情　報　学　基　礎　57－5
自　然　言　語　処　理　136－5
（2000．3．21）

文書自動分類のための分野関連語辞書の構成

呉勇　山田祥　岸本陽次郎
日本大学大学院　工学研究科
〒963-8642 福島県郡山市田村町徳定字中河原 1

E-mail:yamada@cscw00.ce.nihon-u.ac.jp
{wuyong,kisimoto}@ce.nihon-u.ac.jp

あらまし　　オフィスで取り扱う文書量の増大に伴い，文書の自動分類に対する需要が増えている．本稿では，あらかじめ分野分類された文書群から名詞を抽出し，分野内および分野間の出現確率を求めて分野関連語辞書を構築し，自動分類を行う．具体的事例として新聞における政治，社会などの 5 分野の自動分類を取上げて検討した結果，この辞書の容量は 72～82%の統計的な圧縮を行い 1,052～3,067 語で済むこと，分類の再現率および精度は約 80%を確保できることを確認した．さらに，クラスター分析手法を使って，辞書間の距離（非類似度係数）と分類誤り率の関係を明らかにしている．

キーワード　　自動分類，分野関連語辞書，新聞記事，クラスター分析

Building a Dynamic Thesaurus for Automatic Classification

Yong WU　Akira YAMADA　Yojiro KISHIMOTO
Graduate School of Engineering,Nihon University

1 Tokusada aza Nakagawara,Tamura,Koriyama,Fukushima 963-8642

E-mail:yamada@cscw00.ce.nihon-u.ac.jp
{wuyong,kisimoto}@ce.nihon-u.ac.jp

Abstract　　A method of building thesaurus, which consist of keywords extracting statistically from classified documents, is reported for automatically classifying documents. Documents are classified to each field by compare their characteristic vectors and the characteristic vectors of the thesaurus. As an example, fives fields of news (politic, society, economy, sport, international) are investigated. The result shows that the classification precision are about 80% and the thesaurus contain 1,052 to 3,067 words. In addition, a relationship of the classification erratum and the distance between the thesaurus of different fields is clarified by using the cluster analysis.

Keywords Automatic Classification，Thesaurus，Newspaper，Cluster Analysis

1.はじめに

計算機の急速な普及によりテキストの電子化が進み，膨大な情報が計算機上でアクセス可能になりつつある．多くの情報があらかじめある分野に沿って整理されているならば，その中から必要な情報を探し出すことは比較的容易である．このことから，情報を分類することは情報へのアクセスを支援する一つの方法と考えることができる[1]．このため，文献などを対象にして文書を自動的に分類する研究が行われている[2]．

筆者らは，名詞の出現確率を抽出して構築した分野関連語辞書を用いて，文書を自動的に分類することについて研究を行っている[3]．本稿では，文書を分類する際に用いる分野関連語辞書の構築と，自動分類方法を評価した結果について述べる．

2.文書の自動分類における従来の研究と課題

従来の自動分類には，あらかじめ人手で分類されている文書（以下，実例文書と略す）を利用する方法がある．分類したい文書が与えられるとその文書と実例文書との類似度を計算し，もっともよく似ている実例文書の分野を求める分野とする．この方法は大量の実例文書を用いた場合にかなりよい精度で分野を推定することができる[1]．また，文書の意味構造を扱わないため処理が比較的簡単である．しかし，大量の実例文書を必要とするため計算機の記憶容量を多く占有し，分類する際に計算時間がかかるなどの問題点が生じる[2]．

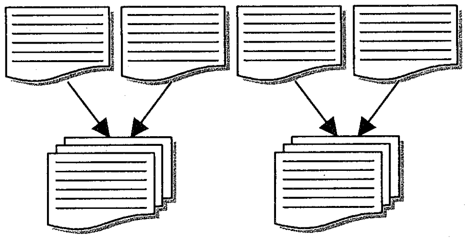
以上のことから，実用的な文書の自動分類の実現を図るために，高い分類効率を維持したまま分野関連語辞書の容量を縮小することを目的とする．

3.分野関連語辞書を用いた文書の自動分類方法

3.1 分野関連語辞書の構築方法

文書を自動分類するために分野ごとに既分類された文書から名詞の種類と出現確率を求め，これらの集合である分野関連語辞書を構築する．文書の自動分類で用いる分野関連語辞書の構築の流れを図１に示す．下記の数字は図１中の数字に対応している．

(1)既分類文書の収集



(2)名詞の抽出と出現回数の計算

~~山に囲まれた~~人口~~約２０００人の~~富山~~県~~山田~~村が、４５８戸の全~~世帯~~に~~パソコン~~を備え、~~電子メール~~で~~回覧~~板を回すなど~~村内~~の~~各家庭~~を結ぶ~~情報交換~~はもちろん、~~インターネット~~で~~世界~~とも~~交信~~しようと~~準備~~を進めている。村が~~アンケート~~をしたところ７０％の~~家庭~~が~~賛同~~し、１０日の村~~議会全員協議~~会に~~報告~~された。今年中には~~実現~~する~~見通し~~だ。~~

名詞	出現回数
円相場	526
ドル高	260
日債銀	203
：	：

(3)名詞の出現確率の算出

名詞	出現確率	
	$Y_{ik}(\times 10^{-2})$	$X_{ik}(\times 10^{-2})$
円相場	0.38	80
ドル高	0.18	100
日債銀	0.14	70
：	：	：

図１ 分野関連語辞書の構築の流れ

(1)既分類文書の収集

分野関連語辞書に利用する文書は，あらかじめ分野ごとに分類されている文書を収集する．分類されている文書の分野が分野関連語辞書の分野となり，自動分類の際の分野となる．ここで，収集する文書には分野関連語辞書を構築するのに十分な文書量を有することが必要である．

(2)名詞の抽出と出現回数の計算

分野ごとに収集した既分類文書から名詞を抽出する．ここでは，形態素解析プログラムChaSen[4][5]を利用して名詞を抽出した．さらに，抽出された名詞から単漢字および数字を取り除き，名詞の出現回数を計数し，出現回数の多い順に並べる．

(3)名詞の出現確率の算出

特定の分野に集中して出現する名詞は，分野を特徴付けるため，分野kの名詞iが全分野の名詞iに占める出現確率X_{ik}は式(1)で表される．

$$X_{ik} = \frac{Y_{ik}}{\sum_k Y_{ik}} \quad \left(\sum_k X_{ik} = 1\right) \tag{1}$$

ここでY_{ik}は名詞iが分野kの全名詞に占める出現確率であり，式(2)で表される．

$$Y_{ik} = \frac{w_{ik}}{\sum_i w_{ik}} \quad \left(\sum_i Y_{ik} = 1\right) \tag{2}$$

但し，w_{ik}は分野kにおける名詞iの出現回数である．

一般に分野関連語辞書は辞書中の名詞i(但し，$i=1,\cdots,n$)を用いて，式(3)で表される．

$$x_k = \left(X_{1k}, X_{2k}, \cdots, X_{nk}\right) \tag{3}$$

n：分野関連語辞書の異なり名詞数

式(3)は名詞空間のベクトルとして表すことができる[6][7]．名詞空間は，分野関連語辞書の特徴空間で，ベクトルx_kは分野の特徴ベクトルである．

3.2 文書の自動分類方法

分野関連語辞書の構築と同様にChaSen[5]を使い，分類対象の文書から名詞を抽出し，出現確率S_iを計算する．出現確率S_iは式(4)で表される．

$$S_i = \frac{si}{\sum_i si} \tag{4}$$

si：分類対象文書中の名詞iの出現回数

分野の特徴ベクトルと同様に，文書の特徴ベクトルは式(5)によって表される．

$$u = \left(S_1, S_2, \cdots, S_n\right) \tag{5}$$

n：分野関連語辞書の異なり名詞数

図2に文書を自動分類するための文書と分野の特徴ベクトル空間を示す．文書の自動分類は，図2に示す特徴空間において，文書の特徴ベクトルuにもっとも近い分野の特徴ベクトルx_kを求める．すなわち，分野kの特徴ベクトルx_kと分類文書の特徴ベクトルuの内積値がもっとも大きい（ベクトル間の角度がもっとも小さい）場合を該当分野とする．分野関連語辞書と文書の内積値を求めるには式(6)で表される．

$$C_k = u \cdot x_k = \sum_i \left(S_i \cdot X_{ik}\right) \tag{6}$$

このためC_kは分野kにおける分類文書の出現確率に相当する．

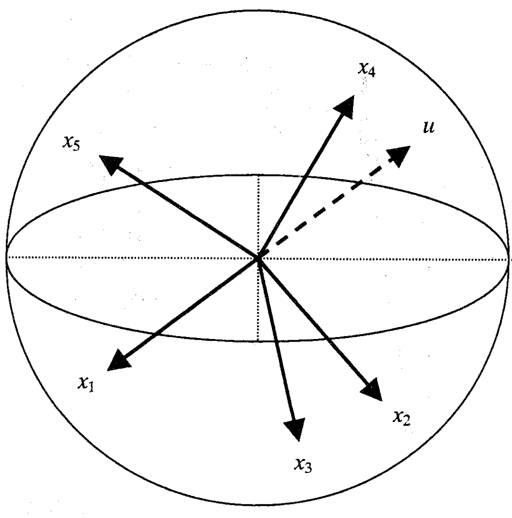


図2 自動分類のための特徴空間

4.分野関連語辞書を用いた分類結果および検討

4.1 分野関連語辞書の構築期間による分類結果

前節で述べた自動分類方法を評価するため，実際に分野関連語辞書を用いて，自動分類を行う．分野関連語辞書に用いる文書および分類対象の文書には，asahi.com[8]の新聞記事を利用した．新聞記事は“経済”，“社会”，“政治”，“国際”および“スポーツ”の5分野にあらかじめ分類されている．そのため，新聞社側の分類結果と自動分類の結果が一致すれば，正しい分類ができたといえる．その際，文献検索における評価を参考に，分類結果の評価は式(7)および式(8)で表すことにする．

$$分類再現率=\frac{自動分類された正分類記事数}{新聞社による分類記事数} \quad (7)$$

$$分類精度=\frac{自動分類された正分類記事数}{自動分類された記事数} \quad (8)$$

asahi.com の 1997 年 10 月～12 月までの 3 ヶ月間の新聞記事を用いて，1998 年 1 月～2 月までの新聞記事 2617 件の自動分類を行った結果を表 1 に示す．表 1 において，縦軸は新聞記事本来の分野を表し，横軸は新聞記事が分類された分野を表している．表 1 より，分類再現率は 75%～99%，分類精度は 82%～91%で，自動分類が可能であることが確認された．なお，表 1 は分野関連語辞書を構築するための新聞記事の収集期間が 3 ヶ月のときの分類結果であるが，構築期間を 1 ヶ月から 21 ヶ月まで変化させたときの分類再現率の結果を図 3 に示す．図 3 において，縦軸は分類再現率，横軸は新聞記事の収集期間を表す．図 3 より，3 ヶ月以上の新聞記事を用いて分野関連語辞書を構築すれば，分類再現率は安定することがわかる．以上のことから，分野関連語辞書を構築するには，少なくとも 3 ヶ月間の新聞記事を収集する必要がある．

表 1 分野関連語辞書を用いた自動分類の分類結果

新聞社による分野		分類された分野[件]					分類再現率[%]
元の分野	記事数[件]	経済	社会	政治	国際	スポーツ	
経済	510	458	14	19	18	1	90
社会	556	33	450	18	14	41	81
政治	507	27	33	381	63	3	75
国際	534	15	44	14	456	5	85
スポーツ	510	0	6	0	1	503	99
記事数[件]	2617	533	547	432	552	553	—
分類精度[%]		86	82	88	83	91	

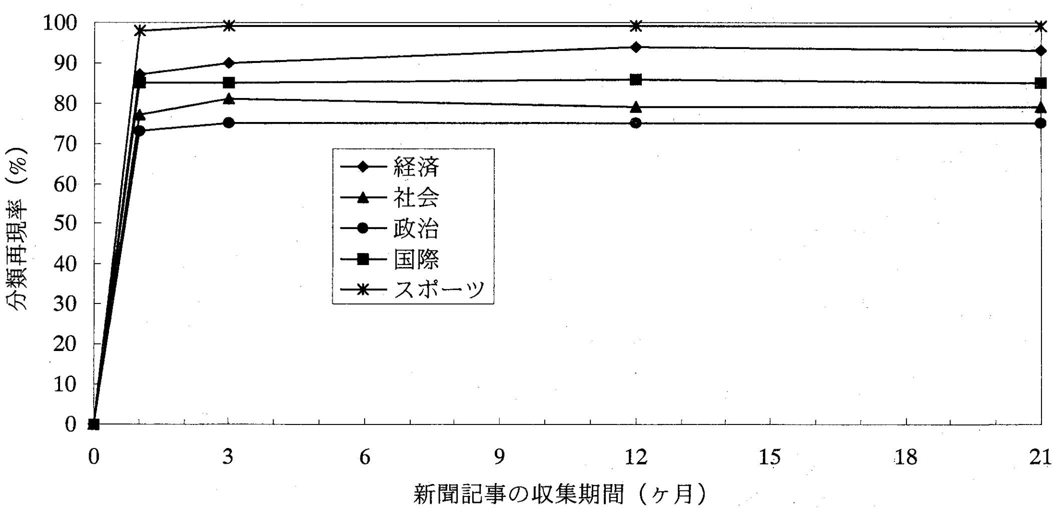


図 3 新聞記事の収集期間と分類再現率の関係

4.2 分離辞書による自動分類の検討

分野関連語辞書は，少ない容量で高い分類効率が求められる．分野関連語辞書を構築するために抽出した名詞は，出現確率の高い名詞と出現確率の低い名詞が存在する．このため，出現確率の高い名詞のみを分野関連語辞書の構築に用いれば，計算機の記憶容量を抑えることができると考えられる．そこで，出現確率 Y_{ik} の高い名詞から出現確率 Y_{ik} を積分し，積分した値がある閾値になるまでの名詞を用いて分野関連語辞書を構築する．

上記の構築方法で文書を自動分類した際の分類再現率を図 4 に，分類精度を図 5 に示す．なお，辞書の構築期間，および分類対象の文書は 4.1 節と同じ条件とする．図 4 において，縦軸は分類再現率，横軸は出現確率 Y_{ik} の積分値を表す．横軸は出現確率 Y_{ik} を積分するため 100%で全名詞を利用したことになる．図 4 および図 5 より，出現確率 Y_{ik} の積分値が 80%に達すると，分類再現率および精度がほとんど変化しないことがわかる．ここでは，出現確率 Y_{ik} の積分値が 80%までの単語を用

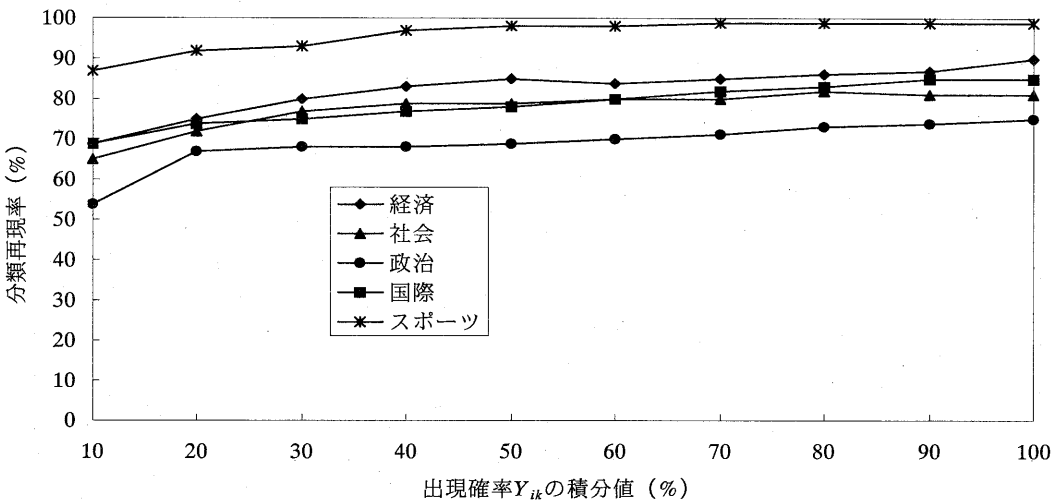


図 4 出現確率 Y_{ik} の積分値と分類再現率の関係

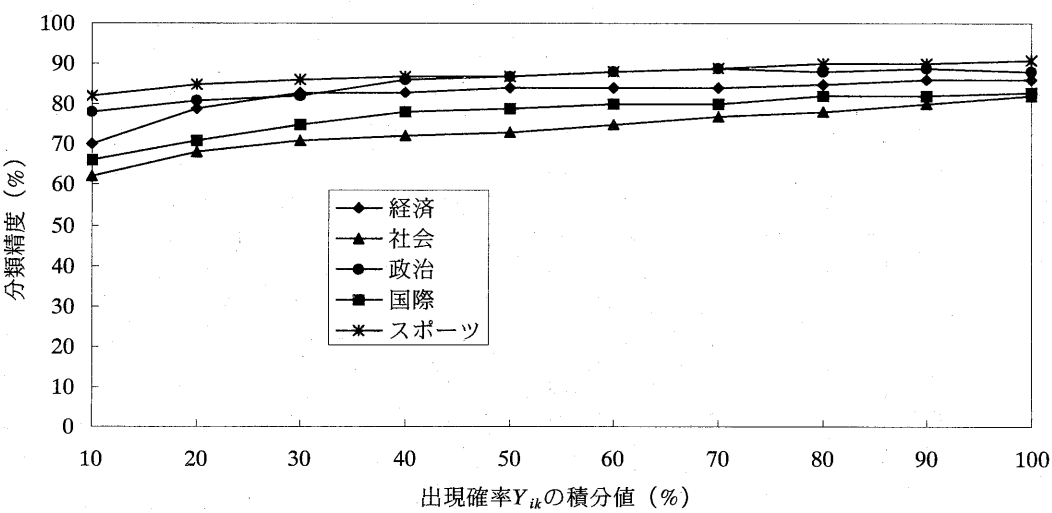


図 5 出現確率 Y_{ik} の積分値と分類精度の関係

いて構築した辞書を分離辞書とする．

各分野における分離辞書の異なり名詞の種類数を表2に示す．分野関連語辞書と比べると，辞書容量が72%～82%縮小されていることがわかる．また，表3は分離辞書を用いて，自動分類を行った結果で，表1と比べるとほぼ同じである．以上のことから分離辞書を用いることで，分類再現率および精度を維持したまま辞書の容量を抑えることができた．ここで，分類再現率および精度の両特性には，他分野との関わりが疎な分野の値は高く，他分野との関わりが密な分野の値は低くなる特性が見られる．そこで，分野ごとの分類再現率および精度のばらつきを明らかにするため，分野間の関連性について検討する．

表2 新聞記事3ヶ月間から抽出した名詞数

		経済	社会	政治	国際	スポーツ
記事数[件]		798	902	802	847	858
名詞数[個]	分野関連語辞書	6027	10958	5699	7472	6396
	分離辞書	1320	3067	1052	1619	1495

表3 分離辞書を用いた自動分類の分類結果

新聞社による分野		分類された分野[件]					分類再現率[%]
元の分野	記事数[件]	経済	社会	政治	国際	スポーツ	
経　済	510	440	28	22	18	2	86
社　会	556	30	456	15	8	47	82
政　治	507	26	41	368	69	3	73
国　際	534	19	56	11	443	5	83
スポーツ	510	0	6	0	0	504	99
記事数[件]	2617	515	587	416	538	561	—
分類精度[%]		85	78	88	82	90	

5.分野関連語辞書の特性

本節では，分野関連語辞書の特性を調べるためにクラスター分析の手法[9]を用いて，各分野関連語辞書間の類似度を調べる．さらに，辞書間の距離という概念を導入するとともに，辞書間の距離と自動分類における誤り率((自動分類された誤分類記事数)／(自動分類された記事数)で表し，以下,分類誤り率と略す)の関係を明らかにする．

各分野辞書間の類似度を計算するため，分野関連語辞書の特徴空間(名詞空間)を属性空間とし，名詞の出現確率X_{ik}を属性値と定義する．ここでいう属性空間とは，属性値が存在する領域を表し，属性値は分野関連語辞書の特性を表している[9]．類似度は分野関連語辞書の特徴ベクトルの内積

$$c_{kl}=\frac{\sum_i\left(X_{ik}*X_{il}\right)}{\sqrt{\sum_i X_{ik}^2\sum_i X_{il}^2}} \tag{9}$$

で表す．類似度c_{kl}より分野$_k$の辞書と分野$_l$の辞書間の距離d_{kl}（非類似度係数）は

$$d_{kl}=1-c_{kl} \tag{10}$$

となり，距離d_{kl}が大きいほど分野間の相関性が低いことを表す．

各分野の辞書間の距離を式(10)より計算した結果を表4に示す．表4の結果より，「群平均クラスター化法」(UPGMA) [9]を使って，各分野の辞書に対してクラスタリングを行った結果の樹形図を図6に示す．図6より，“社会”および“政治”は他分野の辞書との距離が小さく，相関性が高い．一方，“スポーツ”は他分野の辞書との距離が大きく，相関性が低いことがわかる．

表4 分野関連語辞書間の距離の結果

	経済	社会	政治	国際	スポーツ
経済	—	—	—	—	—
社会	0.88	—	—	—	—
政治	0.84	0.89	—	—	—
国際	0.89	0.86	0.89	—	—
スポーツ	0.95	0.91	0.95	0.93	—

次に，表3より求めた各分野の分類誤り率を表5に示す．図7は，表4および表5から求めた分野間距離と分類誤り率の関係を示したものである．両者の相関がほぼ取れており，分類誤り率を約5%以下に抑えるには距離が約0.9以上必要であることがわかる．

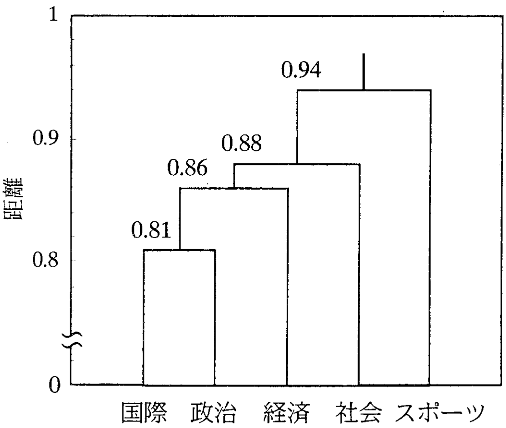


図6 各分野関連語辞書の樹形図

表5 分離辞書を用いた自動分類の分類誤り率

	経済	社会	政治	国際	スポーツ
経済		4.77	5.28	3.34	0.35
社会	5.82		3.60	1.48	8.37
政治	5.04	6.98		12.82	0.53
国際	3.68	9.54	2.64		0.89
スポーツ	0	1.02	0	0	

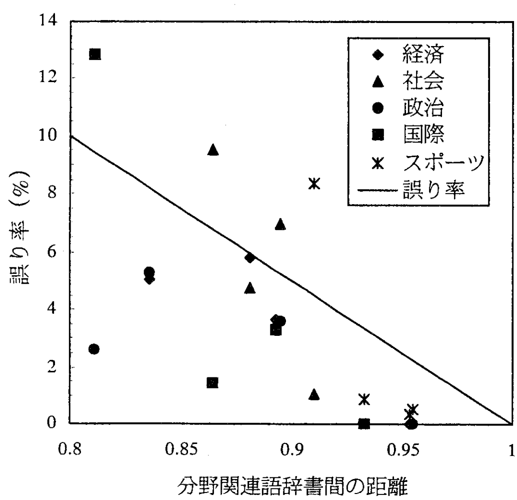


図7 分野関連語辞書間の距離と分類誤り率

6.おわりに

本研究では，新聞記事などの文書の自動分類を行うために，名詞の出現確率を抽出して分野関連語辞書を作成する方法について検討した．分野関連語辞書の構築では，辞書容量の縮小が重要な課題であるが，名詞の出現確率の上位名詞約3割を用いた分離辞書を構築することで，分類再現率と精度は約80%以上を得られることがわかった．また，文書の分類誤りが生じる原因は各分野の辞書間の距離が近いことが原因であり，分類誤り率を5%以下に抑えるには距離が0.9以上必要であることを明らかにした．

今後は，分類再現率および精度に影響を与えている名詞を抽出すること，また構築した分野関連語辞書の使用期限について検討することがあげられる．

参考文献

[1] 長尾他，“自然言語処理”，岩波書店，1996.4

[2] 渡辺他，“χ^2法を用いた重要漢字の自動抽出と文書の自動分類”，情報処理学会研究報告，95-FI-39，pp.25-32，1995

[3] 山田他，“分野関連語辞書を使った文書の自動分類方法”，電子情報通信学会信学技報，OFS98-74，pp.7-12，Mar.1998

[4] 松本他，“日本語形態素解析システムJUMAN”，NAIST TechnicalReport，NAIST-IS-TR94025，July.1994

[5] http://cactus.aist-nara.ac.jp/lab/nlt

[6] G.Salton，“Introduction to Modern Information Retrieval”，McGraw-Hill，NewYork，pp.260，1983

[7] Salton,G.，et al，“A・Theory of Term Importance in Automatic Text Analysis”，J.of American Society for Information Science，vol.26，no.1，pp.33-44，1975

[8] Asahi News Paper flash，http://www.asahi.com/flash/flash.html/

[9] H.C.Romesburg，“実例クラスター分析”，内田老鶴圃，東京，1992