Web-based Case Reports Retrieval System by TF*IDF Method

Shunsuke Doi[a], Tatsunori Tanaka[a], Takahiro Suzuki[b], Toshiyo Tamura[a], Katsuhiko Takabayashi[b]

[a] *Graduate School of Engineering, Chiba University, Japan*
[b] *Department of Medical Informatics and Management, Chiba University Hospital, Japan*

Abstract and Objective

*To learn from similar cases is one of the most important and beneficial processes for clinicians when they encounter a difficult case to make a diagnosis or decide treatment. Most of the databases of case records, however, were not available for comprehensive search. We built a database of more than 15,000 case reports from the Japanese Society of Internal Medicine as well as extraction of case reports from MEDLINE and conducted morphological analysis. Now we can provide to Japanese physicians the option to search for similar cases through the internet by using the TF*IDF method from JSIM homepage.*

Keyword:

Case reports, Clinical laboratory information system,

Methods

Collection of case reports

Clinical medical case reports are very useful material to learn from past similar cases. We collected a huge number of case reports from the database of the Japanese Society of Internal Medicine (JSIM) and MEDLINE. JSIM has stored abstracts of 15,000 cases which were presented in regular meetings for over the past four years. In addition we extracted 100,000 abstracts that included the word "case report" in its title from MEDLINE in the last ten years.

Morphological analysis and Reconstruction of dictionary

Morphological analysis is the most important process in order to compare all the terms in case reports and find similar ones by using the tf*idf method. We extracted index terms by morphological analysis and searched medical terms with our dictionary. Japanese is one of the most difficult languages to perform morphological analysis because Japanese sentences have no spaces between words. We reconstructed a dictionary with some glossaries, such as PHYXAM that has been used as a medical technical dictionary and glossaries of drugs, injections, and diseases used at Chiba University Hospital. In this research, we used MECAB developed as a Japanese morphological analysis tool by the Computational Linguistics Laboratory, Kyoto University and Tree Tagger developed as an English morphological analysis tool by the Institute for Computational Linguistics of the University of Stuttgart.

tf*idf method and Vector space model

Before calculation, we translated case reports from Japanese into English by our Japanese-English and synonym dictionary. After that, we calculated a weight of each word by the tf*idf method. If it sets a document "i" and a word "j", tf*idf is expressed as the following equation.

$$Wij = \frac{tf(ij) * idf(j)}{N(i)}$$

tf(ij): term frequency idf(j): inverse term frequency
N(i): document normalization coefficients
All case reports are expressed as the vectors by tf*idf method. Next, we calculated the degree of similarity defined as inner products between vectors. ($0 \leq s.d \leq 1$)

Retrieval system in internet

We built internet services so that users can retrieve easily. A digitalized user's text of a case record is first inserted into the dialog box morphologically analyzed and compared with all stored cases one by one by calculating inner products. Then they are sorted in order of the degree of similarity. Users can obtain more information in detail and access to the authors.

Results

By using this technique, we demonstrated that we could select similar cases from 115,000 in a few seconds by way of internet services. The internet services were just started and we performed 80 real cases. Most of the highest degree of similarity was usually around 0.15-0.3, even though cases with higher degrees of similarity were not always related to what users wanted. The satisfactory point of this retrieval was evaluated at average 2.2 on a three point scale (1: poor 3: good) in them.

Conclusion

We constructed a similar disease searching system by morphological analysis and the tf*idf method. This system is now functioning and open to Japanese physicians to retrieve similar cases.