2Q-1

潜在的意味解析を用いた出版物データの分析
—エネルギー問題に対する人々の意識変化について—

福田 康平† 松尾 和洋‡

金沢工業大学 工学研究科情報工学専攻† 金沢工業大学 工学部情報工学科‡

1. はじめに

インターネット通販サイトにおける商品や売り上げデータは，「人々が何を求めているか」を直接反映した情報である．そのため，人々の関心やその傾向を分析するための有用な情報源として注目されている．しかし，通販サイトが持つ商品データのような巨大なデータ群から必要な情報の意味的特徴を的確に抽出し，人々の意識やその変化を包括的に分析することは容易ではない．

本研究では，「潜在的意味解析」と呼ばれる自然言語処理技法を用いることで，インターネット通販サイトの商品情報から人々の意識変化の分析が可能か，その有効性を検討する．

2. 分析対象

本や雑誌などの出版物は，インターネット通販サイトが扱う商品の中でも，人々の関心をより反映していると考えられる．そこで本研究では，出版物の内容とその傾向を調べることにより，社会問題に対する人々の意識変化を明らかにする．

分析対象とする社会問題としては，近年関心が高まっている東日本大震災と，それに付随するエネルギー問題を採り上げる．出版物データはインターネット通販サイト Amazon の Web サイト[1]から取得する．

3. 分析の実施

分析を行うために，Amazon から「東日本大震災」や「エネルギー問題」に関連する出版物のデータを収集し，出版年月毎に切り分ける．次に，切り分けた出版物のタイトルおよび内容をコーパスとして潜在的意味解析を行う．その結果から，出版年月毎に出版物の主題や主張の意味的特徴を抽出し，人々の意識変化を明らかにする．

Analysis of publication trends using latent semantic analysis
- Changes in people's attitude toward energy problems -

† Kohei FUKUTA・Kanazawa Institute of Technology, Graduate School of Engineering, Information and Computer Engineering

‡ Kazuhiro MATSUO・Kanazawa Institute of Technology, Department of Information and Computer Science

以下に，具体的な分析方法を述べる．

3.1 データの収集

本研究で対象とする出版物は，洋書を除く書籍および雑誌である．Amazon から収集するデータ項目は，①ASIN（Amazon の商品 ID），②タイトル，③出版年月，④内容（商品説明），⑤商品カテゴリの 5 項目である．

出版物データの収集には，Amazon が開発者向けに提供している商品情報アクセス用の API である Product Advertising API を使用する．収集のための検索キーワードは，「東日本大震災」および「原発」である．しかし，使用する API の制限上，キーワード検索だけでは収集できる冊数に限りがある．そのため，Amazon の商品レコメンデーション機能を利用し，収集した出版物の関連商品も収集する．本分析では，この方法によって 2197 冊の出版物データを収集した．

3.2 データの整形

収集した出版物データを出版年月毎に切り分ける．切り分け方としては，出版年月が 2011 年 1 月から 2012 年 12 月までの出版物は 1 ヶ月毎に切り分け，それ以前の出版物は 1 つにまとめた．2011 年 1 月以降の出版年月毎の収集冊数を図 1 に示す．

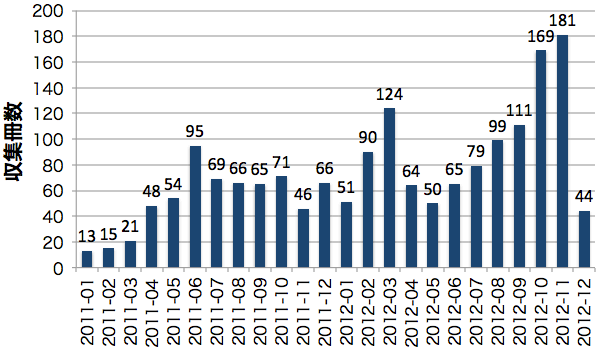

図 1 出版年月毎の収集冊数

次に，切り分けた出版物データのタイトルおよび内容（商品説明）のテキストをそれぞれ結合し，25 個のコーパスを生成する．そして，これらのコーパスから文書-単語の共起行列（文書-

単語行列）を生成する．生成した文書-単語行列の概要を図２に示す．

	単語1	単語2	単語3	…
~2010-12-01	1	0	0	…
2011-01-01	0	2	1	…
2011-02-01	3	1	0	…
⋮	⋮	⋮	⋮	…
2012-11-01	0	0	2	…

図２　文書-単語行列の概要

文書-単語行列の生成のための形態素解析には，形態素解析エンジン MeCab[2]を用いた．その際，ノイズとなる単語を除去するために以下の条件を満たす単語を分析対象から除外した．

①品詞が動詞または形容詞以外の単語
②全コーパス中の出現回数が２回以下の単語
③すべて平仮名で２文字以下の単語
④すべて小文字アルファベットの単語

3.3 潜在的意味解析による分析

潜在的意味解析（Latent Semantic Analysis：以下 LSA）は，ベクトル空間モデルを利用してコーパスを統計的に処理することで，単語の文脈上の意味構造を抽出・表現する自然言語処理手法である[3].

LSA は，文書-単語行列を特異値分解することで，人手やシソーラスなどの事前知識を用いることなく，単語間あるいは文書間の相関関係を明らかにできる手法として知られている．そのため，重要な意味を持つ単語に占める新語の割合が大きいと想定される社会・時事問題の分析においては，より実態に即した分析結果が期待できる．また，分析の客観性についてもある程度確保されると考えられる．

ここでは，3.2 節で生成した文書-単語行列を LSA で分析することにより，出版年月毎に単語の得点を求めた．この得点をもとに出版傾向とその素となる人々の意識を分析する．

4. 分析結果と考察

LSA による分析結果からは，東日本大震災前と震災後でコーパスの意味的特徴に大きな変化が観察できる．例えば，「放射能」や「核燃料サイクル」などの単語は震災後に得点順位が上がっている．しかし，ほとんどのエネルギー問題関係の単語では，筆者が予想していた程の大きな変化は見られなかった．分析結果の例として，2011 年 1 月（震災前）と 2012 年 1 月（震災後）の単語の得点の一部を表 1 に示す．

表１　単語の得点

	出版年月：2011年1月（震災前）		出版年月：2012年1月（震災後）	
順位	単語	得点	単語	得点
1	陰謀論	0.1606588	祈り	0.0155251
2	説	0.0975301	章	0.0084184
3	謀略	0.0765344	予知	0.0077881
4	ノーベル賞	0.0642347	ソーシャルメディア	0.0069060
5	パーセント	0.0642333	確率	0.0065961
6	陰謀	0.0321328	東日本大震災	0.0065961
7	川	0.0321306	事故	0.0057989
8	刺激	0.0321222	子ども	0.0050053
9	判定	0.0321130	薬剤師	0.0049941
10	事例	0.0309469	炉	0.0047115
11	米国	0.0277875	被災地	0.0043436
12	地球温暖化	0.0277676	東北	0.0043389
13	関与	0.0277472	放射能	0.0043368
14	世紀	0.0277472	原発	0.0038720
15	物理学者	0.0243859	十	0.0037884

得点のリストのみでは，出版物の傾向を包括的に分析することが難しい．したがって，現在，単語をクラスタリングすることで，単語間の関係や出版年月毎の推移の分析を進めている．

5. 今後の課題と展望

データ収集の際，API の制限などが原因で十分な量と精度の出版物データの確保が困難だったことが，分析結果に悪影響を及ぼしている可能性がある．API を使わずに，売上げ上位の商品を出版年月毎に収集するなどして，コーパスの品質を改善する必要がある．

また，近年，LSA よりも精度が高いとされている確率的潜在意味解析（PLSA）が提案されている．PLSA を用いた場合の出版物データの分析結果との違いも考察していきたい．

6. おわりに

本研究では，Amazon から収集した出版物データをコーパスとして潜在的意味解析を行い，その結果から社会問題に対する人々の意識変化を分析した．コーパスの品質など，本研究におけるいくつかの課題を改善することで，より有用な結果が得られると考えられる．

参考文献

[1] “Amazon.co.jp”, http://www.amazon.co.jp
[2] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[3] Landauer, T.K., Foltz, P.W., & Laham, D. “An Introduction to Latent Semantic Analysis”, Discourse Processes, 25, pp.259-284 (1998).
[4] 高田明典『潜在的意味解析の原理と数理―女児向けコミック雑誌の意味構造の変遷を題材として―』，フェリス女学院大学文学部多文化・共生コミュニケーション論叢, 5, pp49-62 (2010).