テキストマイニングツール Text Mining Studio による文献データ分析

～J-DreamIII文献情報データを例に～

岩本　圭介

1. はじめに

Text Mining Studio（TMS）は，テキストデータから有益な情報を抽出するためのテキストマイニングツールです。高精度な言語解析処理，長年数理科学に携わってきた当社ならではの信頼性の高い統計･マイニング処理，また言語解析からマイニング・統計処理の根幹，そして可視化機能まで自社で開発を行っていることによる製品サポートのきめ細かさによりユーザーの皆様の好評を得ております。2004 年の販売開始以来，コールセンターに蓄積された VOC の分析，不具合情報や営業日報などの社内文書の分析，またインタビューや調査データを扱うアカデミックな利用の分野など，多くの方面での適用を頂いております。その中でも近年，特許情報や文献データといった技術的なドキュメントの分析を通して自社と他社の技術の関係を把握し，その結果を技術開発における意思決定に役立てるといった動きが高まっており [1]，こういった場面での TMS の適用例が特に増加しております。

本稿では，株式会社ジー・サーチ様よりご提供を頂いた J-DreamIII 文献データの分析例を通じて，TMS の機能と特長について紹介します。

2. 分析対象データ

本稿では，キーワード「自動運転」により検索し得られた文献情報を TMS に取り込み，次の条件でスクリーニングしたものを分析対象としました。

- 日本で発行された，日本語の文献である。
- 日本語の抄録が存在している。
- 特に車両の自動運転への適用について調査する目的で，抄録文中に，「自動車」「車両」「乗用車」などの“車”を含むキーワードが少なくとも 1 回以上存在している。

その結果，1945 件の分析対象データを得ました。文献の発行年は 1980 年から 2016 年にわたっています。

3. TMS の機能と特長

3.1 日本語解析機能

TMS は，csv 形式もしくは Mirosoft Excel 形式といった，表形式で 1 レコードのデータが 1 件とみなせるようなデータを入力とします。今回の分析対象データの，列項目情報の一部を以下に示します。

- 和文／英文標題
- 資料名
- 著者名・所属機関
- 引用数・被引用数
- 発行年
- 抄録

データに存在する列項目のうち，どの列情報をテキストとして扱うかをデータのインポート時に指定します。テキストとした項目が言語解析の対象になります。今回は，和文表題と抄録を対象としました。テキストに指定しなかった項目は属性情報として扱われ，テキストを分析する際の切り口として利用することができます。

日本語解析処理の結果は，図 1 のような表形式で得られます。1 単語が 1 行の情報であり，「見出し語」の列が原文を文節単位に区切った（分かち書きした）もの，そして「原形」の列がその文節内の主要な部分を取り出したものです。特に技術分野別に辞書の整備を行わずとも“高度道路交通システム”“衝突防止”というような連語が自動的に抽出できる点に TMS の特色があります。技術的な分野においては，日々新しい概念が生まれそれに

＊いわもと　けいすけ　株式会社 NTT データ数理システム データマイニング部
〒160-0016　東京都新宿区信濃町 35 番地 信濃町煉瓦館 1 階
E-Mail: iwamoto@msi.co.jp
（原稿受領　2017.1.12）

行ID	文章ID	単語ID	見出し語	原形	置換語	品詞	品詞詳細	係り先	態度表現
1	1	1	高度道路交通システムにおいて	高度道路交通システム	高度道路交通システム	名詞	一般	3	なし
1	1	2	特に	特に	特に	副詞		3	なし
1	1	3	重要であるのは、	重要	重要	形容動	一般	10	なし
1	1	4	衝突防止を	衝突防止	衝突防止	名詞	一般	5	なし
1	1	5	はじめとする	はじめ	はじめ	名詞	副詞可能	6	なし
1	1	6	安全性に	安全性	安全性	名詞	一般	7	なし
1	1	7	関わる	関わる	関わる	動詞	一般	8	なし
1	1	8	技術である	技術	技術	名詞	一般	9	なし
1	1	9	点に	点	点	名詞	助数詞可能	10	なし
1	1	10	異論はないであろう。	異論	異論	名詞	一般	-1	否定

図 1　TMS の分かち書き結果

伴い新語や造語が登場しますが，TMS では自然言語解析処理をもってそれに対処し，ロバストな結果を得ることができます。また，単語の切れ目が望ましくない場合は，辞書機能を用いてカスタマイズすることができます。

図 1 のデータは，文法的情報の付されたキーワード情報であり，分析を進めるにあたってのいわば「生データ」そのものです。TMS はこの「生データ」自体をユーザーに提示することで，ツール全体のインプットからアウトプットまでの透過性を確保しています。

3.2　キーワード集計からの掘り下げ

キーワードの集計を行うことにより，分析対象データを概観するとともに，掘り下げの対象とすべき技術要素の当たりをつけることが可能です。

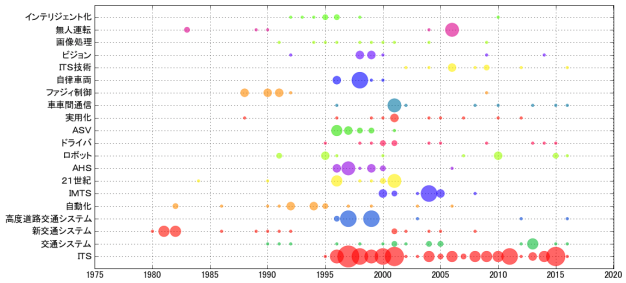

図 2　標題キーワードと発行年

図 2 は，文献の標題部分から抽出したキーワードと，文献発行年とのクロス集計結果を示しています。“ITS”“高度道路交通システム”というキーワードが 1990 年代後半から急激に出現するようになってきたこと，“ファジィ制御”は主に 1990 年前後に多く適用例が発表されていた技術であること，などを読み取ることができます。ここで，集計に先立って，“本稿”“本システム”“取り組み”といった論文データにおいて特有に頻出する単語群，またデータ取得の際の検索キーワードである“自動運転”を含むような単語群は，辞書機能を用いて除外しています。

図 2 で出現したキーワードから“車車間通信”に着目して掘り下げを試みます。TMS 利用時には，グラフ・表といったアウトプットから即座にマウスクリックで対応する原文を表示させ，そのキーワードがテキスト中のどのような文脈で出現しているかを確認することが可能です。

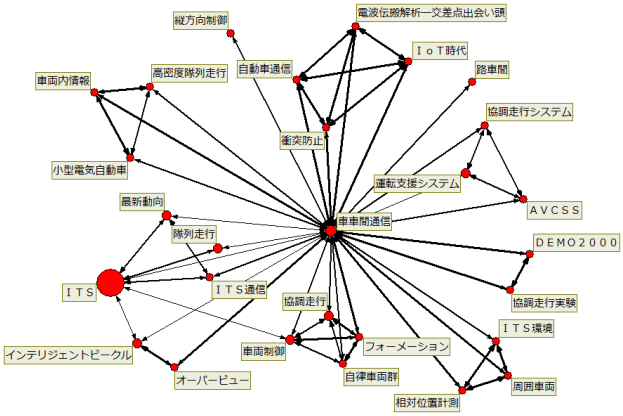

図 3　“車車間通信”との共起単語

また，注目したキーワードが他のどういったキーワードと同時に語られること確率が高いか，また件数が多いかといった単語間の共起情報に着目した図示を行い，背景事情を読み取るといったことも可能です。図 3 に“車車間通信”と他の単語との情報をネットワークグラフにより図示したものを示します。車両群の協調・隊列走行，また出会い頭での衝突防止といった利用目的に関わるキーワードが出現していることがわかります。

3.3　属性情報の利用

データに付随する属性情報を，分析の切り口として用いることが可能です。ここでは文献著者の所属機関情報を利用して，特に文献数上位の機関に着目することとします。それぞれの機関について，どういったキーワード

が特徴的に出現しているか，その偏りを評価して特徴度合いに応じたランキングを作成することができます。

順位	機関-A大学	機関-E大学	機関-I大学
1	交通流	看護師	ゴムタイヤ方式
2	軌道	病院内	新交通車両
3	スライディング制御	医師	高架専用軌道上
4	獲得手法	改善ニーズ	操縦動作
5	車速制御アルゴリズム	エコデザイン	地震時
6	車両追従制御	コ	地震大国
7	横方向制御	モビリティ社会	都市内中量輸送機関
8	制御方法	医療施設内	パワーステアリング装置
9	横方向	患者	マルチボディダイナミクス
10	音声	患者搬送	案内レール

図4　A大学・E大学・I大学の特徴語

文献数1〜10位の機関をA〜Jとし，TMSのアウトプットから，1，5，9位の機関であるA大学・E大学・I大学について，抄録のテキストから抽出された特徴語10語ずつを抜き出したものを図4に示します。どのような分野に車両の自動運転への適用を考えているかがはっきりと見て取れる結果になっています。A大学では交通流や他車両への追従を考慮した車両の制御，E大学では看護師や医師の負荷を軽減するための運搬車両，I大学では新交通システムへの適用に関する文献が発表されているといえます。

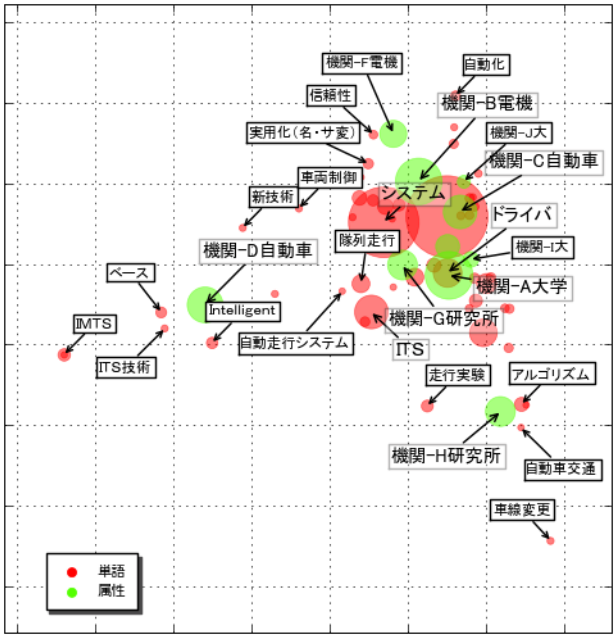

図5　対応分析によるキーワードと機関の関係

また，多変量解析の一手法である対応分析を用いて，属性値（ここでは機関）間の類似関係をマップすることができます。図5に，抄録のテキストから抽出されたキーワードと，機関間の関係を示しました。図5のバブルのそれぞれは1つのキーワードもしくは1つの機関に対応しており，バブルの大きさはキーワード・機関の出現件数に対応しています。キーワードの出現傾向が近い機関同士，及び同じ機関に出現する傾向が高い単語同士が近くにマップされ，さらに機関に対して関連の強いキーワードが互いに近くにマップされるというようになっています。

図5で外れた位置にあるD自動車及びH研究所は，特に他の機関とは異なったキーワードの出現傾向にあることがわかります。H研究所は，“走行実験”“アルゴリズム”“車線変更”といったキーワードで特徴付けられているといえます。こういったアウトプットを基に，キーワードを介した自社と他社との関係や類似性を読み取ることができます。

TMSでは，データに存在する属性情報は自由に分析の切り口として用いることができます。所属機関ではなく，著者，発行国，発行年などを用いて同様の分析が容易に行えます。

4. アドオンモジュール

TMSには，ツール本体とあわせて用いることでより発展的なご利用が可能になる「TextCutter」と「英語アドオン」というアドオン製品があります。後者「英語アドオン」は，日本語テキストを分析する際の操作感そのままに，TMSで英文の分析が可能になる製品です。本章では，前者「TextCutter」の利用例を解説いたします。

4.1　TextCutterの利用

TextCutterは，テキストを話題毎に分割し，「トピック」として話題のラベルを付与する，テキストマイニングの前処理ツールです。1件のテキスト内で様々な話題が語られているような場合，それらを適切に分離してトピック情報を付与すれば，トピック間の比較や特定のト

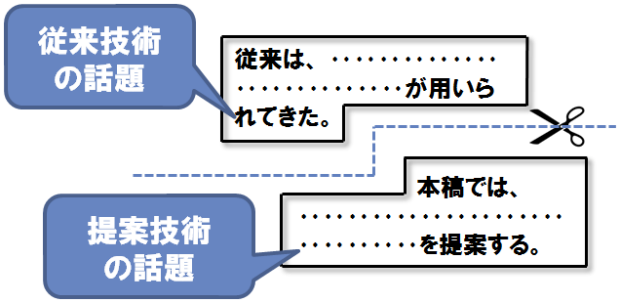

図6　TextCutter概念図

順位	従来技術	提案技術
1	自動車	有効性
2	自動化	モデル
3	システム	実験結果
4	問題点	アルゴリズム
5	ニーズ	試験結果
6	信頼性	実車実験
7	交通事故	制御アルゴリズム
8	プロジェクト	有用性
9	大型化	システム概要
10	研究開発	フレームワーク
11	コンピュータ	ドライビングシミュレータ
12	開発中	シミュレーション結果
13	ヨーロッパ	位置情報
14	ITS	適用例
15	高速道路	パラメータ

図 7　従来技術と提案技術の特徴語

ピックの絞り込みなどが可能になり，TMS でより踏み込んだ分析が可能になります。

TextCutter の概念図を図 6 に示します。論文の抄録においては，「従来は～」などの表現によって語られる従来技術や問題提起を表す部分と，「本稿では～」といった形で語られる提案手法の部分に分かれることが一般的であるため，TextCutter を用いて，問題点となる事項と解決方法となる事項を抽出することを試みます。

図 4 と同様の手法で，「従来技術」「提案技術」それぞれの特徴キーワードを抽出したものを図 7 に示します。大局的な結果ではありますが，「従来技術」の方はやはり“信頼性”“交通事故”が重大な関心事であることが裏付けられ，「提案技術」においては“実車実験”“ドライビングシミュレータ”などによって実際に“有効性”を検証していることがわかります。

5. おわりに

本稿では，文献データの分析例を通じて，TMS の機能の一部を紹介させていただきました。各種分析機能におけるカスタマイズ性にも大きな利点があり，実務において「痒い所に手が届くツール」であると自負しております。是非 TMS がお客様にとって役立てる道具となれますよう，導入サポートや分析コンサルティングも開催させていただいており，好評を得ております。ご興味を頂いた方は，是非当社無料の体験セミナーに足をお運びください。開催要項はサイト http://msi.co.jp/tmstudio/seminarRegular.html にてご確認いただけます。

最後に，分析対象データをご提供下さいました株式会社ジー・サーチ様に多大な感謝を申し上げます。

註・参考文献

1)　豊田裕貴，菰田文男．特許情報のテキストマイニング，ミネルヴァ書房，2011.

Series : Information Analysis Tools, 15: Literature Data Analysis by Text Mining Tool: Text Mining Studio －Using J-DreamIII Document Information Data as Example. Keisuke Iwamoto (NTT Data Mathematical Systems Inc. Data Mining Division, 1F Shinanomachi Rengakan, 35, Shinanomachi, Shinjuku-ku, Tokyo 160-0016, JAPAN)