研究動向分析のための 論文のデジタルテキスト化とマイニングシステム

増田 勝也† 丹治 信† 植松 すみれ† 美馬 秀樹†‡
†東京大学 知の構造化センター ‡東京大学大学院工学系研究科
{masuda,tanji,uematsu}@cks.u-tokyo.ac.jp, mima@t-adm.t.u-tokyo.ac.jp

1 はじめに

本論文では、学会における研究動向分析を目的として、論文のデジタルテキスト化手法、およびそれを利用し分析するためのマイニングシステムについて述べる。論文データベースを対象とした学会における研究動向の分析は、各分野において数多く行われている [6, 8]。一般的にこのような動向調査は多くの時間と労力を要するため、全体を俯瞰し、様々な条件を設定した分析を容易に可能とする枠組が求められる。また、近年の論文に対しては機械処理可能なデジタルテキストが存在しそれを対象としてテキストマイニングを行うことで動向分析が可能であるが、古い年代の論文に関しては、そもそもデジタルデータ、特に本文のテキストデータは存在せず、分析対象が容易にデータ化可能な書誌情報のみを用いたものに限られてしまう。

そこで本論文ではこれまでの言語処理学会年次大会予稿集を対象として OCR を用いたデジタルテキストデータの構築を行い、検索・可視化システム MIMA サーチを用いた研究動向分析のためのマイニングシステムを構築する。東京大学知の構造化センター「岩波書店『思想』の構造化プロジェクト」において構築されたデジタル化手法 [5] を適用し、可能なかぎり人手を用いることなく論文画像から自動的にテキストデータの構築を行う。また、検索・可視化システム MIMA サーチを用いて論文データを分析可能にすることで、学会の論文全体の俯瞰を可能にし、ユーザがさまざまな条件を設定した上での研究動向分析をリアルタイムで行える環境を構築する。

2 関連研究

言語処理学会に対しては、第 10 回までの論文誌・年次大会予稿について研究動向調査が行われている [6]。ここでは、論文誌・年次大会における全体及び研究機関ごとの発表件数、論文タイトル中の形態素を用いた研究分野ごとの発表件数の推移により分析を行っている。本論文ではこのような分析をユーザが自由に条件を変更してリアルタイムに実行可能なシステムを構築する。また研究動向分析のための技術として、論文の表題を自動的に解析し、動向情報を抽出・可視化するシステム [7] や、CiNii データベースを対象とし、検索結果の研究分野の自動分類および技術動向の抽出を行うシステム [4] などが構築されている。

表 1: ブロック属性の分類精度

	適合率	再現率
論文タイトル	0.780	0.799
著者	0.829	0.636
ページ番号	0.913	1.000
ヘッダ	0.875	1.000
フッタ	0.883	0.984
本文	0.975	0.989

3 データ作成方法

年次大会予稿集については ANLP-20 コーパスとして言語処理学会より OCR によるデジタルテキストが提供されているが、本論文では前述のプロジェクトにおいて構築したデジタルテキスト化手法 [5] を用いて、独自にテキストデータの作成を行なった。手法の概要は以下のとおりである。まず対象となる論文画像に対し OCR システムによる認識を行った。その結果認識されたテキストブロックに対し、機械学習手法を用いて属性分類を行った。対象とする属性は「論文タイトル」「著者」「ページ番号」「ヘッダ」「フッタ」「本文」の 6 種類である。機械学習手法は SVM を利用し、以下の情報を素性として利用した。

- タイトルページか否か
- ブロックの座標位置

- ブロックサイズ
- 平均文字サイズ
- 全形態素中の「名詞」の割合
- 全形態素中の「人名」の割合
- ブロック上下左右の余白の長さ
- 論文の発行年 (2010 年以前 or 以後)

全 4,082 論文のおよそ 5%の 239 論文について人手でブロック属性を付与した正解データを作成し、学習に利用した。正解データに対する 10-fold cross validation による精度を表 1 に示す。以降では「本文」として認識したブロック中のテキストを論文本文として利用し、書誌情報は年次大会プログラムの Web ページから取得したデータを利用する。なお本論文でのデジタル化処理は、学習用正解データの作成を除き直接的なデータ修正などには人手を介しておらず、全論文に対する処理はおよそ 1 日で完了した。

4 MIMA サーチによるマイニング

研究動向の分析のためには、論文集合の全体像の把握を可能にすることが必要である。特に単純な数値的集計だけではなく、内容や属性に基づく論文間の関連を抽出し、論文間の関係性を明示することも重要である。また、膨大の論文集合全体の把握には、一定の抽象化が必要となってくる。そして、これらの要件を満たす分析・可視化・操作が、ユーザ個人および任意の視点によりリアルタイムで行えることが重要である。

これらを実現するためのシステムとして MIMA サーチ [2] を利用し、これまでの年次大会予稿集のテキストを実装したサイトを構築した。MIMA サーチは、用語抽出をはじめとした自然言語処理、テキストマイニング、及び可視化の技術を統合したシステムであり、既に東京大学授業カタログ[1]や、工学部シラバス[2]の検索、可視化システムとして実用化されている。また、前述の「岩波書店『思想』の構造化プロジェクト」においても、論文誌『思想』における全体の俯瞰、知識の構造化のためのシステムとして利用されている。

MIMA サーチでは前処理として対象テキストから用語抽出エンジン TermEngine により用語を抽出している。TermEngine では、c-value 手法 [1] により用語をその用語らしさを表すスコアとともに抽出する。対象のデータに加え TermEngine により抽出された用語を用いて、MIMA サーチでは以下の機能を提供することが可能である。

[1]http://catalog.he.u-tokyo.ac.jp/
[2]http://mimasearch.t.u-tokyo.ac.jp/

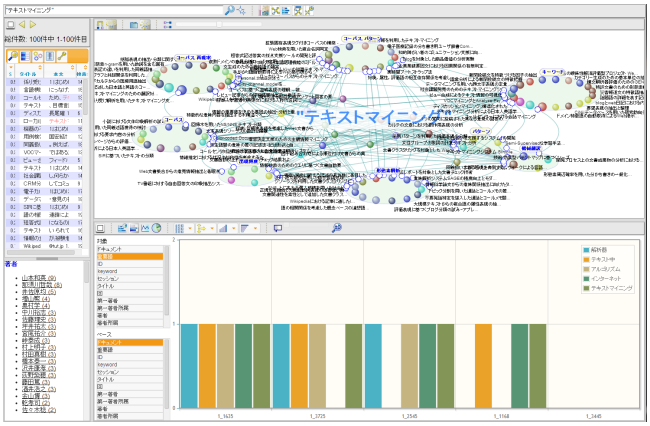

図 1: マイニングシステム画面例

- キーワードや年代等の属性指定による全文検索
- 検索された論文間の関連度の計算
- 上記により計算された関連度を基にした論文のクラスタリング
- 上記のクラスタリングの任意の抽象度での実行 (階層的クラスタリング)
- 検索結果に対するクロス集計

これらの分析結果に対し、ネットワーク表示による可視化、およびクロス集計に対してはグラフ表示による可視化を行うことができる。

図 1 にマイニングシステムの画面例を示す。通常の検索システムと同様に左上のキーワードボックスにキーワードを入力し、検索を行う。検索結果は左側にリスト表示、右側には論文をノードとした、論文間の関連度 (デフォルトでは用語スコアに基づいた関連度) に基づくネットワークが表示される。特に関連度が高い論文についてはクラスタリングが行われ、そのクラスタを代表する用語がクラスタのラベルとして付与されている。ネットワーク表示においてはノードをダブルクリックすることで論文の詳細を見ることができる。また、下部ではクロス集計を行うことができ、現在表示されている検索結果について、集計の対象・ベースの属性を選択し集計することができる。対象・ベースの属性はネットワーク表示にも反映され、対象を特徴量として計算した関連度を基にベースに対するネットワーク構造を表示する。ネットワーク表示、クロス集計表示はいずれかのみを表示することも可能である。また左下部にファセット (絞り込み) 検索用のフィールドがあり、著者、著者所属、年、セッション名などの属性でさらなる絞り込みが可能である。次節では実際にこれらの機能を利用した分析の例を示す。

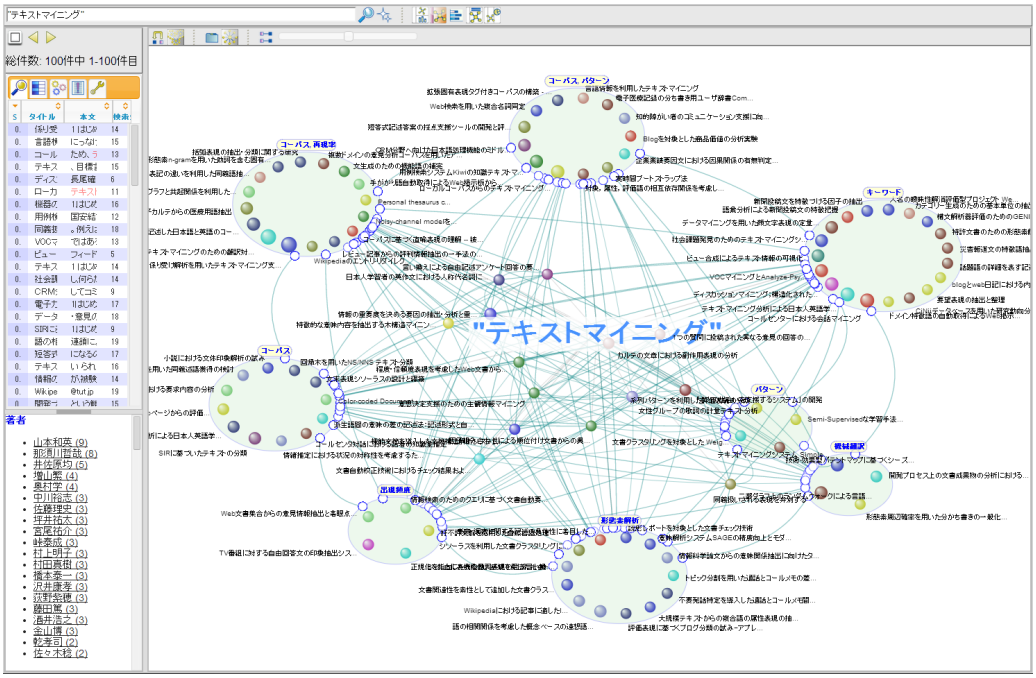

図 2: 「テキストマイニング」を含む論文のネットワーク表示

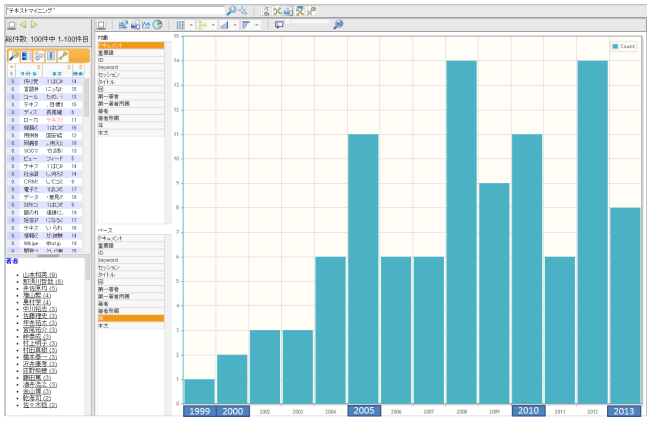

図 3: 「テキストマイニング」を含む論文数の推移

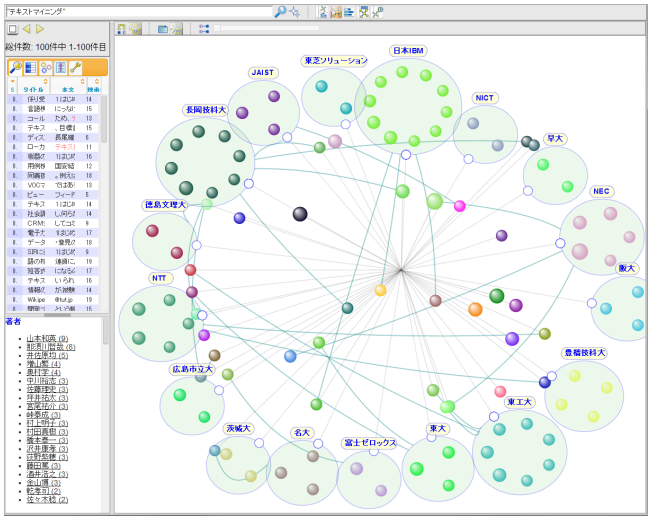

図 4: 「テキストマイニング」を含む論文の著者所属

5 システムを用いた分析例

分析例として、「テキストマイニング」を検索クエリとした場合の分析を示す[3]。システムの検索結果によると、これまでの言語処理学会年次大会では 100 件の「テキストマイニング」という用語を含む論文が発表されている。ネットワーク表示 (図 2) を見ることで、各論文間の関連や「形態素解析」「コーパス」など関連して頻出する用語が見て取れる。次に、クロス集計においてベースを年、対象をドキュメントとして年別の論文数の推移を見ると (図 3)、1999 年に最初の「テキストマイニング」を含む論文があり、以降多少の上下はありながらも徐々に件数が増えていることが分かる。また研究機関について分析するため、対象を第一著者の所属、ベースをドキュメントとしてネットワーク表示を見ると (図 4)、どの機関において「テキストマイニング」についての研究が行われているかを俯瞰できる。

[3]以降の分析はあくまで「『テキストマイニング』という用語を含む論文」を対象としており、必ずしも「主題が『テキストマイニング』である論文」ではないことに注意されたい。

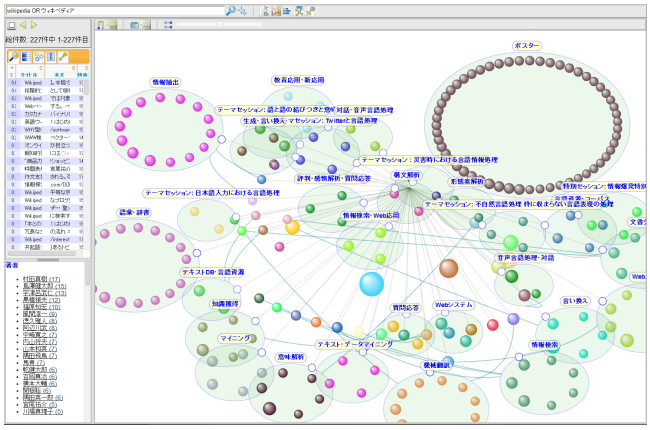

図 5: 「wikipedia」を含む論文のセッションによる分析

また別のトピックとして「wikipedia」をクエリとした場合の分析を示す。ここでは、論文が発表されたセッションをその論文の分野を表すラベルと考え、ベースをドキュメント (論文)、対象をセッションとしてネットワーク表示を行う。図 5 に示すとおり「wikipedia」に関する論文全体を分野ラベル付きで俯瞰することができ、例えば「情報抽出」「語彙・辞書」「機械翻訳」などの研究において利用されていることが分かる。

6 おわりに

本論文では、言語処理学会の研究動向分析を目的として、言語処理学会年次大会予稿集に対してデジタルテキスト化を行い、そのテキストを対象としたマイニングシステムを構築した。また、具体例としていくつかの分析例を提示し、様々な視点からの分析が可能であることを示した。

今後の課題としては、研究動向分析システムとしての用語抽出方法の検討があげられる。現在は用語抽出に c-value 法を用いているため、「コーパス」「モデル」のような用語らしさは高いが、言語処理の分野では一般的な頻出語が多く出現している。研究動向分析のためには、単純な用語ではなく既存研究 [4, 7] で行われているような「動向情報」としての抽出・スコア付けが必要であると考えられる。また、現在は全文を対象としており、用語の文脈を利用していないため、その用語が論文の主題であるかどうかを認識できていない。論文の主題として抽出するには簡単にはタイトルや要旨など論文中の主題を表す部分のみを対象とする方法が考えられる。また、当センターでは現在日本語の深い構文解析器 [3] を開発中であるため、その結果を利用し文脈情報を用語のスコア計算あるいは関連度計算に入れることも検討中である。また本論文のシステムは対象の分野などによらず適用可能であるので、他の学会誌などの動向調査にも適用していきたい。

参考文献

[1] Hideki Mima and Sophia Ananiadou. An application and e aluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. *Terminology*, Vol. 6, No. 2, pp. 175–194, 2001.

[2] Hideki Mima, Sophia Ananiadou, and Katsumori Matsushima. Terminology-based Knowledge Mining for New Knowledge Discovery. Vol. 5, No. 1, pp. 74–88, March 2006.

[3] 植松すみれ, 松崎拓也, 花岡洋輝, 宮尾祐介, 美馬秀樹. 統語・意味コーパスの統合と再解釈による大規模な日本語 CCG 文法の開発. 第 27 回人工知能学会全国大会発表論文集, 2013.

[4] 福田悟志, 難波英嗣, 竹澤寿幸, 武田英明, 相澤彰子, 大向一輝, 宮尾祐介, 内山清子. CiNii データベースを用いた研究動向分析システムの構築. 言語処理学会第 18 回年次大会発表論文集, 2012.

[5] 美馬秀樹, 丹治信, 増田勝也, 太田晋. 近代文献のデジタルアーカイブ化とテキストマイニング-岩波書店「思想」を題材に. 情報処理学会研究報告. 人文科学とコンピュータ研究会報告, Vol. 2012, No. 4, pp. 1–8, 2012.

[6] 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, 2005.

[7] 近藤友樹, 難波英嗣, 奥村学, 新森昭宏, 谷川英和, 鈴木泰山. 論文データベースからの研究動向情報の抽出. 言語処理学会第 13 回年次大会発表論文集, 2007.

[8] 本田由紀, 齋藤崇徳, 堤孝晃, 加藤真. 日本の教育社会学の方法・教育・アイデンティティ：制度的分析の試み. 東京大学大学院教育学研究科紀要, Vol. 52, pp. 87–116, 2012.