専門分野において重要となる新語の特定に向けた基礎研究

辻慶太 (国立情報学研究所 keita@nii.ac.jp)
芳鐘冬樹 (大学評価・学位授与機構 fuyuki@niad.ac.jp)

1 はじめに

本研究は，テキストに新たに出現した新語の中から，今後重要になる語を自動的に抽出できるようになることを目指し，新語に関するいくつかの性質を調査したものである。ここで「重要な語」とは次の２つを考えている。即ち，(1) よく用いられるようになった語（テキスト中の頻度が高くなった語），(2) 専門的な概念と結び付いた語，の２つである。こうした予測が可能になると，まず用語辞書の編纂・改訂の際に，ある語を加えるべきか否かを判断する場面で有用であり，また今後よく普及する語が分かれば，ある分野で今後注目を集める研究・トピックが把握しやすくなり，トレンド分析的な面でも有用であろう。

本研究の背景及び問題意識としては次の２つがある。

(a) 用語抽出における時系列的観点の重要性：

これに関しては，コーパスと専門用語の２つが関連している。まず，コーパスからの専門用語抽出の多くは，過去に作られたコーパスから語を抽出し，抽出した語を現在の観点から評価し，現在における専門用語が抽出できたか否かを測定してきたように見える。従ってそこでは，今作られたばかりのコーパスは入手が難しいという現実的な要請から，ある程度過去のコーパスからでも現在の専門用語は抽出できることを仮定して，用語抽出を行っていると言える。そこではどの程度過去のコーパスまでなら，現在においても有効なのかという検証はあまり行われて来なかった。即ち，コーパスに関する時間的な観点の不足である。次に専門用語についてだが，例えば用語抽出が，用語集の編纂等を目的に行われる場合は，用語集が将来ある期間に有効に機能することが求められ，上記のようなプロセスで抽出した語は，現在においてだけでなく，ある程度将来にわたっても専門用語であることが期待されていると言える。この仮定がどの程度有効なのかといった，専門用語に関する時系列的な検証についても，十分行われてきたとは言い難い。語ごとに，将来における専門用語としてなどの位置付けが予測できれば好ましいであろう。

(b) 用語抽出における新語とそれ以外の区別の重要性：

専門用語抽出では全般に，新語の専門用語とそれ以外とがあまり区別されて来なかった。現実の応用の場面では，新語とそれ以外の語ではニーズや扱われ方に差があると思われる。また新語は，テキスト中での頻度が元々低いといった，用語抽出手法と密接に関連した性質の違いを持っている。従って，現実の応用の観点からも，抽出手法の観点からも，新語を区別して扱うことは有効と思われる。

上記のような観点から，本研究では１つの雑誌論文テキストを，生み出された時点に基づいて新旧の２つに分け，時系列的に異なる２つのコーパスとして調査対象コーパスとする。また新旧の境において現れた新語に焦点を当て，それらの中から，そのテキストに表される分野において重要となる語が，従来の専門用語自動抽出手法あるいは尺度に対して，どのような性質を持っているかを明らかにする。

最後に，本研究では２名詞から成る複合語を抽出対象とする。また重要になるか否かを予測する手がかりには様々なものが考えられるが，今回は主に，語構成要素のテキスト中での統計的特徴に焦点を絞り，言語外的要因や他分野のテキストなどは扱わないことにした。

2 関連研究

テキストを時系列的な観点から眺め，過去にない新しい情報の発見を試みる研究としては，新規トピックの検出に関する Yang et al.(2002)，Zhang et al.(2002)，Allan et al.(2003)，Ma & Perkins (2003) などがある。このうち，Allan et al.(2003) は新規トピックを含んだ文の特定を目指しているが，他は基本的に文献の特定を目指している。そこでは「新規トピック」とは何なのかという定義が曖昧であり，また新規トピックの寿命や，そのトピックはその後新しい流れを作るトピックなのかといった，将来に関する判断はあまり行われていない。一方，新しい流れを作った文献を特定する研究としては，Allan et al.(2000) の First Story Detection (FSD) などがある。本研究との関係で言うと，FSD はある時点から過去を眺めて文献を特定するものであり，将来の予測を目指す本研究とは時間に関する視点が逆になっていると言える。

専門用語抽出研究に関して言うと，これまでに提案された抽出手法は，まずテキスト中に出現した回数を基本にする手法（Damerau (1993)，Daille et al.(1994)，Justeson & Katz (1995)，Pantel & Lin (2001)，及び Frantzi et al.(1998)・Mima & Ananiadou (2000) の C-Value に基づく手法，主にキーワード抽出に用いられてきた tfidf に基づく手法）がある。用語抽出手法としては他に，候補語の出現箇所の周辺に共起する語の性質に基づく手法があり（Hisamitsu et al.(2000)，Maynard & Ananiadou (2000)，合原ら (2000)，山田ら (2000)，竹内 & コリアー (2002)，及び Frantzi et al.(1998)・Mima & Ananiadou (2000) の NC-Value に基づく手法），語構成要素の前後に連接する語の異なり数に注目する手法もある（Nakagawa (2000)，湯本ら (2001)）。また語の表記に注目する手法もある（福田 (1997)，山田ら (2000)）。これらは先述のように，コーパスの時系列的な位置付けや，抽出する専門用語の寿命，新語とそれ以外の区別などがあまり意識されていないが，本研究が目指す重要語の自動抽出に利用できる可能性もある。そこで，本研究では Hisamitsu et al.(2000) と Nakagawa (2000) の尺度，及び tfidf を以下で取り上げる。

時系列的にテキストを扱い，新語の中からその後高頻度で用いられるようになる語の特定を目指した研究としては，辻 & 芳鐘 (2003) がある。そこでは２名詞列の語の語構成要素の経年的な増加率に焦点が当てられている。本研究は用語抽出に用いられてきた統計的な尺度に主な焦点を当てたものである。

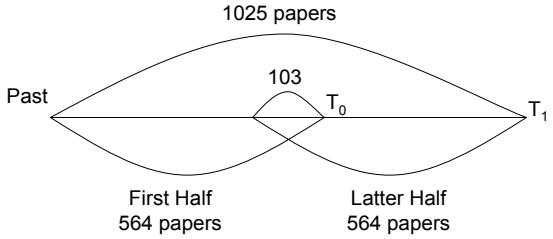

図 1: 本研究が扱う 2 つのコーパスと観察時点

3 調査

3.1 データ

先述の抽出手法を開発するには，(1) 経年的に並べられたテキストを用意し，(2) 現代からある程度離れた過去の時点を，調査対象新語出現期間に設定し，(3) その期間に現れた新語のうち，現代において重要となった語を，調査対象新語出現期間の直後において，その期間より過去のテキストに基づいて抽出する方法を探すというのが，1 つの明解な方向と思われる。本研究もその方向で行う。

本研究では *Journal of the American Society for Information Science (and Technology)* (以下 'JASIST') において，ある期間に発表された論文中の語の集合は，ある 1 分野のある期間の語の集合であるとみなして，分析の対象とした。期間は 1986 年〜 2002 年の計 17 年で，対象部分は 1,025 論文の本文である。これら 1,025 論文のうち，真ん中の 103 論文（1996 年の vol.47，no.8 から 1998 年の vol.49，no.2 に含まれる）を語の初出期間論文とし，その 103 論文を含んだ前半・後半 564 論文ずつを 2 つの調査対象コーパスとした。[1] 図示すると図 1 のようになる。このうち，T0 が調査対象新語出現期間の直後であり，どの新語が重要になるか予測できるようになることを，本研究が目指す時点である。そして T1 がどの新語が重要になったかを把握する時点である。「重要」とは第 1 章で述べたように，1 つは後半における頻度が高かった語である。もう 1 つの「専門的な概念と結び付いた語」としては，Hisamitsu et al.(2000) の尺度が高かった語とする。調査対象語としては，前半最後の 103 論文に出現していて，かつそれ以前には出現していない 2 名詞列 S_1S_2 を，新たに現れた新語として調査する。[2]

本研究では以下を仮定している。即ち，得られる知見・調査結果は，調査対象新語出現期間，即ちある程度過去に関するものとなるが，それらは現代にも通じ，先述の応用場面で有効に機能する。さらに本研究では，全テキスト中の語は，出現した時点は異なっていても，表記が同じならば同じ語であると仮定している。

テキストは Brill tagger で品詞付けを行い，'NN'，'NNS'，'NNP' と判定された語を名詞として扱った。全語の延べ数は約 876 万，名詞の延べ数は約 245 万であった（名詞はすべて単数形に統一した）。

[1] 前半と後半の論文数を同じにしたのは，後述の Hisamitsu の尺度や Nakagawa の尺度を前後半で比較できるようにする為である。これらの尺度にはサンプルサイズに依存する性質がある。

[2] 3 名詞以上の連続列の例は少なかったため今回は取り上げなかった。また単名詞や名詞以外については，後述のように今後の課題とし，今回は取り上げない。

頻度＼ RepL	0 〜 1.0	1.0 〜 1.5	1.5 以上	全体
1	6,248	2,572	3,013	11,833
2 〜 5	813	629	493	1,935
6 以上	33	40	25	98
全体	7,094	3,241	3,531	13,866

表 1: サンプル数

3.2 抽出の尺度

Hisamitsu et al.(2000) の尺度 Rep は以下の通りである。まず $D(T)$ は語 T を含む文献の集合，D_0 はすべての文献を表す。また k_i，K_i はそれぞれ，$D(T)$，D_0 における語 w_i の頻度，$\#D(T)$，$\#D_0$ はそれぞれ $D(T)$，D_0 に含まれる語の数を表す。

$$Rep(T) = Dist(P_{D(T)}, P_0)/B(D_0, \#D(T))$$

ここで，

$$Dist(P_{D(T)}, P_0) = \sum_{i=1}^{n} k_i \log \frac{k_i}{\#D(T)} - \sum_{i=1}^{n} k_i \log \frac{K_i}{\#D_0}$$

であり，$B(D_0, \#D(T))$ は $\#D(T)$ と同じ大きさの文献 D を無作為に抽出して算出したところの $Dist(P_D, P_0)$ である。

Nakagawa (2000) は次の尺度の値が高い名詞列 $S_1S_2...S_k$ を専門用語として抽出することを提案している。

$$Imp_1(S_1S_2...S_k) = (\prod_{i=1}^{k}((Pre(S_i)+1)(Post(S_i)+1)))^{\frac{1}{k^a}}$$

ここで $Pre(S_i)$，$Post(S_i)$ は S_i の前後にそれぞれ連接した名詞の異なり数を表す。本研究では Nakagawa (2000) の実験で最も結果が良かった a=1 を採用した。また 2 名詞列 S_1S_2 に関して，S_1，S_2 の連接名詞異なり数は，S_1S_2 の前半における値を用いた。

キーワードや専門用語の抽出においてよく用いられる $tfidf$ は以下の通りである。まず語 T の $tfidt$ は，語 T の全文献における頻度を f(T)，語 T を含む文献の数を $N_(T)$，全文献数を N_0 とすると，

$$tfidt(T) = f(T) \log \frac{N_0}{N(T)}$$

である。

3.3 結果

先述の 103 論文において初出した 2 名詞列 13,866 個を，後半 564 論文における頻度（1，2 〜 5，6 以上）と Hisamitsu et al.(2000) の尺度 RepL（0 以上 1.0 未満，1.0 以上 1.5 未満，1.5 以上）に基づいて分類したところ表 1 のようになった（以下では Hisamitsu の尺度は，前半に関しては RepF，後半に関しては RepL と表すことにする）。

3.3.1 Hisamitsu の尺度

表 1 のそれぞれのセルに属する語の，前半 564 論文における Hisamitsu の尺度 RepF の値の平均値を調べたところ，表 2 のようになった。また RepF の値が最も

頻度＼ RepL	0 ～ 1.0	1.0 ～ 1.5	1.5 以上	全体
1	0.70	1.43	2.37	1.29
2 ～ 5	0.75	1.33	2.14	1.29
6 以上	0.84	1.50	2.22	1.46
全体	0.71	1.41	2.34	1.29

表 2: RepF の平均値

高かった 500 語が，後半 564 論文において，どの位置に当てはまったかを示したのが表 3 である。表 3 から，RepF が高い語は，後半のコーパスにおいて頻度を伸ばすことは少なく（あるいは表 1 から推定される値より低く），多くは頻度 1 にとどまることが分かる。

後半における Hisamitsu の尺度 RepL が 4 位の語には同じ値 2.83 の語が 194 個現れていた。さらに同じ値 2.73，2.49，2.46 を共有する語はそれぞれ 81，90，96 個あった。これら同じ値を共有する語はすべて，頻度 1 の語であり，それぞれ同じ文献に出現するものであった。これらの語の，前半における Hisamitsu の尺度 RepF の値も高く，最上位の一部を占めるものであった。Hisamitsu の尺度は，特異な 1 文献にのみ現れる語に高い値を付与する。従って Hisamitsu の尺度の値が非常に高い語は，その値を以てそのまま，分野で特異な文脈を持った特殊な重要語とみなすことは避けた方が良く，何らかの尺度と組み合わせて考えた方が良いようである。

実際に観察された語だが，まず頻度が 6 以上で，RepF が 1.0 未満の語には，道具的なものや固有名が目に付いた（例：“HTTP servers”，“FTP sites”，“E-mail address”，“Dublin Core”，“Windows NT”）。また頻度が 6 以上で，RepF が 1.5 以上の語には，研究（の一分野）を表す語が目に付いた（例：“Internet use”，“extraction techniques”，“modeling user”，“usability study”，“interaction models”）。頻度が 6 以上で，RepF が 1.0 以上 1.5 未満の語には，Web 関係の語が多いのだが，まだそれほど Web の中心的な語ではなかった語が目立つ（例：“Web resources”，“search agents”，“Web authors”，“directory service”，“virtual libraries”）。ただしもちろん中間的な枠であるように，“HTML page”のような道具的な語もあるし，“game theory”のような研究の一分野を表す語もあった。

Hisamitsu の値に関しては，分野の流れと関連して興味深い結果が得られた。頻度が最も大きかった語の表 4 にもその一端が現れているが，Hisamitsu の値が前半から後半にかけて増加した語は 324 個，一方前半から後半にかけて減少した語は 13,542 個であり，前者に比べて圧倒的に多かった。このことはある分野で用いられる語の経年的変化は，過去に現れた語や文献に従うものであり，それらを取り込んで「なじませる」方向に進むものであることを意味している。ある語の Hisamitsu の尺度の値が高くなるということは，その分野で用いられる語達が，その語を浮いた存在にする方向にまとまって動くということであり，それは現実になかなかないとも考えられる。

3.3.2 Nakagawa の尺度

先述の 103 論文において初出した 2 名詞列 13,866 個を上記と同様に分類し，前半・後半 564 論文における Nakagawa の尺度 Imp の平均を調べたところ，それぞれ表 5 のようになった。表 5 から，全体的な傾向として，前半における Imp の値が高い方が，後半に頻度が伸びるようである。だが Hisamitsu の尺度が大きくなる傾向はあまり見られない。このことは，前半における Imp の値が最も高かった 500 語が，後半においてどのように分布したかを調べた，先ほど同様の表 6 にも現れている。

頻度＼ RepL	0 ～ 1.0	1.0 ～ 1.5	1.5 以上	全体
1	– (0)	– (0)	3.08 (454)	3.08 (454)
2 ～ 5	– (0)	– (0)	2.96 (42)	2.96 (42)
6 以上	– (0)	5.04 (1)	3.05 (3)	3.55 (4)
全体	– (0)	5.04 (1)	3.07 (499)	3.07 (500)

表 3: RepF の値上位 500 語の，RepL の平均と語数

語	頻度	RepF	RepL	増減
Web site	119	5.042	1.491	-3.552
Web page	110	2.651	1.319	-1.332
Web browser	46	2.365	1.495	-0.869
Internet search	32	2.291	1.519	-0.771
Alta Vista	28	2.568	1.684	-0.884
Web user	28	2.150	1.417	-0.733
Internet use	22	3.263	2.356	-0.907
Web resource	21	1.853	1.339	-0.515
Dublin Core	19	0.844	0.746	-0.098
information visualization	17	0.962	0.892	-0.070

表 4: 頻度が最も大きかった 10 語の Rep と増減

ところで表 5 では，頻度 6 以上で，RepL が 0 以上 1.0 未満の値の低さ（5053.2）が目に付く。この値は，他の頻度 6 以上の値に比べ，有意水準 0.01 で低いことが言えた。このセルに属する語を調べてみると，“FTP site”，“HTTP server”，“metadata element”のように，前の語がそれまであまり用いられず，前半コーパスにおける頻度の低かった複合語が多く，実際，平均出現頻度も低かった。また表 5 から分かるように，このセルの平均値は，後半になっても低いことが分かる。

3.3.3 tfidf

前節同様に，前半・後半 564 論文における tfidf の平均を調べたところ，それぞれ表 7 のようになった。表 7 から，Imp と同様，前半における tfidf の値が高い方が，後半に頻度が伸びるようである。だが同様に，Hisamitsu の尺度が大きくなる傾向はあまりない。このことは，前半における tfidf の値が最も高かった 500 語が，後半においてどのように分布したかを調べた，先ほど同様の表 8 にも現れている。

ところで表 7 では，前半における tfidf に関して，頻度

前半における Imp の平均値				
頻度＼ RepL	0 ～ 1.0	1.0 ～ 1.5	1.5 以上	全体
1	2086.4	1725.1	2425.5	2094.2
2 ～ 5	6651.5	6365.5	6620.2	6550.6
6 以上	5053.2	13219.3	10363.6	9741.0
全体	2623.4	2767.5	3067.4	2770.1
後半における Imp の平均値				
頻度＼ RepL	0 ～ 1.0	1.0 ～ 1.5	1.5 以上	全体
1	3425.1	3087.2	4038.7	3507.9
2 ～ 5	12833.4	13341.9	13426.9	13149.9
6 以上	16227.9	41974.8	30431.3	30360.1
全体	4562.9	5557.3	5536.4	5043.2

表 5: Imp の平均値

頻度＼ RepL	0 〜 1.0	1.0 〜 1.5	1.5 以上	全体
1	30511.8 (150)	25649.7 (56)	31116.1 (86)	29757.3 (292)
2 〜 5	38818.5 (82)	33366.7 (59)	33295.4 (54)	35639.5 (195)
6 以上	26994.2 (3)	57435.5 (7)	37982.4 (3)	45921.4 (13)
全体	33365.4 (235)	31205.5 (122)	32083.1 (143)	32471.6 (500)

表 6: 前半における Imp の値上位 500 語の，後半における Imp の平均と語数

前半における tfidf の平均値				
頻度＼ RepL	0 〜 1.0	1.0 〜 1.5	1.5 以上	全体
1	8.54	8.63	9.98	8.92
2 〜 5	13.92	13.71	21.91	15.89
6 以上	23.71	12.23	28.70	20.30
全体	9.23	9.66	11.78	9.98
後半における tfidf の平均値				
頻度＼ RepL	0 〜 1.0	1.0 〜 1.5	1.5 以上	全体
1	8.54	8.63	9.98	8.92
2 〜 5	23.03	23.29	33.47	25.77
6 以上	96.75	137.60	98.31	113.82
全体	10.61	13.06	13.88	12.02

表 7: tfidf の平均値

6以上で，RepLが 1.0 以上 1.5 未満の値の低さ（12.23）が目に付く。この値は，他の頻度6以上の値に比べ，有意水準 0.01 で低いことが言えた。逆に後半における tfidf に関しては，このセルの値 137.60 は，他よりも有意水準 0.01 で高いことが言えた。前半における tfidf の値が，中の上程度の語は，後半において（tfidf の値が他よりも高くなる程度に）頻度を伸ばすと同時に，Hisamitsu の値で示される文脈的な特徴も持った語になると言えるかもしれない。

頻度＼ RepL	0 〜 1.0	1.0 〜 1.5	1.5 以上	全体
1	54.41 (141)	72.01 (49)	62.64 (124)	60.40 (314)
2 〜 5	62.57 (69)	64.16 (47)	126.16 (53)	82.95 (169)
6 以上	54.20 (9)	38.88 (3)	104.01 (5)	66.15 (17)
全体	56.97 (219)	67.28 (99)	82.27 (182)	68.22 (500)

表 8: 前半における tfidf の値上位 500 語の，後半における tfidf の平均と語数

4 おわりに

本研究では，新たに現れてから多くの論文に現れた 2 名詞列と，Hisamitsu の尺度の値が高い 2 名詞列を調査対象とし，初出時点におけるそれらの性質や，また抽出尺度の値の経年的な変化に関する一側面を明らかにした。これらの調査結果は，新語の中から重要になる語を特定する手法の開発の基礎となり，また従来あまり考慮されてこなかった時系列的観点の導入を用語抽出に行う際に，何らかの示唆を与えるものとなろう。今後の研究方向としては次の４つを挙げたい。即ち，(1) 以前行った語構成要素の増加率に基づく予測可能性（辻 & 芳鐘 (2003)）の利用，(2) 名詞の表記上の特性や初出時の文章表現の特徴といったヒューリスティックスの利用，(3) 他のテキストコーパスの利用，(4) 初出させた著者の特性といった言語外的要因の考慮，の４つである。最後に，今回は 2 名詞列をそのまま語として用いたが，その中にはより大きな複合語の一部であって，構造上，語としてのまとまりに欠ける名詞列があった。それらをデータから除外した形での検証も行い，また単名詞に関しても調査を進めたい。

謝辞

本研究の一部は「科学研究費補助金若手研究 (B)」課題番号 15700216 によるものであり，ここに謝意を表します。

参考文献

[1] Allan, J., Lavrenko, V. and Jin, H. (2000) “First Story Detection in TDT is Hard,” *Proceedings of the Ninth International Conference on Information and Knowledge Management*, p.374-381.

[2] Allan, J., Wade, C. and Bolivar, A. (2003) “Retrieval and Novelty Detection at the Sentence Level,” *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, p.314-321.

[3] Brill tagger http://www.cs.jhu.edu/~brill/

[4] Daille, B., Gaussier, É. and Langé, J. (1994) “Towards Automatic Extraction of Monolingual and Bilingual Terminology,” *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, p.515-521.

[5] Damerau, F. J. (1993) “Generating and Evaluating Domain-oriented Multi-Word Terms from Texts,” *Information Processing & Management*, vol.29, no.4, p.433-447.

[6] Drouin, P. (2003) “Term Extraction Using Non-technical Corpora as a Point of Leverage,” *Terminology*, vol.9, no.1, p.99-115.

[7] Frantzi, K. T., Ananiadou, S. and Tsujii, J. (1998) “The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms,” *Proceedings of the Research and Advanced Technology for Digital Libraries. Second European Conference, ECDL'98*, p.585-604.

[8] Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M. and Takano, A. (2000) “Extracting Terms by a Combination of Term Frequency and a Measure of Term Representativeness,” *Terminology*, vol.6, no.2, p.211-232.

[9] Justeson, J. S. and Katz, S. M. (1995) “Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text,” *Natural Language Engineering*, vol.1, no.1, p.9-27.

[10] Kageura, K. and Umino, B. (1996) “Methods of Automatic Term Recognition: A Review,” *Terminology*, vol.3, no.2, p.259-289.

[11] Ma, J. and Perkins, S. (2003) “Online Novelty Detection on Temporal Sequences,” *Proceedings of the ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, p.613-618.

[12] Maynard, D. and Ananiadou, S. (2000) “Identifying Terms by their Family and Friends,” *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, p.530-536.

[13] Meyer, I. and Mackintosh, K. (2000) “When Terms Move into Our Everyday Lives: An Overview of De-terminologization,” *Terminology*, vol.6, no.1, p.111-138.

[14] Mima, H. and Ananiadou, S. (2000) “An Application and Evaluation of the C/NC Value Approach for the Automatic Term Recognition of Multi-word Units in Japanese,” *Terminology*, vol.6, no.2, p.175-194.

[15] Nakagawa, H. (2000) “Automatic Term Recognition based on Statistics of Compound Nouns,” *Terminology*, vol.6, no.2, p.195-210.

[16] Pantel, P. and Lin, D. (2001) “A Statistical Corpus-Based Term Extractor,” *Proceedings of Advances in Artificial Intelligence, 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (AI 2001)*, p.36-46.

[17] Yang, Y., Zhang, J., Carbonell, J. and Jin, C. (2002) “Topic-Conditioned Novelty Detection,” *Proceedings of the eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, p.688-693.

[18] Zhang, Y., Callan, J. and Minka, T. (2002) “Filtering: Novelty and Redundancy Detection in Adaptive Filtering,” *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p.81-88.

[19] 合原博，宮田高志，松本裕治 (2000) “医学生物学分野からの専門用語の抽出・分類，” 情報処理学会研究報告，NL-135，p.41-48.

[20] 竹内孔一，コリアー・ナイジェル (2002) “生物学文献からの専門用語抽出における機械学習モデルの検討，” 情報処理学会研究報告，NL-150，p.185-190.

[21] 辻慶太，芳鐘冬樹 (2003) “専門用語として普及しそうな語の自動抽出，” 第 51 回日本図書館情報学会研究大会発表要綱，p.105-108.

[22] 福田賢一郎，角田達彦，田村あゆち，高木利久 (1997) “医学生物学文献からの専門用語の抽出，” 情報処理学会研究報告，NL-121 FI-47，p.103-110.

[23] 山田寛康，工藤拓，松本裕治 (2000) “単語の部分文字列を考慮した専門用語抽出と分類，” 情報処理学会研究報告，NL-140，p.77-84.

[24] 湯本紘彰，森辰則，中川裕志 (2001) “出現頻度と連接頻度に基づく専門用語抽出，” 情報処理学会研究報告，NL-145 FI-64，p.111-118.