Discharge Summaries can be diagnosed from extracted index terms by text mining

Takahiro Suzuki [a], Hideto Yokoi [b], Shinsuke Fujita [c], Katsuhiko Takabayashi [a]

a Department of Medical Informatics and Management, Chiba University Hospital, Japan

b Division of Medical Informatics, Kagawa University Hospital, Kagawa, Japan

c Department of Welfare and Medical Intelligence, Chiba University Hospital, Chiba, Japan

Abstract

- Abstract
  - We analyzed electronic discharge summaries by text mining and demonstrated its potential to make an automatical diagnosis and classification. We extracted diagnosis related terminological information by morphological analysis from 4,784 discharge summaries of 14 representative diseases applying vector space model. We calculated the weights of each term for individual diseases and obtained each vector for 14 diseases by tf-idf and entropy methods. The index terms extracted from the discharge summary showed the features of the disease. In comparing tf-idf and entropy methods, the ranking of importance of the terms were different but the extracted terms were approximately the same. When applying these data to new cases, more than 96% of the correct diagnoses could be obtained by both methods. The classification by the clustering showed clear accordance with the code of ICD-9. These results suggest the possibility to classify diseases from medical documents automatically by using the text mining technique.

- Keywords:
  - Text mining, Discharge summary, Electronic medical record.

Introduction

- Recently in Japan, the electronic medical record has become popular, and electronic data has been accumulated in great quantities. At Chiba University Hospital, more than 30,000 electronic discharge summaries are stored. We showed the possibility of disease judgement and extraction of the index terms that characterize diseases by applying the vector space model to the discharge summary in Medinfo 2004 [1].
- In addition, expansion of the target diseases and plural methods of weighting were requested. Therefore we increased the number of target diseases, and added the entropy method to compare.
- Moreover, we performed cluster analysis by similarity computation, and tried automatic classification of the disease.

Materials

- The full text computerization of the discharge summary began at Chiba University Hospital in 1999. A total of 36,335 summaries were registered in four years by May, 2003.
- All the cases have been extracted from representative diseases which consisted of more than 100 cases individually. These diseases were selected from diverse fields of medicine such as gastric cancer in gastrointestinal disease (Table 1). Subsequently, we selected 4,784 discharge summaries from 14 diseases.

Table 1. 14 diseases and number of case

Disease	organ	ICD-9	Case
gastric cancer	GI tract	151	524
Hepatoma	liver	155	483
lung cancer	respiratory	162	687
mammary cancer	mamma	174	363
prostatic carcinoma	genital organ	185	340
renal carcinoma	renal	189	158
malignant lymphoma	hematology	202	153
diabetes Mellitus	endocrine	250	293
schizophrenia	psychiatric	295	104
Cataract	eye	366	777
Angina	circulatory	413	467
Asthma	allergy	493	114
scar contracture	skin	709	133
osteoarthritis	motor organ	715	188

Methods

1. **Obtained vector for discharge summary**

   We extracted diagnosis related terminological information from 4,784 discharge summaries by morphological analysis that resolves the character string to the element like a noun, an adjective, and a particle, etc. is necessary for the analysis of a Japanese natural sentence.
   We used the “ChaSen” that was developed as a morphological analysis system [2]
   The PHYXAM dictionary was used as a medical technical dictionary [3],
   We calculated the weights of each term for individual diseases and obtained each vector for 14 diseases by tf-idf and entropy methods.

2. **Disease judgment**

   We testify to the diseases from a discharge summary based on the calculated vector of the 14 diseases. First of all, we selected 10 cases of discharge summaries from 14 diseases, in total 140 cases as test cases. These summaries were not used to make the vectors according to disease.
   Next, similarity computation between the test case and each of the 14 diseases were calculated by inner products, and visualized by the radar chart. When the disease showed the highest similarity computation corresponding to the correct diagnosis, it was set to "matched".

3. **Cluster analysis**

   We substituted similarity between the diseases for the distance of the Euclidean square, and performed the cluster analysis by the furthest neighbor method.

Vector space model

- We calculated the weights of each term for every disease and obtained each vector for 14 diseases by the vector space model. The vector space model is a technique widely used in the field of information retrieval, which changes the characteristic of the document into the multidimensional vector. [4] In this research, we compared two weighting method (tf-idf and entropy methods).

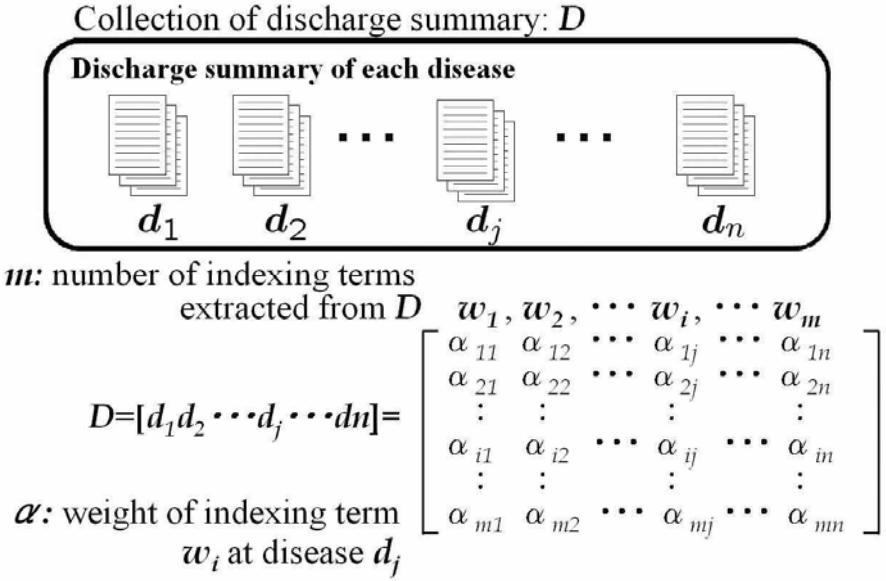

Figure 1- Vector space model

The total targeted discharge summary set is assumed to be D, and discharge summary sets of each disease are assumed to be $d_1, d_2, \cdots, d_j, \cdots, d_n$. Next, the m pieces of index term extracted from D are assumed to be $w_1, w_2, \cdots, w_i, \cdots, w_m$. Weight of index term wi in the discharge summary d_j is assumed to be α_{ij}. As a result, the discharge summary d_j is expressible by the vector that consisted of the m pieces of element α, and defined this as the discharge summary vector of disease j.

Weight of the index term

- tf-idf method
  - In tf-idf method, weight α_{ij} is shown by $\alpha_{ij}=(l_{ij} * g_i)/n_j$.
  - l_{ij} is the local weight, and a big value is given to the index term that appears frequently in the document. g_i is global weight, and a big value is given to the term that appears only in a specific document. n_j is a document normalization coefficient to which removes the influence by the length of documents.

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

tf $$l_{ij} = log(1 + f_{ij})$$

idf $$g_i = log(\frac{n}{n_i})$$

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

f_{ij} · · · Frequency of term w_i in disease d_j
n · · · Total number of the diseases in discharge summary
n_i · · · Number of the diseases which include the term w_i

- Entropy method
  - Entropy decreases when each event occurs in equal probability. The value of H_i/log is close to 1 when the index term appeared in each document at an equal probability, while it closes to 0 when it appeared only in limited documents. Therefore, the following expressions can be used as global weight of an index term w_i.

$$g_i = 1 - H_i = 1 + \frac{1}{\log n} \sum_{j=1}^{n} \frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i}$$

Results: Extraction of the index term

- We extracted 8,375 index terms from the discharge summary set D which is composed of 4,783 cases. The summaries of 14 diseases were vectorized by using the weight of the index term. Table 2 shows the result of permuting the extracted index term of each disease in order of weight. It seems that the index term characterizing each disease was extracted from among the terms actually used in the hospital. tf-idf method and entropy method were compared in each disease. No significant difference was found about medical validity observed in both methods. The correlation coefficient of the weight of all the index terms between both methods indicated a very high value (0.95).

Table 2. Top 10 list of index term order by severity (abstract)

	gastric cancer		hepatoma		Angina	
	tf-idf	**entropy**	**tf-idf**	**entropy**	**tf-idf**	**entropy**
1	IIc	EMR	PEI	HCC	労作時胸痛 (exertional chest pain)	CAG
2	IIa	IIc	CTA	PEI	aVf	PTCA
3	laser	GS	LpTAI	TAE	アセチルコリン(acetyl-choline)	LAD
4	por	GFS	TAI	TAI	Q波 (Q wave)	胸痛(chest pain)
5	胃全摘術(total gastrectomy)	胃癌 (gastric cancer)	Angio-CT	DyCT	RCA	ニトロ(nitro)
6	MDL	胃体(body of stomach)	RHA	RFA	hypokinesis	RCA
7	EMRC	group	lipiodol	CTAP	不安定狭心症(unstable angitis)	LVG
8	CY	幽門 (pylorus)	staining	CTA	絞扼感 constriction	aVf
9	LapMR	小弯(lesser curvature)	PVTT	LpTAI	NTG	舌下 (hypoglottis)
10	Billroth	IIa	RFA	AFP	Extremities	胸部圧迫感 (chest constriction)

Results: Disease judgment

- As shown in Table 3, 137 cases were correctly diagnosed out of 140 cases (98%) by entropy methods and 135 cases (96%) by tf-idf method. There were no significant differences between both methods. All the cases misjudged here were those with ancillary diagnosis and computer selected a complication as a main diagnosis.

Table 3. Judgment result of each disease

disease	ICD-9	tf-idf	entropy
gastric cancer	151	9	9
Hepatoma	155	10	10
lung cancer	162	10	10
Mammary cancer	174	9	10
prostatic carcinoma	185	10	10
renal carcinoma	189	9	9
Malignant lymphoma	202	10	10
Diabetes Mellitus	250	9	9
Schizophrenia	295	10	10
Cataract	366	10	10
Angina	413	10	10
Asthma	493	10	10
scar contracture	709	9	10
Osteoarthritis	715	10	10
Accuracy	709	96%	98%

Results: Cluster analysis of 14 diseases

- We performed cluster analysis from similarity computations between 14 calculated diseases by the nearest-neighborhood method, and visualized it by the dendrogram. According to these results, three diseases became complicated such as diabetes, angina, and cataracts and malignant tumors such as stomach carcinomas and liver carcinomas were established in the small groups.

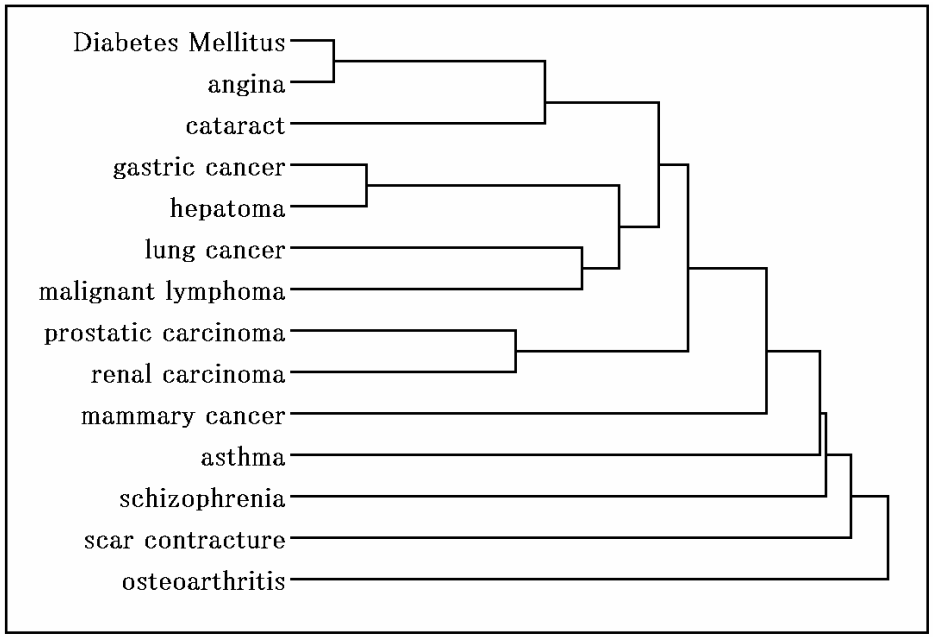

Figure 3- Cluster analysis of 14 diseases

The group that was the furthest from others was osteoarthritis. The results of the classification by clustering were almost the same order of the code system of ICD-9, and showed good accordance from a medical viewpoint.

Discussion & Conclusion

- **Discussion**

  We showed the extraction of the linguistic information and by using it we made a diagnosis of the discharge summary by the vector space model, in Medinfo 2004. [1] However, the some questions arose. Firstly it happens to work only in the combination of specific diseases and did not function when other kinds of diseases were added, such as circulatory diseases. Secondly, since there are various techniques for weighting the index terms, therefore it would be more effective than the tf-idf method.

  We identified diseases even though we expanded the number of diseases. Furthermore, we experimentally analyzed the discharge summaries of other hospitals, and we could identify the diseases as well as the summaries at our hospital. (Data not shown) We believe that this technique is universal when target diseases and facilities expand in the future, and the universal vector data is obtained. In comparing the methods of calculating weights, extracted terms and phrases were almost the same between the tf-idf method and the entropy method. In identifying diseases, there were no significant differences. Furthermore, clear superiority was not observed about medical validity between both methods.

  The research on the medical technical term using the vector space model is performed. We analyzed the documents in Japanese, however, similar results can be expected in English, because we only used the index terms and their weighting for analysis from the characteristics of language. Takemura et al. are experimenting with the automatic diagnosis of the diseases from text mining of the radiograph report [5] and Iwahashi et al. are also researching automatic classification of an incident report [6]. Text mining seems to be applied to all medical documents in the future.

- **Conclusion**

  We extracted the index terms that characterize diseases, and by using them identified more than 96% out of 140 discharge summaries. There were no significant differences between the tf-idf method and the entropy method. The classification by the clustering showed good accordance with the code of ICD-9.

  By collecting discharge summaries in many facilities, we can obtain universal vectors, characteristics for every disease and analysis through for electronic medical records by textmining, therefore we can expect to discover new medical knowledge.

References

1. Ono H, Takabayashi K, Suzuki T, Yokoi H, Imiya A, Satomura Y. Extraction of diagnosis related terminological information from discharge summary. Medinfo. 2004: (CD): 1786.
2. Matsumoto Y, Kitauchi A, Yamashita T, Hirano Y, Matsuda H, Takaoka K, Asahara M. Japanese Morphological Analysis System ChaSen version 2.2.1. Nara institute of science and technology, 2000
3. Fujita S. The interaction of the reason for encounter (ICPC-2) and standaerized physical findings (PHYXAM). Proceedings of 2nd Annual Conference of Japan Association of Medical Informatics 2004:1:908-909
4. Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing. CACM 1975:18:613-620.
5. Takemura T, Matsui H, Kubota H, Sukenobu Yoshiharu, Ashida S. Trial of automating classification of radiation reports by knowledge extraction from natural language. Japan Journal of Medical Informatics 2003: 23：95
6. Iwahashi Y, Ohe K. Trial of automating classification of incident reports. Proceedings of 2nd Annual Conference of Japan Association of Medical Informatics 2004:1:804-805

- **Address for correspondence**

Takahiro Suzuki, MD, PhD, FJSIM

Department of Medical Informatics and Management, Chiba University Hospital. 1-8-1 Inohana, Chuo-ku, Chiba, 260 Chiba, Japan. e-mail: suzuki@ho.chiba-u.ac.jp