自然言語処理 128−23
（1998. 11. 6）

文書頻度を用いた論文データベース内の文書の分類に関する検討

木付 英士* 小舘 亮之** 朱 青** 浦野 義頼*** 富永 英義*
*早稲田大学大学院 理工学研究科
**通信・放送機構 早稲田リサーチセンター
***早稲田大学 国際情報通信研究センター

(抄録)
現代社会の電子化及び情報化の影響からデータベース内のコンテンツを整理することはユーザへの大きな負担となりつつある。ここでは文書データベース内の文書の分類を行うシステムへの需要が高まっていることに注目する。そこで本稿では、知識を持たない状態で文書を理解し、分類するシステムについて検討する。またそこから得られた知識を以後の分類・検索処理に役立てるシステムについても検討する。その手法として、一般には単語の特殊度を示す指標である単語の文書頻度を用いる手法を提案する。シミュレーション実験を行いどのような知識が得られるかについて、特に文書間強度や単語間強度について検討する。

A Study Of The Document Categorization Using Document Frequency

Hideaki KIZUKI* Akihisa KODATE** Qing ZHU** Yoshiyori URANO*** Hideyoshi TOMINAGA*
*Graduate School of Sci. & Eng., WASEDA University
**Waseda Research Center, TAO, JAPAN
***Global Info. and Telecomm. Institute WASEDA University

(ABSTRACT)
To classify contents of database became a burden for user cause of digitization and surfeit of information. We notice the demand of the system categorizing documents in document-database. In this paper, We examine the system understanding and categorizing documents without the knowledge, and the system categorizing and searching the documents using the knowledge from the former. We propose the way using the document frequency. And we examine the knowledge obtained from the simulation of this way, especially the similarity between documents and between terms.

1. はじめに

近年、社会の情報化に伴いデジタル情報の流通が急増している。流通量の増加に伴い各データベース内に蓄積される情報量も増加している。データベース内のコンテンツを使いやすいような形で蓄積するためには、インデックスを付与するなどの整理作業が必要となる。現状のキーワードによるサーチでは必ずしも必要としているコンテンツを得ることが出来ない。そこで蓄積されたコンテンツを解析して真に有用なコンテンツを選び出す作業が必要になる。

そこでデジタル文書の自動分類システムを提案する。このシステムは知識を持たない状態から複数のデジタル文書を解析し分類し、その時得た情報を知識として学習する。その知識を利用し、新たな追加文書の分類を効率的に行う。これを図 1 に示す。

このシステムの実現のためにはどのようなアルゴリズムで分類し、どのような情報を知識として学習するべきかを検討する必要がある。人がこの作業を行う場合、それらの内容を理解する知識が必要になり，また文書数が増えればその作業は容易ではなくなる。従来の技術を用いた分類作業では tf*idf 法を用いるのが一般だろう。しかし、この手法は特殊な単語を優先的に検出するアルゴリズムで、極端に特殊な小さなカテゴリを抽出する可能性が高い。そこで本研究では、複数文書に対し文書頻度(Document Frequency)を用いて数例の文書の

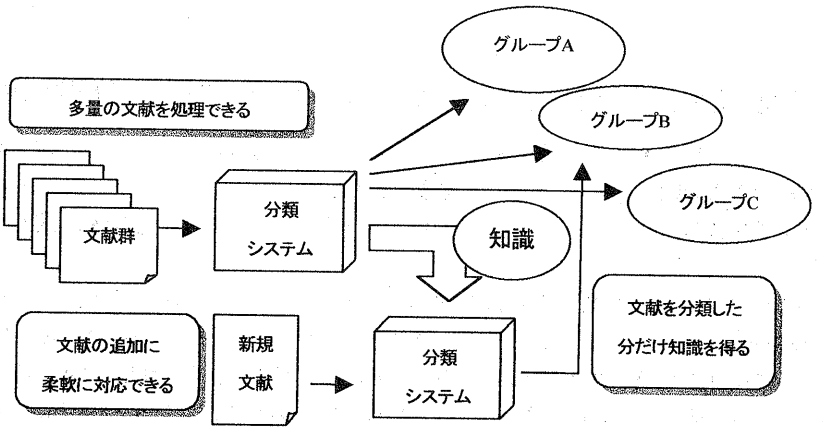

図1 提案システム

分類パターンを例示するアルゴリズムを提案する。この手法を適用することによって各カテゴリに検索に適当な数の文書を配置し、そのカテゴリに合わせたキーワードを抽出出来る。またその分類結果から得られる情報について検討する。これによって分類に必要な知識、つまりシステムが学習するべき知識、について評価・検討することが出来る。

2. 提案手法

まず本研究における単語頻度(Term Frequency)と文書頻度について次の 2 式で定義する。

$$TF(t,d)=\frac{\text{文書d内の単語tの出現回数}}{\text{文書d内の全単語数}}$$

$$DF(t)=\frac{\text{単語tを含む文書数}}{\text{全文書数}}$$

提案手法では全文書を N 個のカテゴリに分類することとする。N は事前に決定しておく必要がある。各カテゴリはそれぞれ異なった一単語t_k(k=1,2,...,N)を含み、極力重ならないように、また極力全ての文書が N 個のカテゴリのいずれかに属するように分類する. 図 2 を用いてこれを説明する。全文書を図のように 3 つのカテゴリに分類する際に複数のカテゴリにまたがる部分(黒い部分)がなるべく小さくなるように、またいずれのカテゴリにも含まれない部分(白い部分)がなるべく小さくなるように分類する。

これを式に表すと以下の様になる。

$$\begin{cases} DF(t_1)+DF(t_2)+\cdots+DF(t_N)\approx 1 \\ DF(t_1\,|\,t_2\,|\cdots|\,t_N)\approx 1 \end{cases}$$

ここで下の式の括弧内に複数の単語が入っているのは、そのいずれかが含まれている、という意味である。この 2 式を条件(1)、条件(2)とする。この条件を満たす単語の組を全て検索する。

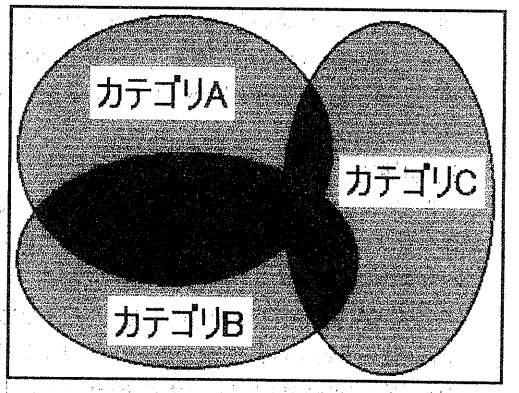

図2 分類ルール

複数の分類パターンが得られたとして、その結果から文書間強度、単語間強度を推定する。

フローチャートの説明

図 3 のフローチャートを用いてアルゴリズムを順を追って説明する。まず、①各文書を形態素解析

によって単語毎に分割し、名詞のみを抽出する。助詞などは共通語にあたるため、本手法には適さない。次に②その各単語を文書毎にカウントして単語頻度を算出する。理由は後述するが、このままでは全文書での異なり語の数が多すぎる。そこで、③単語頻度に下限を設け、ある程度頻出している単語のみを情報として扱う。④その選び出された単語に対して文書頻度を算出する。このとき既に情報削減(③)をされているので本来の文書頻度とは多少異なる。⑤条件(1)、条件(2)を満たす単語の組み合わせを求める。⑥求められた組み合わせから文書間強度、単語間強度を推測する。

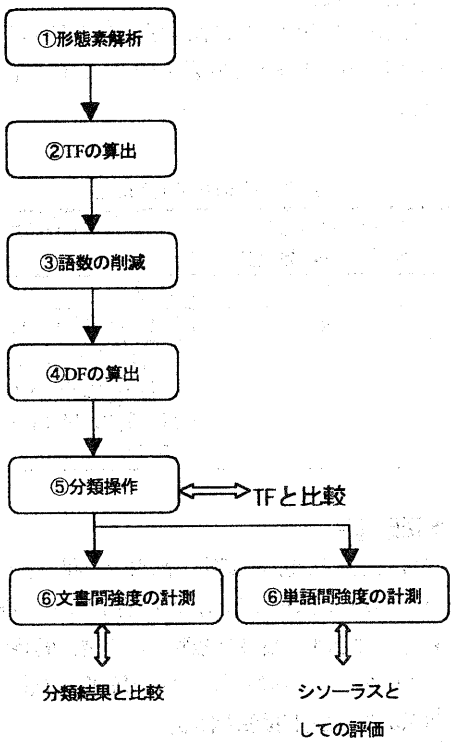

図3 フローチャート

ここで異なり語の数を減らす必要性について言及する。例えば全文書内の異なり語数が300語だったとする。3つのカテゴリに分類するとすると総組み合わせ数は

$$_{300}C_3 = 300 \times 299 \times 298 \approx 2.7 \times 10^7$$

となる。たった300語でも10^7オーダーの検索を行う事になる。その為、どの程度検索精度に影響するかは未知数であるが今回のシミュレーションでは異なり語を減らすこととし、その手法として文書頻度を適用する。

3. シミュレーション実験

実験標本

実験標本として18件のネットワーク用語の解説文を用いる。表4にその用語とファイルサイズを示す。ファイルはすべてShift-JISで書かれたテキストファイルである。

表4 実験標本

番号	用語	ファイルサイズ [bytes]
0	ATM	23,544
1	DBMS	4,370
2	DNS	4,381
3	Eメール	14,918
4	イーサネット	11,561
5	HDLC	4,532
6	ハブ	30,000
7	IEEE802	7,382
8	インターネット	22,633
9	IPX	9,812
10	ISDN	10,486
11	PPP	5,203
12	リピータ	2,169
13	ルータ	13,297
14	SMTP	3,285
15	TCPIP	17,955
16	UNIX	10,814
17	XWINDOW	3,497

形態素解析

形態素解析には茶筌[1]を用いる。文書を単語単位に分解して名詞のみを抽出する。

計算処理結果

各文書の名詞の単語頻度を求め、以下の条件を満たす単語のみを抽出する。

$$TF(t_m, d_n) > 0.012$$

この0.012という数値はチューニングを繰り返した

結果、適当と思われたので採用した値である。ここまでで絞られた単語に対して、全文書に対する文書頻度を求める。ここまでの結果の一部を表 5 に示す。

表 5 計算結果(すべて情報削減した後の値)

情報の種類	値
異なり語の最小数	7(文書 0)
異なり語の最大数	20(文書 12)
TF の最小値	0.012
TF の最大値	0.080
全文書の異なり語数	112
DF の最小値	0.056
DF の最大値	0.611

表 6 分類結果

カテゴリを代表する用語	含まれる文書
ネットワーク	0 4 6 7 8 9 11 12 13 14 15
データベース	1 2
ユーザ	3 8 14 16 17
ネットワーク	0 4 6 7 8 9 11 12 13 14 15
サーバ	1 2 9 17
ユーザ	3 8 14 16 17
ネットワーク	0 4 6 7 8 9 11 12 13 14 15
サーバ	1 2 9 17
システム	3 14 16 17
ネットワーク	0 4 6 7 8 9 11 12 13 14 15
データ	1 5 7 15
システム	3 14 16 17
こと	0 1 2 6 10 11 13 15 16 17
使用	3 4 12
送信	3 5 9 13 14
■	1 4 5 6 11 12 15 16 17
送信	3 5 9 13 14
サービス	7 8 10

分類操作

ここまでの結果を用いて、分類操作を行う。抽出された 112 個の単語を用いる。今実験では 3 つのカテゴリに分類することとする(N=3)。このとき、条件(1)、条件(2)の許容範囲はいずれも±0.15 とする。検索の結果 18 通りのパターンが検出され、その内 6 例を表 6 に示す。

表 6 を見て分かるように異なった分類パターンでも同一のキーワードでカテゴライズされているものがある。よってカテゴリのパターンは 3×18 通りより少なくなって、16通りである。この 16 カテゴリに対して次の測定を行う。

文書間強度の測定

上記の 16 カテゴリに対し、任意の 2 文書が同一のカテゴリに含まれる回数を測定し、その結果を表7に示す。

表 7 文書間強度測定値

同一カテゴリに含まれる回数[回]	文書の組合せ
5	4-12
4	5-15,11-15,16-17
3	1-2,1-15,1-17,,3-14, 6-11,6-12,6-15,11-13, 11-17,13-14,14-17, 15-17

単語間強度

次に文書間強度の測定で高得点を得た組合せのキーワードは何であったかを検索し、その結果を表 8 に示す。これは任意の 2 文書に同時に含まれるキーワード群であり、この単語間に何らかの関係があることが推測される。

また、3 回以上同一カテゴリに含まれた文書に対し、上記の単語間強度の測定を行ったところ、異なるキーワード群に同時に含まれる2単語がある。その単語の組と回数を測定し、表 9 に示す。

表8 単語間強度測定値(1)

文書の組合せ	キーワード群
4-12	ネットワーク、使用、ケーブル、セグメント、■
5-15	データ、アドレス、プロトコル、■
11-15	ネットワーク、こと、プロトコル、■
16-17	システム、ユーザ、こと、■
以下多数	以下多数

表9 単語間強度測定値(2)

検出回数[回]	単語の組合せ
8	■-こと
6	ネットワーク-■, プロトコル-こと
5	ネットワーク,こと
4	ネットワーク-プロトコル, ■-プロトコル
3	ネットワーク-セグメント, セグメント-■
2	■-データ,こと-ユーザ, こと-サーバ, システム-ユーザ

4. 考察

計算結果に対する考察

単語を名詞に限定して更に単語頻度に下限を設けて抽出した結果、表 5 が得られた。文書 0 と文書 12 を比較してファイルサイズは前者の方が大きいのに異なり語は後者の方が多かった。これは単語頻度によるフィルタリングの影響が考えられる。ファイルサイズが大きい文書に、単語頻度の低い単語が多く現れたことが原因と考えられる。また、文書頻度の最大値が低かった理由は、名詞に限定したフィルタリングの影響からでこれはそうなることを想定した結果である。

分類結果の評価

分類パターンに関しては生成される中間ファイルをチューニングの指標としてきたが、結果として 18 通りのパターンが出力された。まず気が付くことは、各カテゴリ内の文書数にやや偏りがあることである。その偏りの影響から、一般的なキーワードや、カテゴリに含まれる文書内ではあまり重要でないキーワードが抽出されているカテゴリもある。しかし、全体的に見て関連のあると思われる文書が同一のカテゴリに含まれている。

また、キーワードの中に「こと」や「■」が含まれている。「こと」に関しては一般的な名詞であり、今後、対策を練る必要があると考えられる。「■」は形態素解析の辞書が名詞扱いをした結果である。これは標本の母集団が大きくなれば影響は小さくなると思われる。

単語頻度と比較

実験の結果出力された 16 のカテゴリを代表するキーワードに関して単語頻度を調べ、その結果を表 10 に示す。

表 10 キーワードと単語頻度

キーワード	単語頻度範囲
ネットワーク	0.012～0.029
データベース	0.021～0.036
ユーザ	0.015～0.027
システム	0.019～0.037
サーバ	0.015～0.045
送信	0.015～0.031
チャネル	0.013～0.025
データ	0.012～0.043
アドレス	0.013～0.024
こと	0.013～0.026
使用	0.016～0.019
プロトコル	0.012～0.032
ケーブル	0.026
セグメント	0.017～0.036
サービス	0.012～0.030
■	0.013～0.036

表 5 の情報と比較すると抽出されたキーワードの単語頻度はあまり高くないことがわかった。

文書間強度の評価

表7の結果の上位について見てみる。4-12 の組はいずれもネットワークのハードウェア的な用語の説明であり、5-15、11-15はプロトコルに近い用語、16-17 は OS に近い用語の説明であることがわか

る。このように判断すると本手法はユーザの知識に近い知識を得、分類をしていると言える。また回数を得点としたところ、最大値が5と低い。よって、知識学習には情報量が少なかったと考えられる。

単語間強度の評価

まず表8の結果から前述の文書のタイトルと同様にハードウェア的なキーワード、ソフトウェア的なキーワードが同一の組に含まれていることが分かる。

また表9の結果から、共起しやすい単語の組が高得点を出していることが分かる。「■-こと」の組はいずれも、今回の文書群内では一般的に用いられている。「ネットワーク-プロトコル」の組に関しても「プロトコル」と言う単語が出れば「ネットワーク」という単語が現れるのは主観的にも予測がつく。同時に一般的な用語は、情報として不要であるが削除が容易ではないことも表9を見て分かる。

5. まとめ

以上の結果をまとめる。①まず本手法による分類結果はユーザの主観に近い分類をすることが出来た。結果が複数出てしまうのでそれを如何に整理するかは今後の課題となる。②分類の過程で削除しきれない一般語があった。これは例外処理として削除するべきなのか、統計的に切り分けることが出来るものなのかは未解決である。③分類に必要な情報量と知識学習に必要な情報量は異なるであろうという推測を得た。検索処理として有効である為には、情報量や検索時間が限られる。しかし、フィルタリングされた情報は知識学習にはやや不足気味であった。④単語間強度の測定実験から関連用語のリストを作れる可能性を見つけた。今回の実験ではそれをモデル化するには至らなかったが、新たな辞書作成アルゴリズムの可能性を見つけた。⑤適当なフィルタを検討する必要があることが分かった。今回の2つのフィルタ、名詞フィルタとTFフィルタ、による情報の損失が実験後半の情報収集にどれだけの影響を与えるか検討する必要がある。

今後の予定

今後、まず検討しなければならない課題として以下の様な事が考えられる。

- ファイルサイズと単語頻度の分布の関係
- 品詞と単語頻度、文書頻度の分布の関係
- キーワードと単語頻度の分布の関係
- 知識学習に適した情報量
- 文書間強度の評価手法
- 単語間強度の評価手法
- 一般的な用語を削除するフィルタ
- 情報の増加による分類処理への影響

これらの課題を調査して文書分類システムの検討することに加え、知的類似検索システムへの情報の転用も検討していく予定である。

参考文献

[1]松田透,小川泰嗣:"統計的確立に基づくキーワード重要度算出モデル",自然言語処理 115-117,情報処理学会(1996)

[2]奈良先端科学技術大学院大学:茶筌 for WINDOWS V1.00,形態素解析ソフト(1997)

[3]木付英士,小舘亮之,浦野義頼,富永英義:"文書頻度を用いた文書の分類手法の検討",ソサイエティ大会 D-4-9,電子情報通信学会(1998)