MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

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

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 [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.

Methods

Discharge Ssmmary

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. Free text which described the history of the present illness was used for analysis in this research. All the cases have been extracted from representative diseases that consisted of more than 100 cases individually. These diseases were selected from diverse fields (Table 1). Subsequently, we selected 4,784 discharge summaries from 14 diseases.

Morphological analysis

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. In this research, we used the "ChaSen" that was developed as a morphological analysis system [2]. The PHYXAM dictionary was used as a medical technical dictionary [3].

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 [4]. In this research, we compared two weighting methods (tf-idf and entropy methods).

Tf-idf method

In tf-idf method, weight *ij* is shown by *ij=(lij×gi)/nj*. *lij* is the local weight, and a big value is given to the index term that appears frequently in the document. *gi* is global weight, and a big value is given to the term that appears only in a specific document. *nj* 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 *Hi*/log is close to 1 when the index

term appeared in each document at an equal probability, therefore, the following expressions can be used as global weight of an index term.

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

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 that were not used to make the vectors, in total 140 cases as test cases. Next, similarity computation between the test case and each of the14 diseases were calculated by inner products. When the disease showed the highest similarity computation corresponding to the correct diagnosis, it was set to "matched".

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.

Results

Extraction of the index term

8,375 index terms have been extracted from the discharge summary set which is composed of 4,783 cases. 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. 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).

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.

Cluster analysis of 14 diseases

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

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, the difference was seen in the order, however extracted terms and phrases were almost the same between the tf-idf method and the entropy method. Furthermore, clear superiority was not observed about medical validity between both methods. In identifying diseases, there were no significant differences.

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]. Text mining seems to be applied to all medical documents in the future.

[1]

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

Table 2 - Top 5 list of index term order by severity (abstract)

	gastric cancer		hepatoma		Angina	
	tf-idf	entropy	Tf-idf	entropy	tf-idf	entropy
1					(exertional chest pain)	
2						
3					(acetyl-choline)	
4					(Q wave)	(chest pain)
5	(total gas-trectomy)	(gastric cancer)	Angio -CT			(nitro)

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		96%	98%

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.

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 encouter (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 Y, Ashida S. Trial of automating classification of radiation reports by knowledge extraction from natural language. Japan Journal of Medical Informatics 2003: 23 95