A Study on the Extraction of Text Similarity from Genome Databases using Vector-space Model

Takao Kataoka

School of Knowledge Science,
Japan Advanced Institute of Science and Technology

March 2000

Keywords: genome database, vector-space model, keyword retrieval, natural language infomation.

abstract

From the discovery of double-helical structure of DNA by Watson and Crick, life science has been rapidly progressed. As the result of it, large amount of data on genes and proteins have been gathered on computers, and they are currently named "genome databases" generically. So, modern biologists and medical scientists are usually accessing fundamental retrieval (keyword search) and analysis (homology search) services on genome databases.

On the other hand, from web browser has been developed, WWW has been explosively grown. As the growth of data on WWW, technology of search engine based on full-text search algorithm became actively studied. By using search engines, we can convert tens of million web pages in the world into useful knowledge. In general, a search engine splits the words in a text, count the frequency of them, and generate index files for fast retrieval of the texts which meet a user's preference, i.e. query keywords. Furthermore, a new type of search engines emerged fwhich employ a more sophisticated and extended approach called "vector-space model". In the approach, a text is converted and represented by the vector of word frequencies, and a user can express his/her preferences by inputting a text as a whole (e.g. a system named ConceptBase developed by Justsystem corporation).

However, there are some difficulties in directly applying such search engines to similarity search and clustering of scientific databases (e.g. genome databases). In this study, An experimental system for clustering data entries in genome databases was developed[1]. It has the following features.

1. It utilizes a freeware called Namazu, which is the most famous search engine in Japan.

2. From the indices generated by Namazu, the system generates keyword vectors for entries.

3. Differences in fields in entries are considered before processing various genome databases.

4. It equips three kind of dictionaries for filtering useful, useless, and harmful keywords for similarity calculation based on vector-space model. That is, dictionaries for technical terms, usual English words, and field names.

Publication

[1] Takao Kataoka and Kenji Satou: A Full-Text Search System Covering the Whole GenomeNet , Genome Informatics 1999, UNIVERSAL ACADEMY PRESS,INC. TOKYO, JAPAN.