原　著

医療情報学　24 (1), 2004: 35–44　35

テキストマイニングによる退院サマリー自動分類の試み

Classification of Discharge Summaries by Text Mining

小野　大樹[1]　高林克日己[2]　鈴木　隆弘[2]
Hiroki ONO　Katsuhiko TAKABAYASHI　Takahiro SUZUKI
横井　英人[2]　井宮　淳[3]　里村　洋一[2]
Hideto YOKOI　Atsushi IMIYA　Youichi SATOMURA

電子化した退院サマリーからテキストマイニングによって疾患別の重要語を抽出し，これをもとに退院サマリーの文章から疾患名を特定できるかについて検討した．千葉大学医学部附属病院病院情報システムに保存された退院サマリーの中から臓器別の代表 13 疾患 4,317 症例を選んで文書を形態素解析し，ベクトル空間モデルを用いて 7,918 の診療に関連する索引語を抽出した．その抽出結果から疾患毎の重要語を選出した．次に 390 症例の退院サマリーから疾患名を特定する実験を行った．その結果 390 症例中 290 症例 (74%) が退院時サマリーの診断と一致した．さらにこれらの例に対してデンドログラムを作成して分類の視覚化を試みたところ，医学的に順当な結果を得た．以上の結果からテキストマイニングにより診療文書内容から疾患の特定，類似症例の検索，さらに疾患の再分類などの可能性が示唆された．
（キーワード：退院サマリー，ベクトル空間モデル，tf×idf 法，データマイニング，テキストマイニング，知識発見）

Objectives: To study the ability of text mining technique for the selection of specific words related to diagnosis and to distinguish the diseases of discharge summaries. ***Materials and methods:*** 4,317 discharge summaries in Chiba University Hospital were selected out of 13 representative diseases. Diagnosis related terminological words were extracted by morphological analysis. Thus, the diseases were compared with each other using tf×idf vector space model and important specific words for each disease were selected. Furthermore, we applied the vector space model for new cases and indicated the vector by a radar chart. ***Results:*** 7,918 words were selected from cases and 74% of 390 cases were properly diagnosed. The maximum-tree problem and dendrogram method demonstrated reasonable

[1] 千葉大学大学院自然科学研究科
〒263–8522　千葉県千葉市稲毛区弥生町 1–33
[2] 千葉大学医学部附属病院企画情報部
〒260–8677　千葉県千葉市中央区亥鼻 1–8–1
[3] 国立情報学研究所／
千葉大学総合メディア基盤センター
〒101–8430　千代田区一ツ橋 2–1–2／
〒263–8522　千葉市稲毛区弥生町 1–33
別刷請求先：千葉大学医学部附属病院企画情報部
高林克日己
E-mail: takaba@ho.chiba-u.ac.jp

[1] School of Science and Technology, Chiba University
1–33 Yayoi-cho, Inage-ku, Chiba 263–8522, Japan
[2] Division of Medical Informatics, Chiba University Hospital
1–8–1 Inohana, Chuo-ku, Chiba 260–8677, Japan
[3] National Institute of Informatics/IMIT, Chiba University,
Hitotsubashi 2–1–2, Chiyoda-ku, Tokyo 101–8430, Japan/
Yayoi-cho 1–33, Inage-ku, Chiba 263–8522, Japan

relationships among 13 diseases. ***Conclusion:*** These results suggest the possibility that text-mining technique is applicable to the automotive classification of medical documents according to the diagnoses.
(Keywords: discharge summary, vector space model, tf×idf, data mining, text mining, knowledge discovery)

1　はじめに

医療情報は病院情報システム (以下 HIS (Hospital Information System)) から電子カルテの時代へと進み，検査データや画像情報と並んで文書情報が大量に電子化されて蓄積される時代を迎えた．千葉大学医学部附属病院 (以下当院とする) においても，電子カルテシステムが 2003 年 6 月より段階的に稼働を始めている．こうした動きにともない，医療データの横断的検索も始まっている．医療現場で発生した大量の情報から医学や医療に有用な知識を発見する研究が行われるようになってきており，例えば，従来から蓄積されている臨床検査値などの数値データを対象にしてデータマイニングが行われるようになった[1〜3]．

一方で，医療以外の分野ではインターネットの急激な発展に伴って，文書を対象とした研究が非常に盛んになっている．和歌データベースを基にして類似和歌を自動抽出する研究[4] やインターネット上のチャットのように，時系列変化する動的テキストから主題識別を行う研究[5] など多岐にわたっている．医療の世界でも検査データなどの数値だけでなく，文書の電子化により今後は文書のテキストマイニングが横断的検索に利用されることが考えられる．

本研究では，この医療文書の 2 次利用可能性を示すために，情報検索の分野で広く用いられているベクトル空間モデル[6,7] の手法を用いて退院サマリーから診断に関連する言語情報の抽出を行った．退院サマリーは，退院された患者の入院中の状況を簡潔にまとめた情報[8] であり，疾患に関する情報を表現していると考えられる．ここでは，当院 HIS に保存されている退院サマリーを疾患毎に抽出し，ベクトル化を行って疾患毎の重要語を抽出した．次に，疾患毎の退院サマリーベクトルをもとに，ある症例の退院サマリーから疾患名を特定できるか否かの検討を行った．これより，医療文書の二次利用の可能性について検討した．また，特定した例に対してデンドログラムを作成し，分類の視覚化を試みた．

表 1　各臓器の代表 13 疾患とその症例数

疾患名	臓器	ICD-9	症例数
胃悪性新生物	消化器	151	524症例
肝，肝内胆管の悪性新生物	肝臓・胆	155	483症例
気管・気管支の悪性新生物	呼吸器	162	687症例
乳房の悪性新生物	乳房	174	363症例
前立腺悪性腫瘍	男性器	185	340症例
腎臓の悪性新生物	腎臓	189	158症例
リンパおよび組織球組織の悪性新生物	血液	202	153症例
糖尿病	内分泌	250	293症例
統合失調症	精神	295	104症例
白内障	眼	366	777症例
喘息	アレルギー	493	114症例
瘢痕拘縮	皮膚	709	133症例
変形性関節症	運動器	715	188症例

2　対象と方法

1)　対　　象

1999 年 3 月から 2003 年 5 月までの約 4 年間に当院 HIS に蓄積された延べ 36,335 症例の退院サマリーを対象とした．疾患の特徴を十分に反映した退院サマリーベクトルを作成するために，ICD-9 による疾患の分類コード[9] をもとに症例数が 100 以上ある 50 疾患を算出し，疾患の偏りを防ぐため，各臓器の代表的疾患である 13 疾患を選定した (表 1)．なお，退院サマリーを抽出する際には，個人を特定できる情報はすべて削除し，当院 HIS データベースシステム U-MUMPS 内でデータ処理を行った．

2)　方　　法

(1)　形態素解析

日本語で書かれた文書を対象とする場合にまず問題となるのは，文字列からの単語認定である．自然言語で書かれた文書の場合，最初に文字列から名詞や形容詞，助詞といった要素に分解する必要がある．この技術は，形態素解析と呼ばれる．現在，いくつかの形態素解析システムが開発されており，本研究では奈良先端科学技術大学院大学松本研究室で開発されたソフトウェア「茶筅 (ChaSen)」[10] を使用した．

また，本研究で対象とする文書は医療文書であり，一般の文書に比べて用語の専門性が非常に高い．したがって，茶筅のもつ辞書では医学用語を十分に抽出することができない．そこで，医学辞書として MEID 辞書[11] (語彙数約 22 万語) を選定し茶筅の辞書に追加した．

(2)　辞書の再構築

本研究では，臨床現場で実際に記載された退院サマリーを扱う．臨床現場では糖尿病を “DM” と呼ぶように，略語が多用されていることが知られている．そのために，実際に退院サマリーを形態素解析する場合，MEID 辞書だけでは退院サマリーから十分に医学用語を索引語として反映させることができない．

そこで，ここではまず茶筅により退院サマリーの形態素解析を行い，MEID 辞書には存在しない単語を “未知語” として抽出した．次にその単語の出現頻度を疾患毎に算出し，上位 50 位までの用語を MEID 辞書に追加することで，MEID 辞書の再構築を行った．この処理により，臨床現場に合った医学用語の抽出を可能とした．

退院サマリーからの索引語の抽出は，再構築した MEID 辞書に存在し，それ自体で意味をもつ内容語のみを扱うこととした．例えば，対象の退院サマリー全体から m 個の索引語を抽出した場合，以下のように示す．

$$w=[w_1\ w_2\cdots w_i\cdots w_m]^{\mathsf{T}} \tag{1}$$

(3)　ベクトル空間モデル

対象とする文書集合を $\boldsymbol{D}$ とし，疾患毎に退院サマリーを $\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_j, \cdots, \boldsymbol{d}_n$ とおく．また $\boldsymbol{D}$ から抽出した m 個の索引語は $w_1, w_2, \cdots, w_i, \cdots, w_m$ とする．さらに，ある疾患の退院サマリー $\boldsymbol{d}_j$ に現れる索引語 w_i に対する重みを α_{ij} とおく．このとき $\boldsymbol{d}_j$ を次のようなベクトル

$$\boldsymbol{d}_j=[\alpha_{1j}\ \alpha_{2j}\cdots\alpha_{ij}\cdots\alpha_{mj}]^{\mathsf{T}} \tag{2}$$

で表現し，これをある疾患 j の退院サマリーベクトルと呼ぶ．また退院サマリー集合全体は，次のような $m\times n$ の行列 $\boldsymbol{D}$

$$\boldsymbol{D}=[\boldsymbol{d}_1\boldsymbol{d}_2\cdots\boldsymbol{d}_n]=\begin{bmatrix}\alpha_{11} & \alpha_{12}\cdots\alpha_{1n}\\ \alpha_{21} & \alpha_{22}\cdots\alpha_{2n}\\ \vdots & \vdots\quad\ \vdots\\ \alpha_{m1} & \alpha_{m2}\cdots\alpha_{mn}\end{bmatrix} \tag{3}$$

によって表現される．この行列を索引語文書行列と呼ぶことにする．

抽出した索引語の中には，退院サマリーの内容と密接に関係したものから，関係の薄いものまで存在する．そこで，その索引語が退院サマリーの特徴を表す上でどれだけの重要度をもっているか示すために，索引語の重み付けの処理が必要となる．ここでは，索引語の重みの計算には以下の tf×idf 法を用いることとし，重み α_{ij}

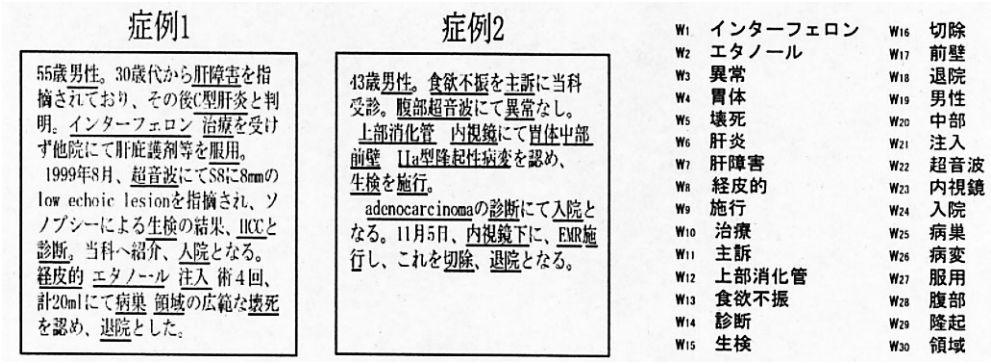

図 1　退院サマリーとその索引語の例
症例 1，症例 2，でアンダーラインが引かれている単語が索引語である．

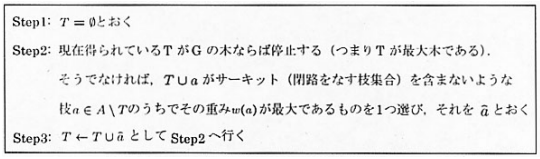

図 2　貪欲アルゴリズム

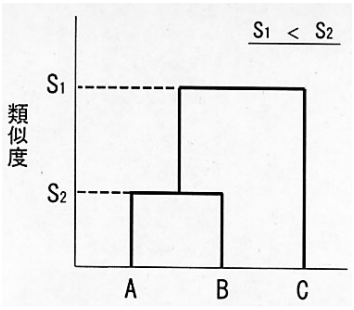

図 3　デンドログラムの例

$$\alpha_{ij}=\frac{l_{ij}g_i}{n_j} \tag{4}$$

を定義した．ここで，l_{ij}, g_i, n_j は以下によって定義する．

$$l_{ij}=\log(1+f_{ij}) \tag{5}$$

$$g_i=\log\left(\frac{n}{n_i}\right) \tag{6}$$

$$n_j=\sqrt{\sum_{i=1}^{m}(l_{ij}g_i)^2} \tag{7}$$

ここで，f_{ij} は索引語 w_i のある疾患の退院サマリー $\boldsymbol{d}_j$ における出現頻度である．また，n は対象とする退院サマリーの疾患数であり，n_i は退院サマリーに索引語 w_i を含む疾患数である．また，退院サマリーの文書の長さは，記載者や診療科，疾患の種類によって大きく異なる．このように文書の長さによる重み付けの影響をなくすために，n_j により正規化を行う．さらに，2つの退院サマリーベクトルの類似度をベクトルの内積 $\boldsymbol{d}_j\cdot\boldsymbol{d}_k^{\mathsf{T}}$ によって定義する．

次に，退院サマリーをベクトル化する例を示す．図 1 は，退院サマリーと抽出された索引語を下線で表している．症例 1, 症例 2 の退院サマリーの下線部から，医学用語である索引語が右側に 30 語抽出された．

これより，前述の手法によって抽出した索引語の局所的重み l_{ij} を算出し，退院サマリーのベクトル化を行うと以下の行列

$$\boldsymbol{D}=[\boldsymbol{d}_1\boldsymbol{d}_2]=\begin{bmatrix}0.080 & 0\\ 0.080 & 0\\ 0 & 0.080\\ 0 & 0.080\\ 0.080 & 0\\ \vdots & \vdots\\ 0.080 & 0\end{bmatrix} \tag{8}$$

が得られる．このとき，2つの退院サマリーベクトルの類似度は 0.045 であることがわかる．

(4)　貪欲アルゴリズムによるデンドログラムの作成

本研究ではグラフ理論やネットワーク理論における基本問題である最大木問題を用いてデンドログラムを作成しクラスタリングを試みた．

連結なグラフ $G=(V, A)$ と枝集合上の重み関数 $w: A \rightarrow R$ が与えられているとする．G の木 $T(\supset A)$ に対して以下で定義される $w(T)$ を木 T の重み (weight) とする．

$$w(T)=\sum_{a \in T} w(a) \qquad (9)$$

また，重みが最大である木を見出す問題を最大木問題 (maximum-tree problem) と呼ぶ．本研究では，各ノードにそれぞれの疾患をおき，各ノード間を繋ぐ木の重みを疾患間の類似度とした完全グラフから，貪欲アルゴリズム (図 2) を用いて最大全域木を求めた．これより，デンドログラムを作成し分類結果を視覚化した．ここでは，クラスタ間の類似度はクラスタ内にあるノードの組み合わせの中で類似度が最大であるものとした．

例として図 3 をあげる．図 3 は A, B, C をノードとする類似度の最大全域木より，類似度が大きい順にグループ化してその類似度を縦軸に表したものである．この図では，類似度が S_1 の時に A と B のクラスと C のクラスに分類されることがわかる．ただし，A, B, C の順序には意味はない．

表 2　13 疾患の索引語とその重み上位 10 位

	胃癌		肝癌		肺癌		乳癌		前立腺癌	
	索引語	α	索引語	α	索引語	α	索引語	α	索引語	α
1	前庭	0.104	エタノール	0.092	右中葉	0.082	乳管	0.195	サドルブロック	0.154
2	胃体	0.101	PHA	0.09	扁平上皮癌	0.078	C領域	0.172	前立腺	0.149
3	フード	0.099	コイル	0.088	気管分岐部	0.075	乳腺症	0.164	前立腺全摘除術	0.146
4	胃透視	0.092	前枝	0.086	肺瘻	0.071	胸筋	0.159	骨盤リンパ節	0.134
5	SE	0.089	Fc	0.083	入口	0.069	ノルバデックス	0.141	PK	0.121
6	胃全摘術	0.089	食道静脈瘤	0.08	肺機能	0.068	上肢挙上	0.138	除睾術	0.117
7	GFS	0.087	門脈	0.079	葉間	0.067	マンモグラフィー	0.125	ホンバン	0.109
8	亜全摘	0.083	アミノレバン	0.077	ブラシ	0.067	癌検診	0.121	タンデム	0.104
9	胃切除術	0.081	右枝	0.077	壁側胸膜	0.067	大胸筋	0.112	側精巣	0.104
10	器械	0.079	完全壊死	0.077	膜様部	0.067	乳房	0.111	直腸出血	0.104

	腎癌		悪性リンパ腫		糖尿病		統合失調症		白内障	
	索引語	α	索引語	α	索引語	α	索引語	α	索引語	α
1	腎腫瘍	0.167	PUVA	0.108	補食	0.11	幻聴	0.122	点眼液	0.211
2	腎盂	0.144	可溶性	0.102	腎症	0.091	拒絶	0.107	眼	0.205
3	尿管腫瘍	0.132	髄注	0.1	硝子体出血	0.089	疎通性	0.101	ミドリンP	0.173
4	右腎盂	0.126	幹細胞	0.097	神経伝導速度	0.087	隔離	0.098	水晶体乳化	0.168
5	腎細胞癌	0.12	耳下腺	0.095	ケトン	0.086	被害妄想	0.095	両眼	0.156
6	右尿管口	0.112	悪性リンパ腫	0.092	グルカゴン	0.082	妄想	0.094	乳化	0.148
7	腎摘除術	0.107	リンパ腫	0.087	ケトン体	0.081	行為	0.093	眼内レンズ	0.142
8	腎部分切除術	0.104	右扁桃	0.083	強化療法	0.079	空笑	0.089	吸引術	0.137
9	拡大率	0.103	躯幹	0.083	マイクロゾーム	0.077	妄想状態	0.087	右眼	0.136
10	上極	0.103	上咽頭	0.079	肥満度	0.072	ダール	0.085	左眼	0.132

	喘息		瘢痕拘縮		変形性関節症	
	索引語	α	索引語	α	索引語	α
1	インタール	0.192	瘢痕拘縮	0.189	裂隙	0.14
2	陥没	0.12	エキスパンダー	0.179	下腿周径	0.125
3	スギ	0.118	プロテーゼ	0.146	内反	0.12
4	ダニ	0.118	植皮術	0.143	荷重	0.118
5	胸骨上窩	0.112	シリコン	0.142	骨棘形成	0.113
6	呼気性喘鳴	0.11	挫創	0.134	左膝	0.112
7	持続吸入	0.108	ケロイド	0.118	CE角	0.11
8	大発作	0.105	左上眼瞼	0.118	外反	0.11
9	ブタクサ	0.103	全層植皮	0.116	脚長差	0.11
10	湿性	0.101	修正	0.112	動揺性	0.11

表 3　各疾患の正診率

疾患名	臓器	ICD-9	診断と一致	複数疾患の疑いあり	診断と異なる判定	判定不明
胃癌	消化器	151	24 / 30	4 / 30	0 / 30	2 / 30
肝癌	肝臓・胆嚢	155	20 / 30	2 / 30	0 / 30	8 / 30
肺癌	呼吸器	162	19 / 30	3 / 30	0 / 30	8 / 30
乳癌	乳房	174	19 / 30	1 / 30	1 / 30	9 / 30
前立腺癌	男性器	185	25 / 30	3 / 30	0 / 30	2 / 30
腎癌	腎臓	189	21 / 30	2 / 30	1 / 30	6 / 30
悪性リンパ腫	血液	202	17 / 30	6 / 30	1 / 30	6 / 30
糖尿病	内分泌	250	19 / 30	7 / 30	2 / 30	2 / 30
統合失調症	精神	295	28 / 30	1 / 30	0 / 30	1 / 30
白内障	眼	366	28 / 30	2 / 30	0 / 30	0 / 30
喘息	アレルギー	493	27 / 30	0 / 30	0 / 30	3 / 30
瘢痕拘縮	皮膚	709	13 / 30	0 / 30	4 / 30	13 / 30
変形性関節症	運動器	715	30 / 30	0 / 30	0 / 30	0 / 30
計			290 / 390 (74%)	31 / 390 (8%)	9 / 390 (2%)	60 / 390 (15%)

3　実際の実験と結果

1)　13 疾患の重要語の抽出

茶筅を用いて 13 疾患の退院サマリーの形態素解析を行った結果，7,918 語の医学に関連する索引語が抽出された．

ここでは，対象とする退院サマリーの文書集合を $\boldsymbol{D}$ とし，疾患毎の退院サマリーを $\boldsymbol{d}_{151}, \boldsymbol{d}_{155}, \cdots, \boldsymbol{d}_j, \cdots, \boldsymbol{d}_{715}$ とする．なお $\boldsymbol{d}$ の添え字 j は ICD-9 コードである．

また $\boldsymbol{D}$ を形態素解析することで抽出した 7,918 個の索引語を $w_1, w_2, \ldots, w_i, \ldots, w_{7918}$ とする．ここで，ある疾患 (ICD-9 コード j) の退院サマリーベクトル $\boldsymbol{d}_j$ における w_i に対する重みを α_{ij} とおく．なお，α_{ij} は ICD-9 コード j の退院サマリーにおける索引語 w_i の相対的な重要度を表現している．上述の tf×idf 法より索引語に対する重み α を求め 7918×13 行列

$$\boldsymbol{D}=[\boldsymbol{d}_{151}\cdots\boldsymbol{d}_{155}]=\begin{bmatrix}0.0114 & 0\cdots0\\ 0 & 0\cdots0\\ \vdots & \vdots\ddots\vdots\\ 0 & 0\cdots0\end{bmatrix} \quad (10)$$

を算出した．

次に 13 疾患毎に索引語の重みである α を算出し，疾患毎に α の大きい順に並び替えることで，医療現場で実際に使われている用語の中から疾患毎の重要語と思われる語を抽出した (表 2)．

2)　退院サマリーからの疾患の特定

算出した 13 疾患の退院サマリーベクトルをもとに，ある退院サマリーから疾患を特定できるか否かの実験を行った．

まず，算出に用いたのとは別の 13 疾患の退院サマリーを各 30 症例，計 390 症例無作為に抽出した．

次に，これらの症例の退院サマリーベクトルを算出し，既に算出した 13 疾患の退院サマリーベクトルとの内積を疾患毎に求めた．これにより算出された 13 疾患に対する類似度を，症例毎にレーダーチャートを用いて表現した．

ここでは，疾患を特定するための類似度の基準値を 0.1 以上とした．類似度が 0.1 以上であり，かつ第 1 診断病名と等しい場合は「退院サマリーの診断と一致」，第 1 診断名と異なる場合は「退院サマリーの診断と不一致」とした．また類似度が 0.1 以上の疾患が複数ある場合，「複数疾患の疑いあり」と判定し，類似度の最大値がどれも 0.1 未満の場合は，「判定不能」とした．

その結果，表 3 で示すように 390 症例中 290 症例 (74%) がサマリーの診断と一致した．複数疾患の疑いありと判定したものは 390 症例中 31 症例 (8%) であった．また，診断と不一致とされたのは 390 症例中 9 症例 (2%)，すべての疾患の類似度が 0.1 未満であったため，判定不能とした疾患は 390 症例中 60 症例 (15%) であった．

図 4 での糖尿病 20 症例の退院サマリーのレーダーチャートにおいて，類似度が 0.1 以上の疾患は

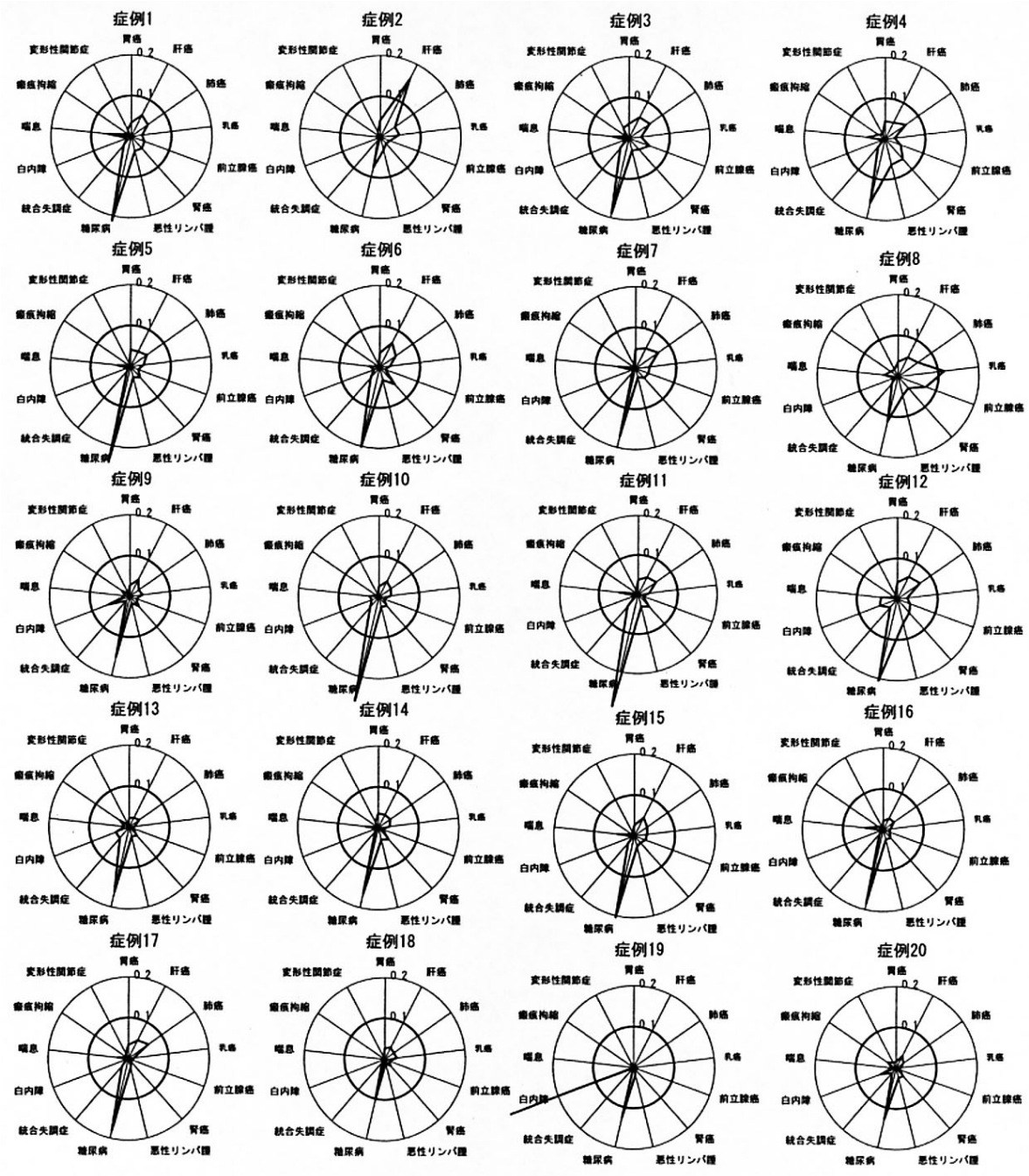

図 4　糖尿病の 20 症例に対するレーダーチャート
ほとんどの症例においてピークの方向が糖尿病へ向いている．ただし，症例 2，症例 8，症例 19 に関してはピークの方向が糖尿病以外の疾患へ向いており，糖尿病と他疾患の合併例である．

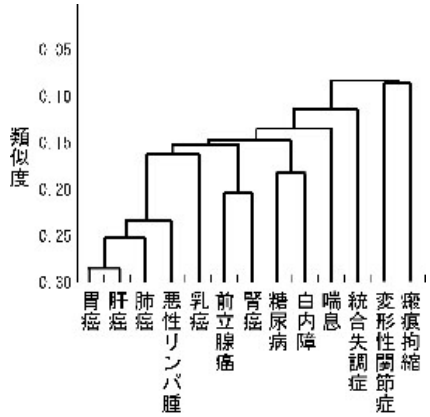

図 5　13 疾患の類似度の最大全域木によるデンドログラム

複数あり，「複数疾患の疑いあり」とする症例が複数存在していた．例えば，症例 2，症例 8 はそれぞれ肝癌，乳癌を合併している．また症例 19 は，糖尿病の合併症として白内障を診断されている患者であった．

3)　13 疾患のデンドログラム

算出した 13 疾患毎の類似度から最大全域木を求めデンドログラムによって視覚化を行った (図 5)．これによれば，この 13 疾患において最も類似しているものは胃癌と肝癌であり，これに肺癌と悪性リンパ腫のグループが類似している，というように表現されているが，左半分が悪性腫瘍として共通している．最も離れているものは，変形性関節症と瘢痕拘縮のグループであった．

4　考　　察

今回利用した手法によって，医療現場で実際に使われている文書から各疾患を特定しうる特徴的な索引語の抽出が可能であることが示された．また，疾患群の退院サマリーベクトルをもとにして退院サマリーから疾患名を特定できる可能性も示された．

退院サマリーは必ずしも完全なものばかりとはいえないが，多数例の解析により疾患を区別できる明瞭な差異が得られたことは興味深い．そして類似度をレーダーチャートで示すことで，退院サマリーの特徴を視覚的に捉えることが可能であった．単一疾患だけで 74%，複数疾患が含まれていた場合を加えると 82% に上り，かなり高い正診率と言うことができる．このレーダーチャートを用いて合併症の可能性を示唆したり，退院サマリーに隠れた新たな医学知識を発見したりする可能性も考えられる．抽出した索引語の中には “フード” や “ダール” のようにそれ自体で意味をもたない用語も抽出されている．これらは，それぞれ “透明フード” と “リスパダール” の一部である．医療現場に見合う辞書への改善が望まれる．

また，索引語にはこの “フード” や薬剤名など当院に固有の用語が存在しており，普遍的なものとは言えない．このことに対しては多施設における広範かつ大量な解析を行うことで，普遍的，絶対的な情報を得る可能性が考えられる．

一方，本方法は複数疾患の対比によって得られた相対的データであり，対比する疾患によっても値が変わってくる．今回は各臓器別の代表疾患の対比を示したが，この予備実験として行った，精神疾患や消化管疾患などの同一のサブスペシャリティにおける近縁疾患間の検討でも，同様に明瞭な識別が可能であることが確かめられている．また，デンドログラムで示されているように肺癌と喘息は類似しておらず，本方法は単に臓器の相違をみているだけではない．さらに相当数の疾患と多数施設の症例を検討することで，ある疾患の退院サマリーからの絶対的なベクトル情報が得られる可能性をもっている．

また，本研究では全体の文書の長さが長い事を考慮して，前述した比較的単純な tf×idf 法を用いたが，重み付けの算出方法はこの他にも様々な手法[7] があり，本方法が必ずしも最善とは言えないかもしれない．例えばもう少し複雑な，索引語の確率分布モデルに基づく重み付けの算出法との比較も考えていく必要があろう．

本方法の応用としては，退院サマリーの診断を自動判別する支援ツールとしての活用が考えられ

る．一方で原因が不明である症候群の解析をデータマイニングと組み合わせて行うことで，症候群の亜分類や，原因がわかっている疾患との類似性から新たな疾患の発見などの応用が期待できると思われる．

またMEDLINEを対象とした研究は，近年盛んに行われており，MEDLINEを対象としたテキストマイニングツール[12]も開発されつつある．一方で，本研究において課題となった，医学専門用語に関する研究も行われており，松本らはMEDLINEのテキストから医学専門用語の抽出や意味クラスへの分類を試みている[13]．

しかし実際の臨床データへのテキストマイニングの応用はまだ緒についたばかりである．われわれは今回退院サマリーを対象にしたが，竹村[14]らは放射線画像レポートのテキストマイニングから疾患の自動診断を考えている．画像レポートの場合は記述がある診断に特化した内容になるから，より知識ベースや潜在的なルールを発見できる可能性が高いという意味で興味深い．このように今後テキストマイニングはあらゆる医療文書に応用され，それぞれの意味をもった研究の展開が今後期待できる．

13疾患の類似度から求めた最大全域木によって作成したデンドログラムの結果も興味深いデータを示している．すなわちこの13疾患に関する限り，医療者から見て順当な位置関係を示している．ただしこれは今後多数例などでの再現性を確認する必要があろう．元来デンドログラムは体細胞や微生物の遺伝子を比較するために考えられた手法であるが，今回のように疾患の類似度の最大全域木にデンドログラムを用いることで，疾患群などの相関関係を視覚化できる点で意味があると思われた．

5　結　　論

茶筅を用いた形態素解析とベクトル空間モデルにより代表的13疾患における退院サマリーから疾患名を高率に自動診断し得た．このことから医療文書の横断的処理により，退院サマリー疾病分類支援や類似症例検索が期待できると考えられた．また，診断不明の疾患をクラスタリングすることによる新しい症候群の抽出支援や疾患の細分類など，新たな医学知識発見に向けた応用の可能性が示された．

なお，本論文の一部分は電子情報通信学会 (IEICE) パターン認識・メディア理解研究会 (PRMU) 2003-77 および医療情報学連合大会2003で口頭発表した．

文　　献

1) 高林克日己，倉沢和弘，縄田泰史，岩本逸夫，齋藤康．データマイニングによる抗リン脂質抗体症候群患者の血栓症の予知．リウマチ 2002; **42**: 343.
2) Takabayashi K, Yokoi H, Hirano S, Tsumoto S. Discovery challenge from temporary data of thrombosis. In APAMI&CJKMI-KOSMI conference Taegu KOSMI, 2003: 181–3.
3) Hirano S, Tsumoto S, Okuzaki T, Hata Y. A clustering method based on rough sets and its application to knowledge discovery in the medical database. *Medinfo* 2001; **10**: 206–10.
4) 竹田正幸，福田智子，南里一郎，山崎真由美，玉利公一．和歌データからの類似歌発見．統計数理 2001; **48**: 289–310.
5) Bingham E, Kaban A, Girolami M. Finding topics in dynamical text: application to chat line discussions. 10th International World Wide Web Conference (WWW10) 2001; Poster Proc 198–9.
6) Salton G, Wong A, Yang CS. A Vector Space Model for Automatic Indexing. *CACM* 1975; **18**: 613–20.
7) 北　研二．情報検索アルゴリズム．共立出版，2002.
8) 周　書義，高柳和江，木村哲彦．退院サマリーの認識に影響を与える要因に関する研究．日医大誌 1999; **66**: 52–60.
9) 厚生省大臣官房統計情報部．疾病 傷害および死因統計分類提要第2巻．財団法人 厚生統計協会，1978.
10) 松本裕治，北内　啓，山下達雄，他．形態素解析「茶筅」version2.2.7 使用説明書．奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座 (松本研究室)，2003.
11) 大島正光，開原成允，里村洋一，及川昭文，石塚隆男．医学用語大辞典．日外アソシエーツ株式会社，1990.

12) Srinivasan P, Rindflesch T. Exploring text mining from MEDLINE. *Proc AMIA Symp* 2002: 722–6.
13) 松本裕治，新保　仁．利用者からの要求を考慮したテキストデータからの知識抽出．情報洪水時代におけるアクティブマイニングの実現．研究成果報告書 2002: 243–53.
14) 竹村匡正，松井弘子，窪田英明，祐延良治，芦田信之放射線読影レポートからの自然言語知識抽出による自動分類の試み．医療情報学 2003; **23**: 95.