退院サマリからの知識発見を目的としたベクトル空間モデル適用実験

○ 小野大樹[1)], 高林克日己[2)], 鈴木隆弘[2)], 横井英人[2)], 井宮淳[3)], 里村洋一[2)]
千葉大学　大学院　自然科学研究科[1)], 千葉大学　医学部附属病院　医療情報部[2)],
国立情報学研究所[3)] / 千葉大学　総合メディア基盤センター[3)]

Hiroki Ono[1)] Katsuhiko Takabayashi[2)] Takahiro Suzuki[2)] HidetoYokoi[2)]
Atsushi Imiya[3)] Yoichi Satomura[2)]
School of Science and Technology, Chiba University[1)] Division of Medical Informatics, Chiba University Hospital[2)] National Institute of Informatics[3)] / IMIT, Chiba University[3)]

Keywords: ベクトル空間モデル tf*idf 法 退院サマリ

1．はじめに

近年の急速な IT 技術の発展に伴い，医療分野の電子化も電子カルテにより進みつつある．千葉大学医学部附属病院（以下当院とする）においても，M 言語の U-MUMPS を用いた病院情報システムにより臨床検査値などの数値データや退院サマリなどの医療文書データ，CT や MRI などの画像データがデータベースに蓄積されつつある．それに伴い，かつての検査値に代表される数値データを対象とした知識発見の研究が，最近では医療文書データを対象とした知識発見の研究に変わりつつある．

2．目的

本研究では，大量の医療文書からの知識発見を実現する為の医療文書処理の手法としてベクトル空間モデルを提案し，本手法を医療文書に適用した場合の実験結果を示す．

3．対象

当院では 1999 年 3 月から 2003 年 5 月までの約 4 年間に蓄積された延べ約 3 万 5 千症例の退院サマリがある．本研究では，症例数が 100 以上あり臓器別の代表疾患である 13 疾患の退院サマリを U-MUMPS によって抽出し実験対象とした．（表 1）

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・胆嚢	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
精神分裂病	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

表 1：抽出した臓器別の 13 疾患の退院サマリ概要

4．索引語の抽出

抽出した退院サマリは自由文で書かれているため，まず文字列から単語を認識する処理が必要となる．その処理は形態素解析と呼ばれ，本実験では形態素解析器にソフトウェア「茶筅」を使用した．さらに医学用語を抽出する為に必要不可欠な医学辞書として MEID 辞書（語彙数約 22 万語）を選定し，茶筅の辞書に追加した．

茶筅を用いて 13 疾患の退院サマリの形態素解析

を行った結果 7948 語の医学に関連する索引語が抽出された.

５．ベクトル空間モデル

本研究では，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の手法を用いた.

ここでは，対象とする文書集合を **D** とし，各々の疾患毎の退院サマリを $\mathbf{d}_{151}$, $\mathbf{d}_{155}$, ･･･, $\mathbf{d}_j$, ･･･$\mathbf{d}_{715}$ とした．なお, **d** の添え字 *j* は ICD-9 コードである.

また **D** を形態素解析することで抽出した 7918 個の索引語を w_1, w_2, ･･･w_i, ･･･, w_{7918} とした．ここで，ある疾患(ICD-9 コード j)の退院サマリ $\mathbf{d}_j$ における w_i に対する重みを α_{ij} とおく．このとき $\mathbf{d}_j$ を次にようなベクトルで表現し，これを疾患 j の退院サマリベクトルと呼ぶ.

$$\mathrm{d}_j = \left[\begin{array}{ccccccc} \alpha_{1j} & \alpha_{2j} & \cdots & \alpha_{ij} & \cdots & \alpha_{7918j} \end{array} \right]^{\top}$$

５．１　tf*idf 法

各索引語の重み α_{ij} を以下に定義する.

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

重み α_{ij} は, ICD-9 コード j の退院サマリにおける索引語 $\mathbf{w_i}$ の相対的な重要度を表現している.

局所的重み　　$l_{ij} = \log(1 + f_{ij})$

f_{ij} は索引語 w_i の ICD-9 コード j の退院サマリにおける出現頻度である.

大局的重み　　$g_i = \log(\frac{n}{n_i})$

なお, n は今回対象とする疾患の総数 13 であり n_i は索引語 w_i を含む疾患数である.

文書正規化係数　$n_j = \sqrt{\sum_{i=1}^{m} (l_{ij} g_i)^2}$

文書の長さによる重み付けの影響を減らすためのものである.

５．２　類似度

退院サマリベクトルの類似度を以下のように内積によって定義した.

$$S_{jk} = \boldsymbol{d}_j \boldsymbol{d}_k^{\top} = \sum_{i=1}^{m} \alpha_{ij} \alpha_{ik}$$

６．結果

６．１疾患毎の索引語と重みの抽出

本手法を用いて対象の 13 疾患の退院サマリをベクトル化し，疾患毎の重要語を抽出した.（表２は各疾患の重み上位 10 位までの索引語である）重みαが大きい索引語は，その疾患に特異度が高く出現頻度の多い索引語である.

６．２退院サマリからの疾患判定

次に，ある退院サマリから疾患を特定できるか否かの実験を行う為に，退院サマリをベクトル化する際には含まれていない，胃癌の症例の退院サマリを10症例無作為に抽出し，疾患毎の退院サマリベクトルと各症例の退院サマリベクトルとの類似度を症例毎に算出した.（表 3）

ここでは，類似度が 0.1 以上であり，かつ第一診断名と等しい場合は「疾患を特定」とした． 次に類似度が 0.1 以上の疾患が複数ある場合は，「複数疾患の疑いあり」とし，類似度の高い順に，第 1 病名，d 第 2 病名とした．類似度の最大値が 0.1 未満だが，突出した疾患がある場合は,「疾患の疑い」ありとした.

その結果，10 症例中 8 症例が胃癌と特定され，残りの 2 症例は胃癌の疑いありと判定された.（図１）

7．考察とまとめ

本手法によって，各疾患を特定しうる特徴的な索引語を抽出する事が出来た．また，13 疾患の退院サマリベクトルを基にして退院サマリから疾患名を特定できる可能性を示した.

しかし，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持たない用語も抽出

されている事が分かった．これらは，当院の診療科に特有の用語であるかもしれない．このことも含めて，抽出されたベクトルが当院に特異的である可能性は否定できない．

今後は，医療現場で多用される略語を反映した辞書作りや同義語に対する処理を加える必要があるが，一方で,一般性を得るためには，他院の退院サマリを対象に加えるべきであろう．

将来的には，病院情報システムには大量の医療文書が蓄積されていくが，本手法を用いてそれらの医療文書を横断的に処理する事により，電子カルテからの類似症例検索や新たな医学知識発見に向けた応用が可能と思われる．

	d_{151}		d_{155}		d_{162}		d_{174}		d_{185}		d_{189}	
	索引語	**α**	**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	腎腫瘍	0.167
2	胃体	0.101	ＰＨＡ	0.090	扁平上皮癌	0.078	Ｃ領域	0.172	前立腺	0.149	腎盂	0.144
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	尿管腫瘍	0.132
4	胃透視	0.092	前枝	0.086	肺癒	0.071	胸筋	0.159	骨盤リンパ節	0.134	右腎盂	0.126
5	ＳＥ	0.089	Ｆｃ	0.083	入口	0.069	ノルバデックス	0.141	ＰＫ	0.121	腎細胞癌	0.120
6	胃全摘術	0.089	食道静脈瘤	0.080	肺機能	0.068	上肢挙上	0.138	除睾術	0.117	右尿管口	0.112
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	腎摘除術	0.107
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	腎部分切除術	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	拡大率	0.103
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	上極	0.103

	d_{202}		d_{250}		d_{295}		d_{366}		d_{493}		d_{709}	
	索引語	**α**	**索引語**	**α**	**索引語**	**α**	**索引語**	**α**	**索引語**	**α**	**索引語**	**α**
1	ＰＵＶＡ	0.108	補食	0.110	幻聴	0.122	点眼液	0.211	インタール	0.192	瘢痕拘縮	0.189
2	可溶性	0.102	腎症	0.091	拒絶	0.107	眼	0.205	陥没	0.120	エキスパンダー	0.179
3	髄注	0.100	硝子体出血	0.089	疎通性	0.101	ミドリンＰ	0.173	スギ	0.118	プロテーゼ	0.146
4	幹細胞	0.097	神経伝導速度	0.087	隔離	0.098	水晶体乳化	0.168	ダニ	0.118	植皮術	0.143
5	耳下腺	0.095	ケトン	0.086	被害妄想	0.095	両眼	0.156	胸骨上窩	0.112	シリコン	0.142
6	悪性リンパ腫	0.092	グルカゴン	0.082	妄想	0.094	乳化	0.148	呼気性喘鳴	0.110	挫創	0.134
7	リンパ腫	0.087	ケトン体	0.081	行為	0.093	眼内レンズ	0.142	持続吸入	0.108	ケロイド	0.118
8	右扁桃	0.083	強化療法	0.079	空笑	0.089	吸引術	0.137	大発作	0.105	左上眼瞼	0.118
9	躯幹	0.083	マイクロゾーム	0.077	妄想状態	0.087	右眼	0.136	ブタクサ	0.103	全層植皮	0.116
10	上咽頭	0.079	肥満度	0.072	ダール	0.085	左眼	0.132	湿性	0.101	修正	0.112

表 2：疾患毎の索引語とその重み上位 10 位

類似度S	症例 1	症例 2	症例 3	症例 4	症例 5	症例 6	症例 7	症例 8	症例 9	症例 10
d_{151}(胃癌)	0.1587	0.1155	0.1470	0.1261	0.1596	0.1314	0.0747	0.1091	0.0729	0.1563
d_{155} (肝, 肝内胆管癌)	0.0714	0.0218	0.0504	0.0354	0.0442	0.0466	0.0304	0.0369	0.0254	0.0275
d_{162} (肺癌)	0.0217	0.0230	0.0322	0.0187	0.0321	0.0545	0.0246	0.0372	0.0096	0.0136
d_{174} (乳癌)	0.0446	0.0129	0.0241	0.0186	0.0117	0.0205	0.0084	0.0099	0.0044	0.0146
d_{185} (前立腺癌)	0.0063	0.0062	0.0100	0.0122	0.0115	0.0251	0.0178	0.0156	0.0035	0.0147
d_{189} (腎癌)	0.0179	0.0063	0.0137	0.0181	0.0204	0.0305	0.0149	0.0116	0.0018	0.0101
d_{202} (悪性リンパ腫)	0.0302	0.0256	0.0065	0.0107	0.0168	0.0439	0.0326	0.0199	0.0256	0.0165
d_{250} (糖尿病)	0.0121	0.0194	0.0423	0.0057	0.0275	0.0100	0.0133	0.0163	0.0004	0.0049
d_{295} (精神分裂病)	0.0040	0.0098	0.0025	0.0035	0.0500	0.0021	0.0081	0.0142	0.0004	0.0082
d_{366} (白内障)	0.0000	0.0008	0.0011	0.0000	0.0111	0.0039	0.0064	0.0027	0.0000	0.0018
d_{493} (喘息)	0.0000	0.0057	0.0057	0.0057	0.0057	0.0057	0.0057	0.0057	0.0057	0.0057
d_{709} (瘢痕拘縮)	0.0011	0.0024	0.0067	0.0145	0.0142	0.0015	0.0054	0.0040	0.0008	0.0016
d_{715} (変形性関節症)	0.0007	0.0014	0.0097	0.0001	0.0065	0.0007	0.0029	0.0011	0.0000	0.0014

表 3：疾患毎の退院サマリベクトルと胃癌 10 症例の退院サマリとの類似度 S

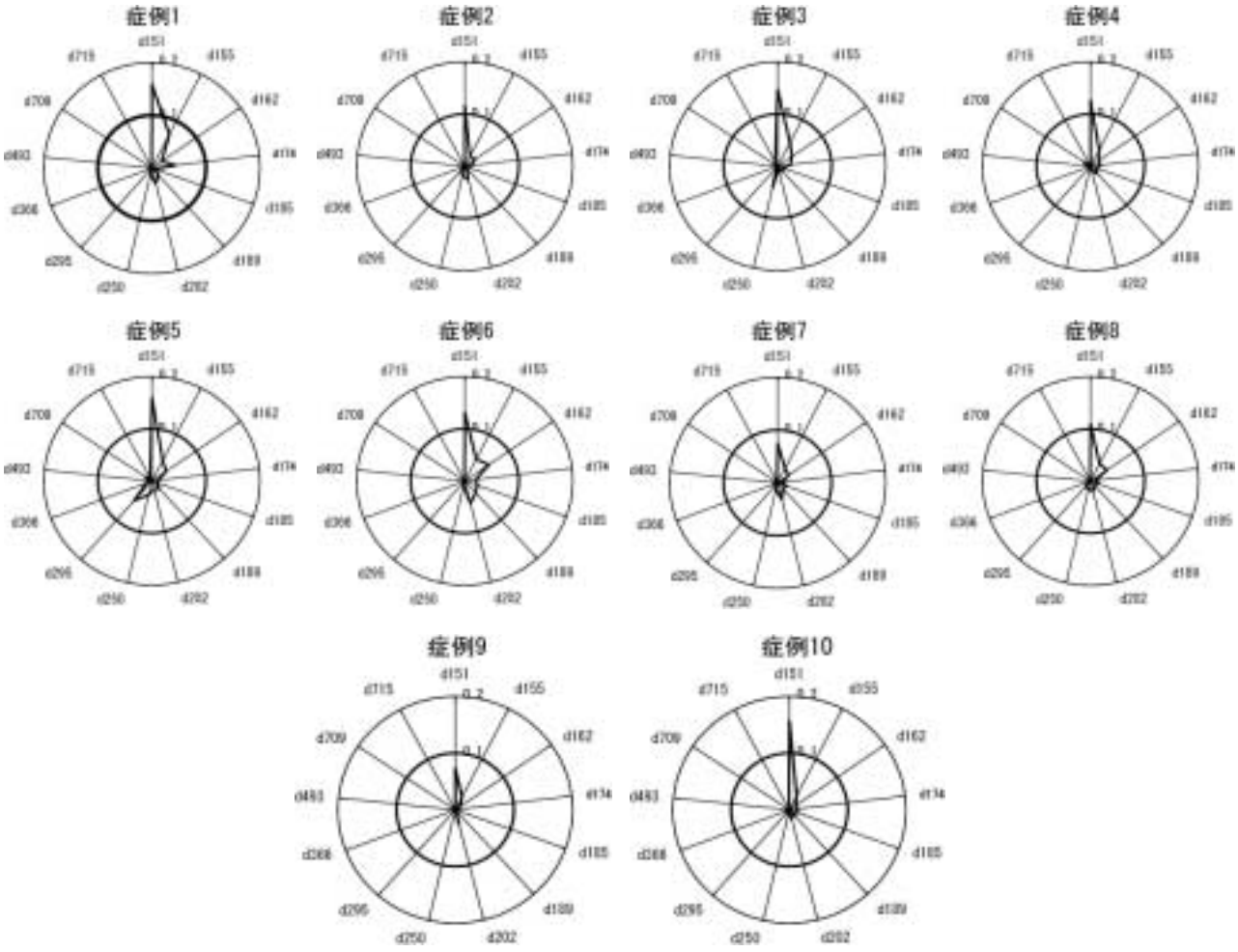

図１：症例毎のレーダーチャート