情 報 学 基 礎 47−16
(1997. 9. 12)

分野判定のために用いられる単語リストの作成と応用

石田 栄美† 石塚 英弘† 根岸 正光‡ 山本 毅雄†
†図書館情報大学
〒305 茨城県つくば市春日 1-2
‡学術情報センター
〒112 東京都文京区大塚 3-29-1

先に発表した、大きな分野を対象とする大量の自然語テキストデータの検索・分類のための単語リスト作成方法を詳しく検討した。情報処理学・農芸化学・土木学などの学会予稿抄録の一部を単語リスト作成用データとし、他を検索・分類用データとした。単語リストの選択基準に含まれる二つのパラメタを変えて検討したが、単語リストの大きさや内容は大きく変わるものの、互いに離れている上記 3 分野間では、分野推定の結果は安定していることがわかった。さらに、分野に重なりのある電子情報通信学を加えて影響を調べた。また、これらの単語リストを応用して文献の分野関連度、文献集合の分野関連度などを求める方法を示した。

Formulation and Application of Word Lists for Classifying Texts into Large Fields

Emi Ishida† Hidehiro Ishizuka† Masamitsu Negishi‡ Takeo Yamamoto†
†University of Library and Information Science
1-2,Kasuga,Tsukuba,Ibaraki,305,Japan
‡National Center for Science Information Systems
3-29-1,Otsuka,Bunkyoku,Tokyo,112,Japan

A formerly developed method for statistically formulating word lists for large fields, which can be used for classification and retrieval of Japanese texts, is examined in detail. Two parameters studied, used for the selection of words in the list, were found not to influence assignments of abstract texts into three mutually independent separate areas: information processing, agricultural chemistry and civil engineering. The effect of adding the fourth field, electronics, information and communication engineering, was also examined. Formulas for calculating affinity to each field of an individual text and also of a body of texts were given.

1 はじめに

ある分野に関するテキストを集めたい、検索したい、あるいはどの分野に属するか分類したい場合、比較的小さな分野であれば、いくつかの特徴的な単語の組み合わせを含む、あるいは含まないこと、あるいはその組み合わせ（語の出現-Bool演算法）で行うことが普通に行われている。大きな分野であっても、テキストの量が全体として少ないか、件数が少ない場合、内容を人が読んで決定することができる。しかし、対象が大きな分野であり、テキスト件数もきわめて大きい場合、これをいかに実行するかが問題である。

先にわれわれ[1]は、情報処理学、農芸化学、土木学のような大きく、比較的独立した分野を推定できるような単語リストを作成し、その妥当性を評価した。本論文では、単語リストに含める単語の選択基準について2種のパラメータを変え実験を行い、選択基準の変化による分野推定への影響を調べた。また、上の3つの分野の少なくとも一つと大きな重なりをもつ分野（電子情報通信学）を加えた4分野で単語リストを作成し、分野を増やした場合の影響を調べた。さらに、論文中での各分野の単語リスト中の語の出現頻度から、文献および文献集合の分野関連度を求めた。分野関連度は、文献あるいは文献集合が分野とどの程度関連しているかを示すものである。

本研究で作成する単語リストは、文献集合中における単語の実際の出現回数を測定した出現確率を比較し、選択することによって求める。単語を単語リストに含める選択基準には二つの要素がある。一つは相対出現率の比であり、もう一つは出現回数0のかわりに与えられる出現回数である。前者は分野間で単語の出現率が何倍以上の差がついたときに単語リストに含めるかを決める値であり、後者は分野内で単語が出現しないときに出現確率が推定できず、他の分野と出現率の比較ができないため、便宜的に与えられる値である。本論文では、まず、この2つの変数により、単語リストがどのように変化するかを調べた。

2 方法の定式化

- 分野$1,2,\cdots,i,\cdots,I$のI種の分野の文献を単語リスト作成用とする。分野$1,2,\cdots,f,\cdots,F$のF種の分野の文献を単語リストの評価用とする。本報告では、$I=F=3$、および$I=F=4$の場合を調べる。
- テキストからMHSA[2][3]によって切り出した語（未知語として切り出した語を含む）を“単語”とする。
- 分野iに属する文献中、M_i個の文献の集合$D_i=\{D_{i1},\cdots,D_{ij},\cdots,D_{iM_i}\}$を単語リスト作成用に、別の$M'_f$個の文献の集合$D'_f=\{D'_{f1},\cdots,D'_{fj},\cdots,D'_{fM_f}\}$を手法の評価用と応用に用いた。
- 単語リスト作成用と評価用の全文献中の語彙の大きさをNとし、語彙を$\{T_1,T_2,\cdots,T_N\}$とする。分野iの全語彙$\{T_{i_1},T_{i_2},\cdots,T_{i_n}\}(n<N)$は、当然ながら$\{T_1,T_2,\cdots,T_N\}$の部分集合である。
- 文献D_{ij}における単語T_kの出現回数をc_{ijk}とすると、文献集合D_i内での単語T_kの総出現回数c_{ik}、および修正総出現回数a_{ik}は、

$$c_{ik}=\sum_{j=1}^{M_i}c_{ijk} \tag{1}$$

$$a_{ik}=\begin{cases} s & (c_{ik}=0) \\ c_{ik} & (c_{ik}\neq 0) \end{cases} \tag{2}$$

である。

- 文献集合D_i内での単語T_kの修正総出現率A_{ik}は、

$$A_{ik}=\frac{a_{ik}}{\sum_{k=1}^{n}a_{ik}} \tag{3}$$

である。

- 文献D'_{fj}における単語T_kの出現回数をc'_{fjk}とすると、文献D'_{fj}における相対出現率A'_{fjk}は、

$$A'_{fjk}=\frac{c'_{fjk}}{\sum_{k=1}^{N}c'_{fjk}}\times 100 \tag{4}$$

で求める。

- 単語リストの作り方 p は、s と u の組み合わせ (s,u) である。s は、式 (2) において単語 T_k の i 分野の文献集合 D_i における総出現回数 c_{ik} が 0 であるとき、そのかわりに与えられる $0 \leq s \leq 1$ の値である。u は、後述のアルゴリズムによって単語リストに含める単語を選択するときに用いられる、相対出現率の比 R_{lmk} の下限である。

- 単語リストの作り方 p による分野 i の単語リストを Salton[4] にしたがって重みベクトル $W_i^p = (w_{i1}^p, \cdot\cdot, w_{ik}^p, \cdot\cdot, w_{iN}^p)$ で表わす。この場合、単語 T_k がリストに入っていれば $w_{ik}^p = 1$、入っていなければ 0 である。

- 分野 l、および m の文献集合 D_l、D_m における同じ単語 T_k の相対出現率 A_{lk}、A_{mk} の比 $\frac{A_{lk}}{A_{mk}}$ を R_{lmk} とする。l を固定し、m を 1 から I まで動かしたとき、$R_{lmk} > u$ となる m があれば $w_{lk}^p = 1$ とする。すなわち、l 分野の単語リストに単語 T_k を入れる。

- 文献 D'_{fj} における分野 i の単語リスト W_i^p の総出現率 V_{fij}^p は、

$$V_{fij}^p = \sum_{k=1}^{N} w_{ik}^p A'_{fjk} \qquad (5)$$

である。

- 文献集合 D'_f の中で総出現率 $V_{fij}^p = x$ である文献の個数を $g_{fi}^p(x)$ とする。

- 単語リスト W_i^p の出現率の累積分布関数 $G_{fi}^p(x)(100 \geq x \geq 0)$ は、

$$G_{fi}^p(x) = \frac{\int_{100}^{x} g_{fi}^p(x)dx}{\int_{100}^{0} g_{fi}^p(x)dx} \times 100 \qquad (6)$$

である。

- 累積分布関数 $G_{fi}^p(x)$ の逆関数を、$G_{fi}^{p-1}(y)$ とする。

3 方法の評価

本研究で用いたデータは、学術情報センターの学会発表データベースのうち情報処理学会と日本農芸化学会と土木学会と電子情報通信学会の 1994 年の抄録である。本研究では、ある学会で発表された文献はその分野に属するものとする。そのうち、それぞれ 1500 件を単語リスト作成用とし、残りのそれぞれ 1740 件、1051 件、2424 件、7961 件を単語リストの評価用に用いた。

3.1 3 分野間での方法の評価

情報処理学、農芸化学、土木学の 3 分野間で単語リストの作り方 p を変えて単語リストを作成し、それを用いて評価用文献の属する分野を推定し、この方法の安定性を検証した。p は、$s = \{0.1, 0.5\}$、$u = \{5, 10, 20, 30\}$ のすべての組み合わせで 8 通りである。

分野推定には、$V_{fij}^p (i = 1, \cdot\cdot, I)$ の中で V_{fij}^p が最も大きい分野 i を文献 D'_{fj} の分野と推定する方法を採った。f 分野に所属する文献のうち、この分野判定法により i 分野であると判定された文献数を H_{fi}^p としたとき、推定率 h_{fi}^p は以下の式で求める。

$$h_{fi}^p = \frac{H_{fi}^p}{\sum_{i=1}^{F} H_{fi}^p} \times 100 \qquad (7)$$

文献 D'_{fj} はもともと f 分野に所属する文献であるので、$f = i$ のときは正しい推定であり、$f \neq i$ のときは不正確な推定をしたことになる。

表 1 は、作成方法 p が $s = 0.5$、$u = 10$ のときの単語リストを用いた分野推定の結果である。表の上段は H_{fi}^p であり、下段は h_{fi}^p である。

正しく推定された文献の割合を平均すると、

$$h^p = \frac{\sum_{i=1}^{I} H_{ii}^p}{\sum_{f=1}^{F}\sum_{i=1}^{I} H_{fi}^p} \times 100 \qquad (8)$$

単語リストの作成方法 p に対する h^p の変化を表 2 に示す。表 2 から、作り方 p を変化させても正しい分野推定は 95.9%〜96.6%の間にあり、作り方 p にほとんど依存せず比較的高い確率で正しく推定されることがわかる。

表 1: 分野推定結果の一例 (s=0.5,p=10)

文献の属する分野	推定された分野		
	情報処理学	日本農芸化学	土木学
情報処理学会	1703 (97.9)	4 (0.2)	33 (1.9)
日本農芸化学会	7 (0.7)	1037 (98.7)	7 (0.7)
土木学会	118 (4.9)	41 (1.7)	2265 (93.4)

表 2: 正しく分野推定された文献の割合

作成方法 p		正しく分野推定された文献の割合
s	u	
0.1	5	96.2
0.1	10	96.2
0.1	20	96.5
0.1	30	96.6
0.5	5	96.4
0.5	10	96.0
0.5	20	95.9
0.5	30	96.2

単語リストの作り方 p による単語リストの種類数の変化を表 3 に示す。$s = 0.1$ の場合に比べて、$s = 0.5$ の場合の方が単語の種類数が大幅に少なくなっている。これは、$s = 0.1$ ではほぼ他分野に出現しない単語が約 2 回出現すれば単語リストに含まれるのに対して、$s = 0.5$ では約 6 回出現しなければ単語リストに含まれないためである。また、u が大きくなるにつれて単語の種類が少なくなっている。これは、u の定義からして当然である。

以上のように単語リストの作り方 p を変化させると単語の種類数に大きな変化が見られるが、それにもかかわらず表 2 のような安定した結果が得られる。そこで、各分野の単語リストの内容を検討した結果、以下では $p = (0.5, 10)$ を採用することにした。

3.2 4 分野間での方法の評価

分野数を増やしたときの方法の安定性を実証するために、先の 3 分野に情報処理学と比較的近いと思われる電子情報通信学を加え、4 分野間で単語リストを作成し、3 分野間で作成したときと比較した。

図 1〜3 は、3 分野間で作成した単語リスト中の語の、各分野の文献における出現率の累積分布グラフである。図 4〜7 は、同じく 4 分野間で作成した単語リストを用いた累積分布グラフである。図 1 と 4、2 と 5、3 と 6 をそれぞれ比較すると、電子情報通信学を加えても、情報処理学、農芸化学、土木学の文献におけるこれら 3 分野の単語リスト中の語の出現率の累積分布に大きな差はみられない。これは、新しい分野を加えてもこの方法が安定した結果を与えることを示す。

表 4 は、4 分野間で単語リストを作成したときの種類数である。表 3 の $p = (0.5, 10)$ の行の対応

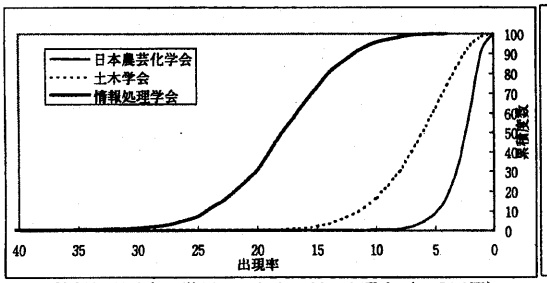

図1　情報処理学会の単語リスト中の語の出現率（３分野間）

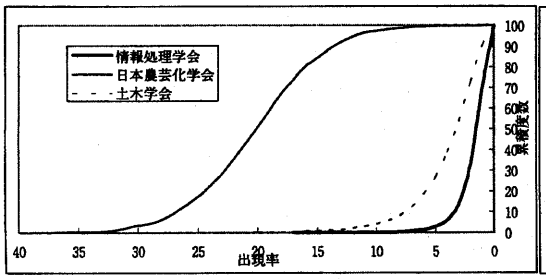

図2　農芸化学の単語リスト中の語の出現率（３分野間）

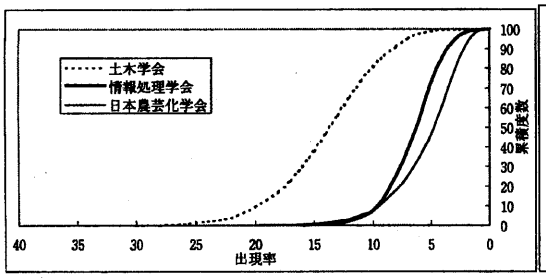

図3　土木学の単語リスト中の語の出現率（３分野間）

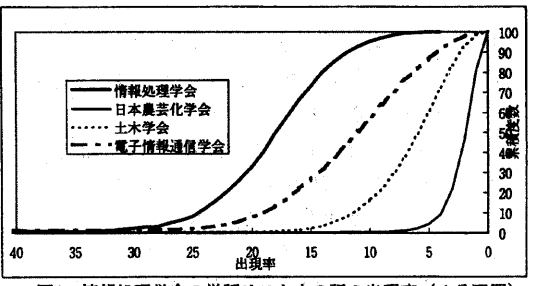

図4　情報処理学会の単語リスト中の語の出現率（４分野間）

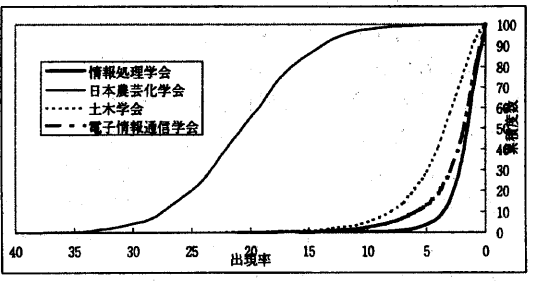

図5　農芸化学の単語リスト中の語の出現率（４分野間）

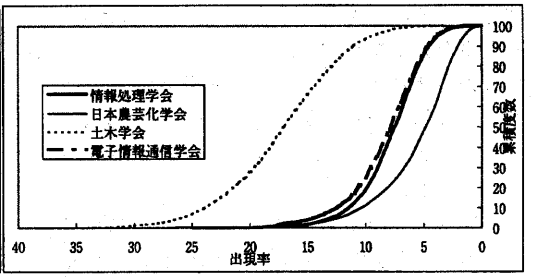

図6　土木学の単語リスト中の語の出現率（４分野間）

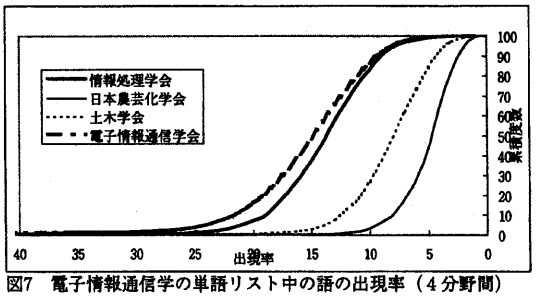

図7　電子情報通信学の単語リスト中の語の出現率（４分野間）

表 3: 3分野間で作成した単語リストの種類数

作成方法 p		単語リストの分野		
s	u	情報処理学	農芸化学	土木学
0.1	5	10718	14811	9288
0.1	10	9918	6504	8789
0.1	20	4815	4106	4220
0.1	30	3116	2979	2775
0.5	5	4008	4588	3638
0.5	10	2156	2138	2062
0.5	20	947	1062	1081
0.5	30	572	745	704

表 4: 4分野間で作成した単語リストの種類数

作成方法 p		単語リストの分野			
s	u	情報処理学	農芸化学	土木学	電子情報通信学
0.5	10	2191	2193	2133	2159

する値と比較すると、最大3.4%の差であり、この結果からもこの方法の安定性が示唆される。

4 文献の分野関連度

分野 i の文献集合を同じ分野の単語リスト中の語の出現率の順に配列する。べつに、分野 f の文献 D'_{fj} 中での分野 i の単語リスト中の語の出現率 V^p_{fij} を求める。この文献を仮に先の配列中においたとき、その何%点に位置するかを、この文献の i 分野との分野関連度とする。式で表現すれば、文献 D'_{fj} の分野 i との関連度 y_{fij} は、

$$y_{fij} = 100 - G^p_{ii}(V^p_{fij}) \qquad (9)$$

となる。文献 D'_{f_1i} と分野 i との分野関連度を「y_{fij}%点」と表わす。このような方法で、文献のそれぞれの分野との分野関連度を求めた結果の例を以下に示す。

000205014
東京23区の配電用変電所の電力需要データを分析し、配電エリアの電力需要曲線の特徴により、住宅地型・オフィス街型・工場地型・繁華街型の4パターンに分類した。そしてどのエリアの電力需要もこの基本4パターンの重ね合わせで表せると仮定し、それぞれの需要パターンの平均需要曲線を最小自乗法で求めた。さらに、地域の特性と電力需要特性を関連づけるために、寄与率という概念を用いて、そのエリアの電力需要に占める基本4需要パターンの割合を求めた。寄与率を用いると、電力需要特性から見た地域特性の評価や、時間ごとのパターン別電力消費量の予測が可能となる。

文献 1

000185094
航空交通流管理システムは、平成6年10月に本格的な運用を開始しようとしている。本システムは、福岡の航空交通流管理センター内に中央処理装置を設置し、全国4ケ所の航空交通管制部および羽田、成田、関西の主要空港事務所に端末装置を設置して、航空交通流管理業務を支援することとなる。本稿は、初期の航空交通流管理システムのシステム構成、機能概要および交通量集計、出

表 5: 文献集合の分野関連度（3分野間）

文献の出所	単語リストの分野		
	情報処理学	農芸化学	土木学
情報処理学会	90	0.1	11.0
日本農芸化学会	0.4	90	9.3
土木学会	8.5	1.3	90

表 6: 文献集合の分野関連度（4分野間）

文献の出所	単語リストの分野			
	情報処理学	農芸化学	土木学	電子情報通信学
情報処理学会	90	0.1	11.9	87.9
日本農芸化学会	0.5	90	8.0	5.5
土木学会	8.2	1.3	90	34.3
電子情報通信学会	44.9	0.6	16.9	90

発制御時刻の算出等の処理方式について述べている。また、航空交通流管理業務の役割にふれてはいるものの、その範囲を限定せず、基本的な意見を述べるにとどめている。今後とも航空交通流管理システムの機能は、航空交通流管理業務の発展とともに追加、変更が加えられ、航空交通の安全と空域の有効利用に寄与するものと期待される。

文献 2

文献1は、土木学会の抄録であり、土木学、電子情報通信学、情報処理学、および農芸化学との分野関連度は、それぞれ49.2%点、34.2%点、2.2%点、0.0%点である。この抄録は、土木学の文献集合中では土木学との分野関連度はほぼ中位にあり、他に電子情報通信学との関連度が比較的強い。

文献2は、電子情報通信学会の抄録であり、土木学、電子情報通信学、情報処理学、および農芸化学との分野関連度は、それぞれ89.6%点、67.6%点、45.1%点、0.0%点である。すなわち、この抄録は土木学との関連度が最も強く、文献と同じ分野の電子情報通信学の文献集合中でも比較的上位になり、農芸化学との関連度はないといえる。

5 文献集合の分野関連度

3節では、ある学会で発表された文献はその分野に属するとして、分野の単語リストを作成した。しかし、実際には図1〜7から明らかな通り、これらの文献の自分野への関連度には分布があり、その分布は分野によって異なる。しかしどの分野 (i) でも、この分野 (i) の単語リスト中の語の出現率順に文献を配列したとき、$G^p_{ii}(x)$ が90までに含まれる文献は、"その分野 (i) の文献らしい文献"とみることができよう。また、分野 (f) の文献であっても、分野 (i) への分野関連度が上のような90までに含まれる文献は、"この分野 (i) と強く関連している"とみなすことができる。このような文献の、文献集合 D'_f 中での割合を文献集合 D'_f の i 分野との分野関連度 Y_{fi} と定義する。式に表わすと、

$$Y_{fi} = G^p_{fi}(G^{p-1}_{ii}(90)) \qquad (10)$$

となる。

3分野間および4分野間で作成した単語リストを用いた Y_{fi} の結果をそれぞれ表5、6に示す。

表6の情報処理学会、日本農芸化学会、土木学会の文献集合とこの3分野との分野関連度に関する部分は表5の対応する値とよく一致している。す

なわち、新しい分野を加えて異なる単語リストを作成しても文献集合の分野関連度は安定しているといえる。

表5から、土木学会の文献集合の、情報処理学への分野関連度が8.5%、情報処理学会の文献集合の、土木学への分野関連度が11.0%であり、相互にほぼ同程度の関わりといえる。それに対し、土木学会の文献集合の、農芸化学への分野関連度は9.3%、日本農芸化学会の文献集合の、土木学への分野関連度は1.3%である。これは日本農芸化学会の文献集合には土木学と関連するものが比較的多いが、土木学の文献集合には農芸化学と関連するものが少ないことを示す。また、全体的にみるとこの3分野は比較的互いに独立な分野といえる。とくに、情報処理学会および土木学会の文献集合の、農芸化学への分野関連度はごく低い。

表6から、電子情報通信学会の文献集合の情報処理学への分野関連度は44.9%、情報処理学会の文献集合の、電子情報通信学への分野関連度は87.9%であり、この2分野間に強い関係があることを示している。さらに、土木学会の文献集合と電子情報通信学との分野関連度も34.3%と相当大きい。

6 おわりに

単語の統計的手法を用いた自動分類、索引語の研究については、Baysian モデルを適用して自動分類を試みたMaron[5]、条件付き確率を用いてタイトルからの自動分類を試みたHamil ら[6]、因子分析による自動分類を試みたBorko ら[7]、専門用語を抽出するために、文献中の単語の分布を2-ポアソンモデルであらわし重みづけをしたHarter[8]などが古くから行われている。日本語の文献に対しても、分野別用語集の選択基準に相対出現率を用いた加藤ら[9]、単語の出現傾向からχ^2値を求め重要語の自動抽出をした長尾ら[10]、漢字の出現頻度情報を用いて日本語文献を自動分類した細野ら[11]の研究がある。また、simple query に relevance feedback を組み合わせ、多数の単語を用いたStanfill ら[12]の研究がある。

検索語の選択や分類に用いる単語の選択に関する研究では、対象となる小さい集合と平均的な集合との単語の出現率の比較で単語を選択する場合が多い。本研究では、分野間での単語の相対出現率の比で選択している。これは、一般的な単語はどの分野にも均等にあらわれると考えるからである。この方法の特色は、分野の数は少なくても、また分野間に重なりがあってもよいこと、単語リスト作成用テキストが多い時に適することである。

本研究では、日本語抄録から大きい分野の単語リストを作成し、その単語リストを用いて、高い精度で文献の分野判定を行った。また、文献の分野関連度および文献集合の分野関連度を求める方法を示した。

参考文献

[1] 石田栄美 ほか. 多数の語を用いた検索質問の作成と評価, 情報知識学会, 47-52, 1996.

[2] 中本賢一. 複数ハッシュふるい分け法の日本語情報システムへの応用. 情報システム研究会. 48-7. 1994.3.

[3] Kigen Hasebe, Ken'ichi Nakamoto, Takeo Yamamoto. An Information Retrieval System on Internet for Languages without Obious Word Delimiters. Proceedings of International Symposium on Digital Libraries 1995. 181-185. 1995.8.

[4] G.Salton,A.Wong,C.S.Yang. A Vector Space Model for Automatic Indexing. CACM. Vol.18. No.11. 613-620. 1975.

[5] M.E.Maron. Automatic Indexing: An Experimental Inquiry. JACM. Vol.8. No.3. 404-417. 1961.

[6] K.A.Hamill,A.Zamora. The Use of Titles for Automatic Document Classification. JASIS. Vol.31. No.6. 396-402. 1980.

[7] H.Borko,M.Bernick. Automatic Document Classification. JACM. Vol.10. 151-162. 1963.

[8] S.P.Harter. A Probabilistic Approach to Automatic Keyword Indexing. JASIS. Vol.26. No.4. 197-206. 1975.

[9] 加藤 緑 ほか. 分野別用語集のための語の選定方法に関する実験的な検討. 第7回情報科学技術研究集会発表論文集. 319-326. 1969.

[10] 長尾 真 ほか. 日本語文献における重要語の自動抽出. 情報処理. Vol.17. No.2. 110-117. 1976.

[11] 細野 公男 ほか. 漢字の出現頻度情報を用いた日本語文献の自動分類. 自然言語処理. Vol.47. No.7. 47-54. 1985.

[12] C.Stanfill. Parallel Free-text Search on The Connection Machine System. CACM. Vol.29. No.12. 1229-1239. 1986.