Clustering of Documents by Multiple Correlation

Takehiko HIRATA*

(Received December 4, 1973)

Abstract

In information storage and retrieval systems, the clustering of input information is necessary. Various definitions and methods of clustering have been put forward. However, here it is defined as the grouping of almost the same kinds in the case of documents, and the method of clustering by multiple correlation coefficients is proposed.

Introduction

Various definitions and methods for the clustering of documents have been given[1),2)], but here it is defined that the clustering of documents is the grouping of almost the same family of documents, and the method of clustering by multiple correlation coefficients is proposed.

Definition of multiple correlation coefficient

Firstly, in the correlation coefficients between two documents **A** and **B** that have been defined variously[1)], the directional nonsymmetric correlation coefficient is used here as follows:

$$r_{AB}=\frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i} \tag{1}$$

$$r_{BA}=\frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} b_i} \tag{2}$$

where, $a_i=1$ or 0 ($i=1, 2, \ldots, n$) represents the existence of an index in a document **A**. b_i is similar to a_i.

Next, when K documents **1**, **2**, ..., **K** are given, the multiple correlation coefficient between document **1** and the others **2**~**K** is written as follows:

$$r_{1\cdot 23\ldots K}=\sqrt{1-\frac{\Delta}{\Delta_{11}}} \tag{3}$$

* Department of Electrical Engineering

where, Δ is the determinant which elements are the correlation coefficient r_{IJ}, and Δ_{IJ} is the cofactor of the IJ element of Δ.

Clustering

Now, given the set of M documents $\Omega = \{\mathbf{1}, \mathbf{2}, \ldots, \mathbf{M}\}$, let's consider the clustering of this set. First, forming the power set of Ω, then 2^M elements are generated. Next, an element $\omega_K = \{\mathbf{1}, \mathbf{2}, \ldots, \mathbf{K}\}$ which is a subset of Ω is taken out, and then for the individual elements of ω_K the multiple correlation coefficients in the subset, are computed. Let the results be $r_{1\cdot 23\ldots K}, \ldots, r_{K\cdot 12\ldots(K-1)}$. If in K values there exists even one less than a certain threshold value t, this subset ω_K is taken off for the reason that it cannot constitute the cluster for the threshold t. Further, when all K values are greater than threshold t, and if in the elements of a power set of ω_K there exists even one that cannot constitute the cluster, ω_K is taken off for the reason that it cannot constitute the cluster.
That is to say, the cluster is the greatest subset with multiple correlation coefficients between an element and all combinations of other elements, over threshold t.

Conclusion

A brief experiment according to the above definition was made with 8 books on electrical engineering. The result was satisfactory except for the problem of how much threshold value to set. The problem of the threshold will be investigated by the author.

Acknowledgment

The author acknowledges the assistance dueing the experiment, of 3 research students of the Department of Technology at the University of Yamaguchi, Mr. Akira Kanekiyo, Mr. Yasuo Koyama and Mr. Nobuo Miki.

References

1) G.Salton: "Automatic Information Organization and Retrieval," (McGraw-Hill) p. 133–148, 236–243 (1968)
2) K. Samuelson: "Mechanized Information Storage, Retrieval and Dissemination", (North-Holland) p. 73–107, 225–234 (1968)