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分析

A Method to Detect Remarkable Trends of Technical Phrases in Research Documents

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

*1島根大学
Shumane University

In this paper, we propose a method for detecting temporal trends of technical terms 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. As for the documents, they are often published daily, monthly, annually, and irregularly for each purpose. 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 combined a method to extract terms, importance indices of terms, and trend identification based on linear regression analysis. Empirical results show that our method detected emergent and subsiding trends of extracted terms in a corpus of a research domain.

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，時系列に沿って発行される種々の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のみを扱っており，単語や単語間の関係度合いとそれぞれの傾向検出のための指標が別々に議論されていない．このため，単語毎の傾向が全く異なる場合，複数の単語からなるフレーズの傾向を考察しようとしても，扱われている重要度がフレーズに対応していないため解釈が困難となるなど，課題がある．

これに対し，我々は，従来別々に扱われてきた，辞書に依らない用語の抽出，単語やフレーズの重要度指標，時系列の変化である傾向の抽出を統合したフレーズの傾向抽出手法を提案する．本稿では，専門的文書群として，データマイニング関係の国際会議である ICDM[ICD] の 2002 年から 2008 年にかけてのアブストラクトとタイトルについて，2 種類の傾向を示すフレーズが同定可能であることを示す．以上の結果から，テキストマイニングで広く利用される複数の重要度指標がフレーズの傾向検出において利用可能であることを示す．

2. 重要度指標に基づく文書中フレーズの傾向同定手法

本節では，以下の処理を統合した文書中のフレーズの傾向同定手法を提案する:

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. 時系列の傾向検出

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，抽出された用語の中から 2 つ以上の単語から成るフレーズを選定し，各フレーズについて，時点毎の文書群における重要度を算出する．この結果，各フレーズを行，各時点の重要度の値を列とするデータセットを作成される．生成されたデータセットに対し，時系列の傾向を抽出する手法を適用して，各フレーズの傾向を抽出する．手法の概観を図 1 に示す．

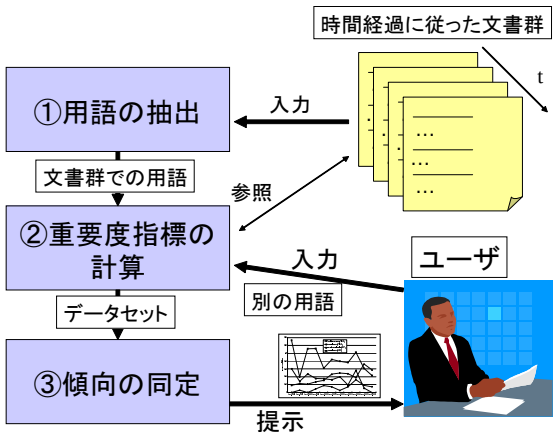

図 1: 重要度指標に基づく文書中フレーズの傾向同定手法の概観

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 [?] を用いた．この手法では，抽出候補となる単語

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

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 が左に出現する頻度，$FR(N_i)$ は N_i が右に出現する頻度を表している．

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

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．

用語抽出手法によって得られたフレーズについて，各時点の文書群における重要度算出する．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[Jones 88]，複数の単語の共起性を表す Jaccard 係数 [Anderberg 73] を対象とする．

各フレーズ t の tf-idf 値 $TFIDF(t)$ は以下のように計算される．

$$TFIDF(t) = \frac{TF(t)}{log_e \frac{|D|}{DF(t)}}$$

ここで，$TF(t)$ は，サイズ $|D|$ の文書における t の出現頻度を表し，$DF(t)$ は t を含む文書数を表している．また，単語数 $L \geq 2$ のフレーズ t の Jaccard 係数の値は，各単語 w_i が用いられたうち 1 つのフレーズとして用いられた割合として，以下のように計算される．

$$Jaccard(t) = \frac{h(w_1, w_2, ..., w_L)}{h(w_1)h(w_2)...h(w_L)}$$

各フレーズの重要度の指標としては，これらの他に文書中の共起度合いを表す n-gram[Shannon 48] など言語モデルに基づく指標，情報検索における評価尺度などが適用可能である．

そして，フレーズ毎に各時点の重要度指標の値について，時系列の傾向を同定する．データセットとして種々の時系列分析手法が適用可能であると考えられるが，ここでは線形回帰に基づく傾きと切片を評価基準として用いた．

各フレーズ t の傾き $Deg(t)$ は，各時点の重要度指標の値 y_i につて，時間経過 $x_1, ..., x_n$ に対して，以下のように算出される．

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

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

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

これらの結果は，ユーザに示され，評価を受ける．そして，さらに必要な単語やフレーズの入力を受け，所与の時系列に従う文書群での傾向を提示する．

以上の処理の繰り返しにより，本手法は，ユーザが目的とする傾向をもつ，より的確な用語の同定を可能にする．

3. 実験

本実験においては，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フレーズの傾向の同定を行う．文章群として，データマイニング分野の国際会議である ICDM の 2002 から 2008 まで，各年のアブストラクトと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．アブストラクトについては，各アブストラクトを一文書として扱う．タイトルについては，各タイトルを一文書とした．

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][*1]によって，用語を抽出した．

次に，抽出された各用語について，各年での tf-idf 値および Jaccard 係数を算出し，データセットを生成した．これらの値について，線形回帰による傾きと切片を算出し，これらを評価基準として並び替えを行って新興および沈静化の傾向を同定する．

3.1 用語の抽出

表 1 に ICDM の 2002 年から 2008 年までのアブストラクトおよびタイトルの文書数および単語数について示す．

表 1: ICDM のタイトルおよびアブストラクトの文書数と単語数．

年	アブストラクト		タイトル	
	文書数	単語数	文書数	単語数
2002	112	18,916	112	960
2003	125	19,068	125	1,040
2004	106	15,985	106	840
2005	141	20,831	141	1,153
2006	152	24,217	152	1,307
2007	101	16,143	101	782
2008	144	22,971	144	1,136
合計	881	138,131	881	7,218

これら 7 年分の全アブストラクトについて，単語の組み合わせは無意味なものも含めて 2^{138131} となるが，このうち 21,599 用語を抽出した．このとき，$FLR(t) > 1.0$ となる用語を求めた．同様に，全タイトルからは，1,912 用語が抽出された．

3.2 自動抽出されたフレーズの傾向同定

アブストラクトおよびタイトルから抽出された用語について，より特徴的なフレーズを抽出するため，単語数が 2 以上 9 以下のフレーズを選択した．これは，単語数が多い用語では，途中で交換可能な単語を含む可能性が増大し，構文と関連したパターンを抽出する方法 [Mei 05] が適すると考えるためである．

これらのフレーズについて，線形回帰による各フレーズの重要度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基準として，以下のように傾向を同定する．

- 新興
  - 降順に傾きの値を整列
  - 傾きが同順の場合は切片が小さいものを優先
- 沈静化
  - 昇順に傾きの値を整列
  - 傾きが同順の場合は切片が大きいものを優先

*1 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）の英文でのストップワード除去を適用した．

表 2: 2002 年から 2008 年までの ICDM のアブストラクトにおける「新興」「沈静化」の各傾向で上位 10 位のフレーズ.

順位	新興						沈静化					
	tf-idf			Jaccard係数			tf-idf			Jaccard係数		
	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)
1	Experimental results	0.0050	0.0182	collaborative filtering	0.077	0.284	data mining	−0.021	0.222	association rules	−0.055	0.457
2	social network	0.0032	−0.0020	upper bound	0.065	0.089	association rules	−0.011	0.057	association rule	−0.053	0.421
3	text mining	0.0028	−0.0014	social networks	0.054	−0.035	experimental results	−0.009	0.072	web pages	−0.043	0.480
4	real world	0.0027	0.0040	social network	0.053	0.027	association rule	−0.006	0.033	nearest neighbor	−0.042	0.501
5	feature extraction	0.0024	−0.0029	gene expression	0.051	0.479	data sets	−0.006	0.084	frequent itemsets	−0.039	0.367
6	real−world data	0.0016	0.0054	matrix factorization	0.047	0.043	frequent itemsets	−0.005	0.039	naive Bayes	−0.034	0.266
7	background knowledge	0.0016	−0.0004	Support Vector Machines	0.045	0.143	clustering algorithm	−0.003	0.025	experimental results	−0.032	0.322
8	social networks	0.0014	−0.0015	anomaly detection	0.038	0.027	mining algorithm	−0.003	0.016	dynamic programming	−0.027	0.218
9	synthetic data	0.0014	0.0048	random walk	0.038	0.057	clustering algorithms	−0.002	0.025	outlier detection	−0.022	0.217
10	real datasets	0.0013	0.0016	computational cost	0.037	−0.024	association rule mining	−0.002	0.012	feature selection	−0.017	0.268

表 3: 2002 年から 2008 年までの ICDM のタイトルにおける「新興」「沈静化」の各傾向で上位 10 位のフレーズ.

順位	新興						沈静化					
	tf-idf			Jaccard係数			tf-idf			Jaccard係数		
	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)
1	Data Streams	0.0010	0.0029	Collaborative Filtering	0.125	0.101	Association Rules	−0.00333	0.01638	Event Sequences	−0.107	0.607
2	Active Learning	0.0006	0.0000	Nonnegative Matrix Factorization	0.095	−0.048	Data Mining	−0.00111	0.01173	Association Rules	−0.089	0.583
3	Nonnegative Matrix Factorization	0.0005	−0.0007	Random Walk	0.095	0.048	Data Sets	−0.00104	0.00559	Decision Trees	−0.077	0.637
4	Collaborative Filtering	0.0005	0.0002	Dimension Reduction	0.089	0.018	Decision Trees	−0.00057	0.00404	Latent Semantic Indexing	−0.060	0.393
5	Text Categorization	0.0005	−0.0006	Hidden Markov	0.071	−0.071	Unsupervised Algorithm	−0.00048	0.00206	Experimental Evaluation	−0.054	0.375
6	Social Networks	0.0005	−0.0004	Belief Propagation	0.071	0.000	Web Page Classification	−0.00048	0.00206	Bayesian Network	−0.054	0.280
7	Dimension Reduction	0.0004	−0.0001	Taxonomic Research	0.071	0.071	Dimensional Data	−0.00040	0.00232	Document Categorization	−0.054	0.280
8	frequent itemsets	0.0004	−0.0002	Similarity Measure	0.065	−0.077	Time Series	−0.00038	0.00797	Fast Algorithm	−0.048	0.238
9	Document Clustering	0.0004	0.0003	Pairwise Constraints	0.051	−0.068	Latent Semantic Indexing	−0.00037	0.00227	Utility Itemsets	−0.045	0.313
10	Sequential Pattern Mining	0.0003	−0.0002	Link Prediction	0.050	0.021	data mining	−0.00037	0.00207	data mining	−0.042	0.244

以上の操作によって各フレーズの傾きを同定した結果を表 2 および表 3 にそれぞれの傾きと切片の値と共に示す.

以上の結果から，本手法によって同定された各フレーズの傾向は，それぞれの文書群において同一の場合と，異なる場合が見られた.

“Social Network” あるいは “Social Networks” は，各文書群において，２種類の重要度指標の傾向が「新興」を表す結果となった．これは，このフレーズが頻度および共起の特異性の両方が増大してきたことを表している．このことは，当該の研究分野において，著者および査読者が注目している研究対象を調べる上で有用な情報となる.

一方，“Matrix Factorization” というフレーズは，Jaccard 係数の傾向において「新興」としてより上位に位置づけられた．このフレーズのように，共起の特異性のみにおいても時間経過に従った文書から注目されているものが同定されうる，と考えられる．したがって，場面において各重要度指標が表す意味について，評価者の専門知識と併せて得られた傾向を評価する必要がある.

また，決定木や相関ルールといったデータマイニングにおける基本技術を表すフレーズの使われ方が沈静化の傾向を示している．しかし，文書群における重要度の沈静化はこれらの技術に重要性が当該研究分野で低下したとは言えない．むしろ，過去の盛んな研究によって基本技術が確立され，各問題領域に適用した方法や，より別の方向からの研究として扱われるようになった，と言える.

4. おわりに

本稿では，フレーズの抽出について重要度指標に基づく傾向抽出手法を提案した．専門的文書である ICDM のタイトルとアブストラクトにおいて，各年の重要度指標の値に対する線形回帰による傾きと切片を基準にフレーズの傾向が同定可能であることを示した.

評価実験においては，フレーズの出現回数に依存する tf-idf とフレーズに含まれる単語同士の共起に依存する Jaccard 係数の傾向を用い，複数の重要度指標において同一の傾向を示すフレーズの多くが一致する結果となった．ただし，異なるフレーズが上位となる相違については，それぞれの重要度指標の特徴を踏まえて，分析を進める必要がある.

今後は，重要度指標についてテキストマイニングに限らず，情報検索における評価指標についても利用可能性を検討する．また，時系列の傾向を得るため，時系列クラスタリングの適用なども行っていく．さらに，各重要度指標の傾向がどのようなイベントと関連付くのか，という視点からルール生成を行っていく予定である.

参考文献

[Anderberg 73] Anderberg, M. R.: *Cluster Analysis for Applications*, Monographs and Textbooks on Probability and Mathematical Statistics, Academic Press, Inc., New York (1973)

[ICD] IEEE International Conference on Data Mining:, `http://www.cs.uvm.edu/~icdm/`

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

[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)

[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)

[Shannon 48] Shannon, C. E.: A Mathematical Theory of Communication, *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656 (1948)

[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