析した．ここでは，学術データベースから入手した発表論文の書誌情報を利用し，論文中の単語の出現頻度とその推移を調査した．1145 件の論文書誌情報から抽出した 21 のキーワードについて出現頻度の推移を調査し，研究テーマの変遷を示すとともに，今後の展望を述べた．

本稿で用いたデータは，テキストマイニングの対象としては小規模なものであり，意外性のある結果を得るには不十分であると考えられる．25 回分のワークショップの研究テーマの変遷は，人手でもある程度把握可能な規模であるだろう．しかし，本稿での分析を通して，24 年におよぶワークショップの歴史を振り返り，未来を見据えた議論の土台を構築できたようにも思う．今後，蓄積され続けていく論文の本文を含めた分析，他分野との関連の考察など，より詳細な分析を期待したい．

参考文献

[1] DPSWS2017: 第 25 回マルチメディア通信と分散処理ワークショップ，情報処理学会 DPS 研究会（オンライン），入手先 ⟨http://www.dpsws.org/2017/⟩（参照 2017-06-20）．

[2] 村田真樹，一井康二，馬　青，白土　保，井佐原均：過去10年間の言語処理学会論文誌・年次大会発表における研究動向調査，言語処理学会第11回年次大会発表論文集，pp. 77–80 (2005).
[3] 那須川哲哉，西山莉紗，吉田一星：学術文献のテキストマイニング，言語処理学会第20回年次大会発表論文集，pp. 800–803 (2014).
[4] 藤井章博：IEEE論文に基づくIoT研究動向の計量書誌学的調査，科学技術動向，No. 149, pp. 19–24 (2015).
[5] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Reserach*, Vol. 3, No. 2, pp. 993–1022 (2003).
[6] 情報学広場：情報処理学会電子図書館，情報処理学会（オンライン），入手先 ⟨https://ipsj.ixsq.nii.ac.jp/ej/⟩（参照2017-06-20）.
[7] Janome: Japanese morhological analysis engine written in pure Python, mocobeta (online), available from ⟨https://github.com/mocobeta/janome⟩ (accessed 2017-06-20).
[8] Gesim: topic modelling for humans, RadimRehurek (online), available from ⟨http://radimrehurek.com/gensim/⟩ (accessed 2017-06-20).