3A3-4

専門用語の用法に関する計量化指標の時系列パタン分析

Analyzing Temporal Patterns of Importance Indices Related to Technical Term Usages in Research Documents

阿部 秀尚[*1]　　津本 周作[*1]
Hidenao ABE　　Shusaku TSUMOTO

[*1]島根大学
Shumane University

In this paper, we describe a method for detecting temporal patterns of technical term usages based on importance indices and clustering methods. In text mining, importance indices of terms such as simple frequency, document frequency including the terms, and tf-idf of the terms, play a key role for finding valuable patterns in documents. Although the purposes of each set of documents are not changed, roles of terms and the relationship among them in the documents change temporally. In order to detect such temporal changes, we developed a method to extract temporal patterns as clusters of importance indices of technical terms. Empirical results show that our method has availability for assigning abstracted sense of technical terms by considering their temporal usages based on the linear trends of the temporal clusters.

1. はじめに

近年，各分野における情報システムの普及に伴い，電子的に蓄積される文書が日々増加している．これらのテキストデータから有用な知見を獲得するため，種々のテキストマイニング手法が開発されてきた．特に，定期的に発行される刊行物をはじめ非定時的な文章群である電子掲示板や検索サイトに於けるキーワードを対象として，新興の単語や複合語の検出が世論の動向を捉える方法として注目されている [Swan 00]．しかしながら，従来の新興単語の傾向検出手法 (ETD: Emerging Trend Detection)[Lent 97, Kontostathis 03] では，対象が単語のみ，あるいは個別の指標の傾向や確率的状態遷移 [Kleinberg 03, Mei 05] を扱っている．このため，単語やフレーズの指標について，時間経過に対する連続的な類似性を議論することが困難である．

これに対し，我々は，先行研究 [阿部 09] において，辞書に依らない用語の抽出，単語やフレーズに対する重要度指標，時系列方向の線形傾向によるフレーズの傾向抽出手法を提案した．ここでいう重要度指標とは，用語やその構成要素である単語の出現頻度を基に算出される指標であり，本稿ではこれを用語の計量化指標と呼ぶ．

本稿では，傾向検出を改善するため，抽出された用語の各時点での計量化指標について系列を作成し，系列間の類似性に基づいた時系列クラスタリングを導入した．用語の傾向は，各用語が属する時系列パタンである各クラスタの代表元の線形傾向として割り当てる．実例を用いた実験では，人工知能学会全国大会の 2003 年から 2009 年にかけての発表題目（タイトル）について，本手法を適用する．本実験の結果について，用語の基本統計量である出現文書数と 2 種類の計量化指標による傾向を比較する．

2. 計量化指標に基づく

本節では，以下の処理を統合した文書中の用語に関する傾向分析手法について述べる．

連絡先: 阿部 秀尚，島根大学，〒 693-8501 島根県出雲市塩冶町 89-1，電話番号 (0853)20-2174，FAX(0853)20-2170，abe@med.shimane-u.ac.jp

1. 辞書に依らない文書群からの用語の抽出
2. フレーズあるいは単語の重要度指標
3. 時系列パタンの生成
4. 時系列パタンの傾向割り当てによる用語用法の傾向分析

本手法では，まず，全時点の文書群あるいは一部の文書群を対象として，用語を抽出する．次に，各用語について，時点毎の文書群における計量化指標の値を算出する．この結果，各用語を行，各時点の計量化指標の値を列とするデータセットを作成される．データセットの生成は用意した計量化指標の数の分だけ繰り返される．生成されたデータセットに対し，各用語の時系列について類似度を算出し，時系列パタンを生成する．生成された時系列パタンについて，時間方向の傾向に意味[*1]を割り当てて各時系列パタンを解釈する．

手法の概観を図 1 に示す．

2.1 用語の抽出

まず，本手法では，辞書に依らない用語の抽出手法を用いるが，これは新興の単語や概念が既存の辞書とのマッチングでは得られないことを防ぐためである．また，これら新興の概念は，新規の単語の組み合わせや全く新たな単語として現れることが多い．このような用語抽出手法として，今回は中川らによる用語抽出手法 [Nakagawa 00] を用いた．この手法では，抽出候補となる単語数 $L \geq 1$ の複合名詞 CN について，スコア $FLR(CN)$ を算出して，ユーザが与える閾値を越えた複合名詞を用語として抽出する手法である．ここで，$FL(N_i)$ は単語 N_i の左に異なる語が出現する頻度，$FR(N_i)$ は N_i の右に異なる語が出現する頻度を表している．

$$FLR(CN) = f(CN) \times (\prod_{i=1}^{L}(FL(N_i)+1)(FR(N_i)+1))^{\frac{1}{2L}}$$

本手法のほかに，用語抽出の手法としては χ^2 統計量に基づく隣接共起単語抽出 [Matsuo 04] など，他の手法も同様に適用可能である．

*1　ここでいう時系列パタンの意味とは，対象とする期間における用語用法の類似性に割り当てる意味ラベルである．

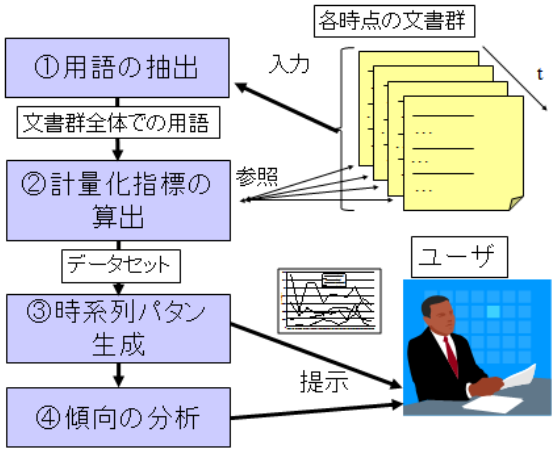

図 1: 用語の計量化指標に基づく時系列パタン分析手法の概観

2.2 用語の計量化指標の算出

用語抽出手法によって得られた用語について，各時点の文書群における計量化指標の値を算出する．用語やその用法は，用語自体の出現頻度や構成要素の単語と用語の出現頻度の差異として計量化される．これらを総称して計量化指標と呼び，用語に関する計量化指標の変化では，タグクラウドをはじめとする用語の出現頻度の変化が注目される．用語 $term$ のある一定期間の文書 D_{period} 中の出現頻度としては，用語自体の出現頻度 $TF(term, D_{period})$ や用語を含む文書数 $DF(term, D_{period})$ がある．

本稿においては，用語自体の出現頻度および用語が含まれる文書数から算出可能な次の 2 指標を定義し，計量化指標として扱う．

テキストマイニングにおいて単語（あるいはフレーズ）の重要度として広く用いられる tf-idf[Sparck Jones 88] は，各用語 $term$ のある一定期間での文書 D_{period} に対する tf-idf 値 $TFIDF(term, D_{period})$ として以下のように計算される．

$$TFIDF\ (\ term, D_{period}) = TF\ (\ term, D_{period}) \times log_e \frac{|D|}{DF(term, D_{period})}$$

ここで，$TF(term)$ は，サイズ $|D|$ の文書における $term$ の出現頻度を表し，$DF(term)$ は $term$ を含む文書数を表している．本指標は，TFIDayF[多田 09] と類似し，任意の時点における文書内での用語の出現度合いを各時点での文書数に対する用語を含む文書数の割合によって重み付けした値となる指標である．

また，用語の出現頻度である $DF(term, D_{period})$ について，各時点における出現確率 $p = DF(term, D_{period})/|D|$ を考慮して，これが各時点において特異となるかどうかを計量化する指標としてオッズを用いた．用語 $term$ に関するオッズ $Odds(term, D_{period})$ は，以下のように定義する．

$$Odds(term, D_{period}) = \frac{p}{1-p} = \frac{DF(term, D_{period})}{|D| - DF(term, D_{period})}$$

計量化指標を用いて，各時点における値が算出された各用語の時系列について，指標ごとにデータセットを生成して，時系列間の類似性に基づくパタンを生成する．

2.3 時系列パタンの同定と傾向分析

前節の操作により，各用語の用法や注目度の時間方向の変化は，計量化指標の値として計量化されている．この時系列について，時系列パタンを抽出する．本稿における時系列パタンの生成では，時系列間の類似性に基づくクラスタリングを適用する．

最後に，時系列クラスタリングによって得られた時系列パタンの傾向を同定する．時系列パタンの傾向として，本手法では，各クラスタの代表元について線形回帰に基づく傾きと切片を基準として，各パタンの解釈を割り当てる．

各時系列パタン c の傾き $Deg(c)$ は，代表元の値 y_i について，時間経過 $x_1, ..., x_n$ に対して，以下のように算出される．

$$Deg(c) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

この傾きとそれぞれの平均 $\bar{y}, \bar{x}$ を用いて $Int(c)$ は，以下のように算出される．

$$Int(c) = \bar{y} - Deg(c)\bar{x}$$

また，上の直線的傾向を表す値を各用語 $term$ にも適用し，時系列パタンのメンバであるそれぞれの用語の傾向とする．

3. 実例を用いた実験

本節では，2. 節で提案した手法を用いて，実際の専門的文書群における用語用法に関する傾向の分析を行う．実例として，人工知能学会全国大会の 2003 年から 2009 年まで，各年の発表タイトルをそれぞれ時間経過に従った文書群の集まりとして扱う．文書と発表タイトルの対応は，一発表のタイトルを一文書と見なすこととした．

これら 7 年分の各文書群から，隣接頻度に基づく用語抽出手法 [Nakagawa 00][*2]によって，用語を抽出した．用語中の単語抽出については，MeCab[MeC] を用いた．

次に，抽出された各用語について，各年での tf-idf 値およびオッズを算出し，それぞれデータセットを生成した．

3.1 用語の抽出

2003 年から 2009 年までの人工知能学会全国大会の発表のうちタイトルとして対象とした数は，順に 259,288,297,275,336,410,402 の計 2,267 である．これら各発表のタイトルを一文書として，$FLR(term) > 1.0$ となる用語を求めた．さらに，各用語の出現回数が複数回となる（$\sum DF(term, D_{period}) > 1$）1,148 語を対象とした．

これらの用語について，用語を含む文書数，tf-idf を基にした指標，用語を含む文書数を基にしたオッズのそれぞれを年ごとに算出した．年ごとに算出された値については，元となる文書数が異なるため，以下のように 0 から 1 までとなる正規化を適用した．

$$\hat{y_i} = \frac{y_i - min(y)}{max(y) - min(y)}$$

各指標について，列に各年の指標の値とし，行方向が各用語の時系列となるデータセットを作成した．

*2 公開された実装である TermExtract モジュール（http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html にて配布）を適用した．

表 1: 2003 年から 2008 年までの人工知能学会全国大会の発表タイトルに対する時系列パタン生成結果.

Cluster ID	用語を含む文書数(DF)					tf-idf					オッズ				
	メンバ数	term	Deg(term)	Int(term)	解釈	メンバ数	term	Deg(term)	Int(term)	解釈	メンバ数	term	Deg(term)	Int(term)	解釈
1	3	支援	0.034	0.702		10	利用	0.014	0.640		3	支援	0.038	0.675	
2	252	学習者	0.000	0.012		115	情報編纂	0.004	0.009		259	学習者	0.000	0.010	
3	63	モデル化	0.012	0.015		69	***性能評価***	0.013	−0.009	新興	12	開発	0.034	0.098	
4	149	移動支援	−0.004	0.027	沈静化	118	情報抽出	0.008	0.075		230	移動支援	−0.003	0.020	沈静化
5	107	構造化	0.004	0.007		177	***協調行動***	0.009	0.000	新興	83	情報抽出	0.008	0.015	
6	44	人間	−0.007	0.086	沈静化	165	移動支援	−0.011	0.063	沈静化	40	適用	−0.006	0.087	沈静化
7	244	化学構造	−0.001	0.010	沈静化	155	学習支援システム	−0.004	0.040	沈静化	232	化学構造	0.002	0.005	
8	26	開発	0.013	0.155		30	開発	0.007	0.350		17	設計	0.001	0.165	
9	213	***性能評価***	0.004	−0.002	新興	135	コミュニケーション支援	0.001	0.028		232	***性能評価***	0.004	−0.001	新興
10	14	利用	0.039	0.264		63	応用	−0.002	0.199	沈静化	11	利用	0.042	0.252	
11	33	可視化	0.009	0.080		111	化学構造	−0.002	0.039	沈静化	29	応用	0.015	0.047	

3.2 自動抽出された用語の時系列パタン分析

用語を識別子とする計量化指標の時間ごとの値は，用語の計量化指標に関する時系列データと見ることができる．本手法では，これら各用語の時系列の類似性から時系列パタン生成を行う．類似性に基づくパタン生成として，多くの類似性が定義可能であるが，ここでは 2 つの時系列 $\mathbf{x}$ と $\mathbf{y}$ について以下のように定義されるユークリッド距離尺度を用いる．

$$Sim(\mathbf{x},\mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

この類似度を基に，k-means クラスタリングを適用した．このとき，生成パタン数 k は，全用語数の 1%である $k = 11$ とした．

生成された時系列パタンについて，代表元に線形回帰による計量化指標の時間方向に対する傾きと開始時点における切片を算出した．これら傾きと切片を基準として，以下のように解釈を与える．

- 新興：傾きが正（$Deg(term) > 0$），かつ，切片が負（$Int(term) < 0$）
- 沈静化：傾きが負（$Deg(term) < 0$），かつ，切片が正（$Int(term) > 0$）

「新興」とした用語については，研究コミュニティにおいて注目が集まり，利用頻度や相対的な重みが上昇傾向であることを表す.「沈静化」とした用語については，さらなる研究トピックへの発展などによって，その用語自体の利用が減ったことを表す．なお，傾き・切片ともに正である時系列パタンについては，対象とする期間に恒常的に利用された用語群である．

表 1 に各指標での時系列パタンについて，各パタンに含まれる用語数（メンバ数），代表元に直近の用語，線形傾向である傾きと切片の値を各パタンの解釈と示す．指標間の相違をみると,「性能評価」を代表元とする時系列パタンが 3 つの指標によって「新興」と解釈できる．tf-idf に基づく指標では，時系列パタンの傾向が強調されるため,「新興」「沈静化」に割り当てられるパタン数が増えている．オッズでは，用語が含まれる文書数に対して，直線的な傾向が小さくなる傾向があり，どちらかの解釈が割り当てられるパタン数が減少する．ただし，これら 3 指標によって割り当てられたパタンの内訳を見ると，大きな相違は認められない[*3].

*3 データ数が多いため検出力が大きいと考えられるが，χ^2 統計量によると $p < 0.01$ となる．

3.3 tf-idf に基づく指標による時系列パタンの詳細

表 1 において，tf-idf に基づく指標による時系列パタンでは，用語を含む文書数 DF とは異なる用語を代表とした「新興」となる時系列パタンが見出された．ここでは，本指標による時系列パタンについて，詳細を述べる．

図 2 に示すように，それぞれの時系列パタンは横軸に示した時間経過とともに，表 1 に示した直線的な傾きをもって上昇，あるいは下降している．それぞれの年での値は，0 から 1 に正規化されているため，出現頻度の大小にかかわらず，各年間での比較が可能である．

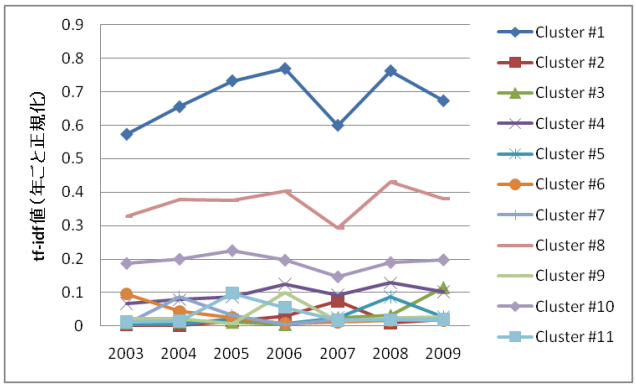

図 2: 人工知能学会全国大会（2003 年から 2008 年）の発表題目における tf-idf に基づく指標による時系列パタン．

次に，図 3 に，DF とは異なる用語が代表元となる「新興」のパタン（Cluster #5）について，決定係数がもっとも大きくなる 10 の用語を代表元の値とともに示す．これらの用語のうち，もっとも傾きが大きい「日常生活行動」を例とすると，傾きは 0.0272，切片は −0.0574 であり，この直線に対する決定係数は 0.62 であった．

ここで各用語の傾き $Deg(term)$ と切片 $Int(term)$ から指標についての予測値を $py_i = Deg(term) \times \hat{y}_i + Int(term)$ とすると，決定係数 R^2 は以下のように求められる．

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - py_i)^2}{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}$$

この結果から,「文書集合」「情報流通」「知識継承」「多重オントロジー」など，現在の情報化社会に対応した話題に関する用語や人間の身近な行動の支援に関する用語などが得られた．これらの用語は，近未来チャレンジセッションをはじめ，複数の発表セッションに分かれているが，現在の人工知能研究に関して注目を集めている研究と関連した用語と考えられる．

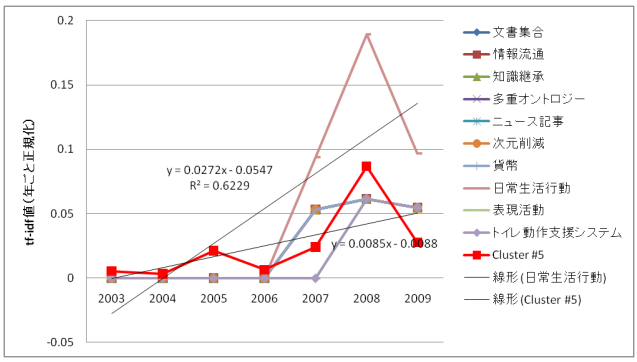

図 3: tf-idf に基づく指標による時系列パタン (Cluster #5) と各用語の傾きと切片に対する決定係数上位 10 位までの用語の値.

以上より，時間軸方向に対する傾きと開始点の切片を考慮することにより，時系列パタンの解釈が可能となった．これらの時系列パタンから得られる各用語の傾向は，計量化指標の値によってグループ化されたものであり，個々の用語を逐一判断するよりも早く目的の傾向となる用語を見出すことが可能である．また，これらの時系列パタンは，用語そのものの意味からではなく，テキストデータ中への出現度合いを計量化指標によって表される．このため，テキストデータ中への出現度合いの類似した未知の用語の同定も可能となると考えられる．

4. おわりに

本稿では，用語の計量化指標に基づく時系列パタンの傾向分析手法を提案した．計量化指標として，用語を含む文書数，時点毎の文書を対象とした tf-idf，用語を含む文書数に対するオッズを利用して，本手法を実装した．

評価実験においては，同一目的の文書群である人工知能学会全国大会の概要とタイトルを対象として，用語の時系列が類似した時系列パタンを k-means 法によって求め，各クラスタの代表元にたいする線形回帰による傾きと切片を基準に用語群の傾向を同定した．この結果，時系列パタンの解釈から人工知能学会全国大会における研究動向と関係する用語を見出すことができた．また，各計量化指標による時系列パタンの差異を分割表としたところ，各指標による時系列パタンに属する用語は統計的に同一であるという結果となった．

今後は，用語の計量化指標について開発を行い，より多くの計量化指標の算出を可能とする．さらに，各計量化指標の時系列パタンがどのようなイベントと関連付くのか，という視点から数値予測および分類学習によるモデル生成を行う予定である．

参考文献

[Kleinberg 03] Kleinberg, J. M.: Bursty and Hierarchical Structure in Streams, *Data Min. Knowl. Discov.*, Vol. 7, No. 4, pp. 373–397 (2003)

[Kontostathis 03] Kontostathis, A., Galitsky, L., Pottenger, W. M., Roy, S., and Phelps, D. J.: A Survey of Emerging Trend Detection in Textual Data Mining, *A Comprehensive Survey of Text Mining* (2003)

[Lent 97] Lent, B., Agrawal, R., and Srikant, R.: Discovering Trends in Text Databases, pp. 227–230, AAAI Press (1997)

[Matsuo 04] Matsuo, Y. and Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information, *International Journal on Artificial Intelligence Tools*, Vol. 13, No. 1, pp. 157–169 (2004)

[MeC] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, `http://mecab.sourceforge.net/`

[Mei 05] Mei, Q. and Zhai, C.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining, in *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 198–207, New York, NY, USA (2005), ACM

[Nakagawa 00] Nakagawa, H.: "Automatic Term Recognition based on Statistics of Compound Nouns", *Terminology*, Vol. 6, No. 2, pp. 195–210 (2000)

[Sparck Jones 88] Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval, *Document retrieval systems*, pp. 132–142 (1988)

[Swan 00] Swan, R. and Allan, J.: Automatic generation of overview timelines, in *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 49–56, New York, NY, USA (2000), ACM

[阿部 09] 阿部 秀尚, 津本 周作：重要度指標に基づく専門的テキストデータからのフレーズ傾向抽出, 2009 年度人工知能学会全国大会 (第 23 回), pp. 1C3–2 (2009)

[多田 09] 多田 知道, 岩沼 宏治, 鍋島 英知：イベント系列マイニングを目的とする新聞記事からの時間情報に基づく単語抽出, 人工知能学会論文誌, Vol. 24, No. 6, pp. 488–493 (2009)