２Ｉ０３

ＩＴ業界のコンセプトトレンドの分析手法

○片岡利枝子，神田陽治，内平直志，井川康夫（北陸先端科学技術大学院大学）

第1章　はじめに

インターネットの普及により、我々が日々扱う文書データは通常の知識情報に加え、論文や特許のような技術情報、メールやツイッターで個人が発信する感情や意見など、多種多様な形で増え続けている。その結果、せっかくの文書データが十分に活用できていない懸念が生じている。

しかしながら、菰田(2013)[1]によると、膨大な文書データのすべてに目を通すことは不可能であるが、テキストマイニングを使いこなすことによって、その中から有効な情報を見いだし活用することが可能になる。上田(2008)[2]は、テキストマイニングはデータマイニングと異なり、文書データ中に記述されている内容を、その言語表現から分析するものであると定義する。

テキストマイニングは通常のアプリケーションソフトウエアと異なり、単に文書データを入力すれば有効な結果が得られるものではない。那須川(2006)[3]が述べるように、どのように分析し、その出力特徴をどう読みとるかによって結果が大きく変わってくる。

これらの特徴を踏まえ本研究では、膨大な文書データからテキストマイニングによって、“話題の経時的な変化”という、時間軸を含んだ知見を有効に抽出する手法を提示することを目的とする。

話題は時代を反映し、その背景と共に存在する。このため、変化を語る際には、その前後関係や絶対時期を正確にとらえる手法が求められる。単に話題の時間的移り変わりを収得するのみならず、話題間の関係性も同時に把握することによって、変化を推移として把握することができる。

通常、テキストマイニングのアプリケーションプログラムは時間の検出も可能であるが、有効な結論を導くためには目的に合わせた活用方法が求められる。本論文では、IT 業界の文書データからコンセプトトレンドの推移を収得する手法について提案する。

第２章　先行文献のレビュー

2.1　テキストマイニングの分析方法について

テキストマイニングを駆使して文書データやインターネット記事を解析した例は多数存在するが、本研究では、話題の変化（トレンド）の分析に注目した論文を中心に調査を行った。

脇森（2013）[4]は、話題の変化を文書データで語られている単語の経時的な変化ととらえ、単語の出現する頻度の増減から消費者トレンドの変化を検知する試みを行っている。さまざまな単語の出現件数を時間軸と共に記述し、出現頻度の変化の大きい単語に着目すれば話題が変化したことを察知できるとしている。

本文献ではトレンドが動いたと判断するための単語出現の増減を検知する「感度基準」が議論されているが、トレンドを代表する単語間の関係性までは問われていない。

白井（2009）[5]らは、文書データからトレンド情報を抽出のためには、重要なキーワードがあ

らかじめ抽出されていることが必要であるとし、キーワードの抽出方法を目的に応じて選択することによってマイニング環境の整備を行っている。テキストの中から名詞と認定された候補から専門用語を抽出する形態素解析による抽出手法、および、特定の言い回しが頻繁に用いられるようなテキストや、特定のフォーマットが定義されたテキストを対象にする場合に有効な、パターンマッチングによる抽出がある。

奥和田（2009）[6] らは、キーワード抽出において、前出の形態素解析を用いて実施し、さらにそれらを分野別特徴によって分類することによって、自由記述の文書データから自動的に効率よく目的とする結果を収得する手法を示している。

山本（2009）[7] は、特許や論文の文書データにテキストマイニングを駆使することによって、特定の技術分野に存在する単語を単語群としてグループ分けを行い、ひとつの図上にマップした。それぞれの単語群に出願年を明記することで、年代ごとの単語群をとらえているが、図の座標軸には意味はない。ある時代を代表する単語は明らかになるが、単語群と単語群との間の相関関係を表すものではなく、従って、単語群を超えた単語と単語についても同様である。

この他、テキストマイニングによって文書データから単語を抽出し、それら関係性をマップするための手法は多数みられるが、いずれも単語の経年変化を追跡する形には至っていない。

本提案方式は、座標軸に時間的な意味を持たせて単語間の相関をとらえながら単語の推移を捉えるものである。すなわち、表出する単語は、ある時期に単独で出現したのではなく、基本的に前後の単語と強い相関を持つ場合のみマップされる。

2.2　テキストマイニングを使う技術について

一方、テキストマイニングを使う技術についての文献は少ない。第一章で述べたように、テキストマイニングは通常のアプリケーションソフトウエアと異なり、単に文書データを入力すれば有効な結果が得られるものではない。

那須川[3]は、テキストマイニングによって有効な結果を導くためには，それを使いこなす技量も重要な要素であるとして、過去の経験から得た知見を基に、分析を行う際のプロセスを「トライアルフェーズ」、「本格化フェーズ」、「結果の活用フェーズ」の３段階にまとめて提示している。

ただし、この手順はあくまで指針を述べたものであるから、実際に分析を進める上では、個々の詳細な作業内容はそれぞれの事象によって検討が必要である。

戦略策定にテキストマイニングを活用する技術も紹介しておきたい。Bose (2008) [8] は、テキストマイニングを使うことによってコンセプトの連関（Linkage）情報を取得し、企業の業界における戦略策定の分析に利用するための手法を述べている。Bose はテキストマイニングの分析結果を得ることが目的ではなく、ここで取得した情報をどう役立てるかが重要であると述べ、戦略策定を行うまでの全体的なプロセスを提示している。

テキストマイニングで得られた結果を有効に活用するための手法を具体的に述べたものであり、また、時間的な概念は含まれていない。

本研究では、那須川の提示するフレームワークをもとに対比を行いながら、業界雑誌などの文書データから業界のコンセプトトレンドを分析するための具体的な手法を提示する。アプリケーションツールとしては、2012 年発売の IBM Content Analytics Version 3.0（以下，ICA3.0）を採用した。

第３章　分析手法の提案

3.1　トライアルフェーズ

トライアルフェーズでは、「データを使って何を実現したいのか？」という目的設定と並行して、対象データの全体像を把握する。そしてある程度の方向性が見えてきたら、本格化フェーズに向けて、分析の目的に応じたマイニング環境を整備するものである。

今回の分析では全体の傾向を掴むために、テキストマイニングのアプリケーションが搭載しているさまざまな出力形式を駆使して、文書をひとつひとつ読み込みながら、ファセット項目（名詞、動詞・・などの種類）や時系列メモリを切り替えることによって分布を確認することとした。ICA3.0 の Redbook[9]より、ツールの機能一覧を図１に示す。

図１．ICA3.0 の機能一覧

本提案方式では、ファセット分析を行った際に文意の把握に役立たない一般語が多数存在する点に着目した。ファセット分析は選択した分類項目の用語について、用語間の相関係数が自動的に数値として検出されるものである。従って、一般名詞を選択した場合には、頻繁に登場する「部長」「会議」・・などの一般的なビジネス用語もキーワードとして集計されてしまう。これらの単語をマップ上に表示しても不要な情報となるため除外する必要がある。このため、次フェーズで本格的な分析を行うための前処理として、分析の中心とするキーワードの抽出が必須であるとの結論に至った。

文書データの中から抽出する手法については、先行文献の中にもいくつかの手法が紹介されている。しかし、産業界に特化した文書データであるような場合には、あらかじめキーワードとして選定されているコンセプト用語を採用する方が、トレンドについて精度の高い分析結果が得られると判断する。例えば、IT 業界であれば、(1) 技術用語として、きちんと説明しておく必要があるもの、(2) 技術トレンドとして、押さえておくべきもの、(3) いま、世の中で大事な言葉として提案するもの、として定義されているキーワードである。

3.2　本格化フェーズ

本格化フェーズは、目標設定がほぼ完了して本格的な稼働に入る段階である。当然、前処理やマイニングを行うための中心となるキーワードの選定も完了している段階である。

本提案方式では具体的に、トライアルフェーズで抽出したキーワードを中心に、それらの偏りや変化を正確に検出する作業を行った。抽出した全キーワードについてテキストマイニングのアプリケーションの偏差分析機能（時系列偏差）を使って各キーワードの偏りや変化について、そしてファセット分析機能を使ってキーワード間の相関についての詳細な確認を行うこととした。分析画面の一例を図２、図３に掲載する。

菰田（2014）[1] らによると、テキストマイニングでは、あるデータが何件存在するかということではなく、その増減や分布の偏りの意味するところを読み取ることが重要である。すなわちテキストマイニングの価値は、基本的には比較による特徴の検出にある。

図２．偏差分析例（クラウドコンピューティング）

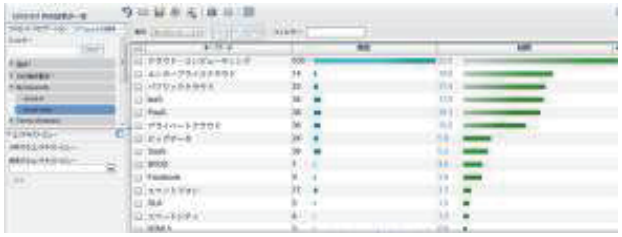

図３．ファセット分析例（クラウドコンピューティング）

3.3　結果の活用フェーズ

本フェーズは、本格化フェーズで明らかになった特徴を踏まえて、活用シナリオを検討する段階である。

筆者らは、本格化フェーズで得られた分析結果から、業界の文書データにおいては、キーワードは一時期に同時に現れたのではなく数年の時間を要して時系列に関係構築されながら出現していったと判断した。従って結果の活用シナリオとして、前フェーズのファセット分析で収得したキーワード間の相関値と、偏差分析で明らかになった時間軸のデータを基にキーワード間を結ぶことによって、業界のトレンドの推移が分かる図式として表出させることを試みた。

ICA3.0 では、ファセット分析にて数値化されたキーワードの相関関係はコネクション分析として出力することができる。本提案方式では、これに偏差分析で収得した時間軸データを加えてプロットすることにより、キーワード間のトレンドについて図４のようなマップを行った。楕円の大きさは、そのキーワードが登場した頻度に応じて描くことが可能である。また、単語間を結ぶ線に、太さの違いを持たせることで相関の強さを表現することもできる。相関の強い線をたどっていくことによって推移がわかる。

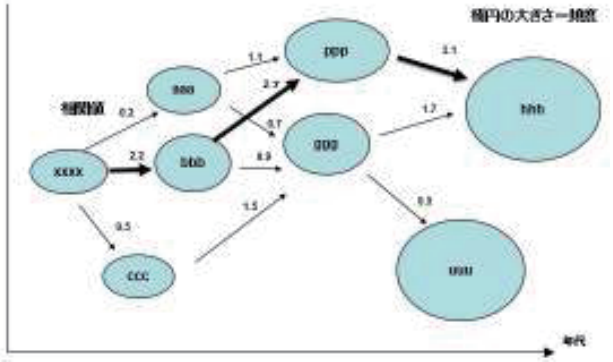

図４．トレンド推移の出力イメージ

第４章　考察

本提案方式では、業界のトレンドをより精度よく捉えるために、あらかじめ、その業界で何らかの意味があるとされるコンセプト用語を、テキストマイニング分析を行う際のキーワードとして抽出することを提案した。そして、テキストマイニングの分析結果として得られるキーワード間の相関値に時系列の概念を導入することによって推移としてマップする手法を述べた。

テキストマイニングによって、ある事象ついて何らかの知見が得られた場合、そこからどのような結論を導くかが重要である。従って、第３章で導き出した業界トレンドの推移を、その周辺に付随するキーワードの発生と共に時間軸で抽出したものは、さらに、時代を背景にしたさまざまな状況やビジネス環境などの要素を加味してその意味を捉える必要がある。このため、テキストマイニングで得られた知見を、その目的に応じて、実際のさまざまな状況に照らし合わせながら検証していく仕組み作りや、手法の検討しておくことが求められであろう。

第５章　まとめ

テキストマイニングは膨大な文書データを分

析することによって、本来、手動分析では容易に見えなかった知見を得ることを理想とする分析技術である。今回はテキストマイニングを有効活用するために、その分析手法に焦点をあてた。

まず、トライアルフェーズにて特定キーワードの抽出の必要性を認識し、より分析の精度を上げるためのキーワード抽出方法を採用した。そして最後の活用フェーズの段階で、キーワード間の相関値に時間的概念を加味してマップさせることによって、業界トレンドの推移として表出する手法を述べた。相関の強さを線の太さなどで表し、その軌跡を追跡いくことで、ある出発点のキーワードがどのように推移していったのかを把握することができる。

筆者らは、実際に本提案手法を用いて IT 業界の業界誌の分析を行い、クラウドコンピューティングに関するコンセプトの発展段階についてモデルの生成を行なっている[10]。

テキストマイニングの分析結果から得られる情報を適切にとらえ、現実の事象に何らかの活用を行うことによって成果を有効に活用したい。

主な参考文献

［1］ 菰田文男, 那須川哲哉, 技術戦略としてのテキストマイニング, 中央経済社（2014）。

［2］ 上田太一郎, 事例で学ぶテキストマイニング, 共立出版（2008）。

［3］ 那須川哲哉, テキストマイニングを使う技術／作る技術, 東京電機大学出版局（2006）。

［4］ 脇森浩志, ビッグデータに対するテキストマイニング技術とその適用例, Unisys Technology Review 第 115 号（2013）。

［5］ 白井康之, 小関悠, 小池亜弥, テキストマイニングによるトレンド情報抽出環境の構築, MRI 技術レポート 5. 14, p110-123 (2009)。

［6］ 奥和田久美, 白井康之, 小関悠, 分野別の自由記述から科学技術政策上意味ある意見を自動抽出する試み, 研究・技術計画学会 年次学術大会講演要旨集, Vol.22, 69205 (2007)。

［7］ 山本外茂男,「産業連携のマッチング性分析におけるテキストマイニングの有効性」, 情報の科学と技術, 59 巻 6 号（2009）。

［8］ Bose, R. Competitive intelligence process and tools for intelligence analysis, Industrial Management & Data Systems Vol.108 No.4, 2008.

［9］ Introducing OmniFind Analytics Edition, Customizing Text Analytics An IBM Redbooks publication, Developer Works, IBM Corp. (2012).

［10］ 片岡利枝子, 井川康夫, 内平直志, テクノロジーコンセプトのサービスコンセプトへの進化プロセス―クラウドコンピューティングの事例研究, 研究・技術計画学会 第 28 回年次学術大会講演予稿集, 2G19 (2013)。