九州大学総合理工学研究科報告
第11巻 第3号 327 - 337頁 平成元年12月

Engineering Sciences Reports, Kyushu University
(KYUSHU DAIGAKU SOGORIKOGAKU KENKYUKA HOKOKU)
Vol. 11, No. 3, pp. 327 - 337, December 1989

A new method of modeling and clustering for citation relationship

Tatsuki SAITO* and Chooichiro ASANO**

(Received August 22, 1989)

Recently some researcher-databases have been constructed to obtain detailed information for their citation relationship. In the present paper, a fundamental methodology of modeling is proposed for a scientific article-space with the relational structure among scientific articles, where the techniques of clustering are studied on the basis of a-valued and directed graphs and the asymmetrical similarity matrix.

1 Introduction

There exists impotant and characteristic information involved in various relations among scientific articles. However, most of bibliographic databases in market do not give such accurate information[1]. Consequently, the direct acquisitions are not able to help the users on essential information involved in the originsl set of articles. From this viewpoint, a researcher-database[2),3)] is newly-proposed to obtain detailed information regarding citation relations among original articles. Basing on the application of such a database, it may be expected to grasp some dynamic trends of further development of scientific researches in the field.[2),3),13)~18)]

The purpose of the present paper is to propose a new methodology to make clear the structure of relationship among articles to promote the above study. Also the clustering methods are given with graphic consideration. The ordinary graphs applied are dual-directional and are easily processed[19),20)]. However, information involved in scientific articles is one-directional, i. e. one-way between a citing article and a cited article. Therefore, in order to represent such one-dimensional graphs with strength values and to construct more precise model of article relationship, a method is studied with the validity of representing relationship based on asymmetric similarity matrices among articles. Combinatorial cluster analyses are ordinarily suitable[21)~27)], when the relation matrix is symmetric. In this case, however, it is required to transform the matrix to a symmetric matrix. Thus a new method of clustering is obtained for asymmetric similarity matrix, aborting such a trouble. the clustering method with binary reiationship is based on the strength among articles as a criterion, and actually this method has shown the better performance than the combinatorial methods.

2. Modeling by citation relation graph

2.1 Valued graph

Graph consists of a non-empty finite set P with P points, and it consists of a specified

*Research Assistant, Faculty of Engineering, Hokkaido University
**Interdisciplinary Graduate School of Engineering Science, Dept. of Information Sciences

set L that has nonordered q pairs to belong to P. The pair $l=(i, j)$ of points i and j belongs to the set L, and it is called a line of a graph. A graph that has p points and q lines is called an (p, q) graph, or a graph $G=(P; L)$. Graph is applicable to model structural object. Namely it is a relational graph that a point is corresponding to an objective article and a line is corresponding to the relation. The nondirected graph is thought to be the special directed graph that always accompanies a directed line of reverse direction. In case of directed graph, the finite point set P that is not empty and specified set L that has ordered pairs of two different points are dealt simultaneously. Scientific articles by the same author are connected by lines. The relation of author is not directed, but the relation of citation is directed. Therefore, the citation relation is represented by a directed graph. While directed graph can express the presence or absence of a relation among articles, the strength of relation cannot be expressed only by it. Consequently, valued graph is introduced to enable an expression of the strength of relation. Valued graph (P; L; Y) is the graph that a line of the set L of lines is accompanied with the value r that is mapped onto a real number set Y. Expressing the value that is accompanied with a line (i, j) as $r(i, j)$, it is called a value of the line. The production of the valued graph and its relational similarity matrix are describad in 2. 2.

2.2 Modeling by similarity matrix

Fundamental Procedure for Modeling

The fundamental procedure for modeling is as follows,

1. Consider a binary citation relation between articles i and j.
2. Represent the binary relation as a relational graph expressed by points i and j.
3. Generate a directed valued-relation graph that corresponds to citation relations among articles.
4. Represent those graphs as similarity matrices.

The generation of direct citation graph and the formation of its similarity matrix are described as below,

Direct Citation Relation Matrix

Define direct citation relation matrix $A=[a_{ij}]$. This is so-called an adjacency matrix, where $a_{ij}=1$; if article i cites article j.

0; otherwise.

Citation Relation Matrix

Considering a total citation relation that involves indirect citation, define citation relation matrix $S=[S_{ij}]$, where

$$S_{ij}=\sum_{k=1}^{n} w_{k\cdot k}a_{ij}\,, \tag{1}$$

k means the length of a directed walk, and ${}_{k}a_{ij}$ means the number of a directed walk of which the length is k between article i and article j. The upper bound n is less than $Max\ (k)$ (maximum length of directed walk), w_k is a weight and takes a value such as $1/k$ or $1/k^2$. The ${}_{k}a_{ij}$ can be obtained by kth power of matrix A as resetting diagonal elements 0. An efficient algorithm that calculates only the nonzero elements of a matrix A has been developed.

3. Cluster analysis for relational graph model

3.1 Characteristics of relational graph model

In this section, we introduce new methods of clustering which analyze the relational graph represented by the similarity matrix. Traditional techniques of the combinatorial cluster analyses are discussed in 3. 2 and the present method is described in 3. 3.

As mentioned in the previous section, the analyzed matrix is given by

$$R=[r_{ij}],\ r_{ij}=(S_{ij})^P, \tag{2}$$

where P is an enhancing factor (ordinarily $P=1$).

The characteristics of this matrix are summarized as follows,

1. Each element of the matrix has a positive real quantity.

Element s_{ij} is the similarity as a correlation-like measure, then element r_{ij} has a positive real quantity.

2. The matrix is asymmetric.

The similarity matrix of the citation relation is asymmetric, because there is a time sequence in the incidental relation between a citing article and a cited article. In consequence, the matrix R becomes asymmetric. When the methods of the combinatorial cluster analyses discussed in 3. 2 are applied, it is necessary to change to the symmetric matrix. However, it is not desirable due to occurring the distortion of the original matrix. We propose a new clustering method for the asymmetric matrix in 3. 3.

3.2 Application of combinatorial clustering analysis

The combinatorial cluster analysis is also called the method of hierarchical cluster analysis, because it constructs a tree. All computer programs of the following combinatorial methods have been implemented and applied for the analysis of the relational graph model. Since the original similarity matrix of the relational graph model is asymmetric, it is necessary to change to a symmetric matrix, because the combinatorial methods are only applied to symmetric matrices.

Seven hierarchical clustering methods are examined. The nearest neighbour method is based on single linkages, because clusters are joined at each stage in view of the single shortest or strongest link among them. The furthest neighbour method is called the complete-linkage method, since all articles in a cluster are linked to each other in view of some minimum similarity. The median method adopts the middle value of the nearest neighbour value and furthest neighbour value. The method of average linkage within the new group is not influenced by extreme values for clustering and cannot make any statement about the minimum or maximum similarity within a cluster. Average linkage between amerged group, called the group average method, evaluates the potential merger of clusters i and j in terms of the average similarity between the two clusters. The centroid method uses both the mean values of similaries and number of articles for the merger. The minimum variance method, called Ward method, is generally reasonable, because those merger give the minimum increase at each stage for the total within group error-sum of squares.

When these combinatorial methods are applied, a symmetric similarity matrix R is

changed to symmetric matrix by (3).

$$r_{ij}=(Max(S_{ij}, S_{ji}))^{P}, \ (i>j). \tag{3}$$

3.3 Cluster analysis for relational graph model

Though combinatorial clustering methods are versatile, when it applies to the asymilarity matrix, the matrix must be symmetrized. Consequently the initial information space to be clustered is distorted, so that the precision of analysis often decreases. We propose a new method of clustering for asymmetric similarity matrix. It is called binary relationship cluster analysis that classifies objects by the binary relationship between articles.

Binary Relationship Cluster Analysis (BRCA)

Let $od\ (i)$ denote the outdegree at an anticle i, which is obtained by $\sum_{j=1}^{t} a_{ij}$, where t is the total number of articles, and let y_{ij} be

$$y_{ij}=\frac{r_{ij}}{(od\,(j))^{e}}, \tag{4}$$

where e is an enhancing exponent ordinarily $e=1$.

If the combinatorial analysis of total relation is required, let z_{ij} be

$$z_{ij}=r_{ij}, \tag{5}$$

otherwise let a matrix $Z=[z_{ij}]$,

$$\text{where } z_{ij}=\sum_{k=1}^{t} w_k {}_k y_{ij}, \tag{6}$$

and w_k is a weight (e. g. $w_k=1, \frac{1}{k}, \frac{1}{k^2}$), and ${}_k y_{ij}$ is (i, j) element which is a powered matrix $[y_{ij}]$ of k. The quantity z_{ij} is construed as the quantity in which article i has influence on article j. Comparing z_{ij} with z_{ji} pairwisely, when $z_{ij} \geqq z_{ki}$, article j is linked under article k $(i \neq k)$.

This clustering procedure is realized by the pollowing way.

Procedure 1 : Searching $max(z_{ij})$ row-wisely $(j=1, 2, \ldots, m)$ and denote it z_{ik}.

Procedure 2 : *If* z_{ik} is greater than or equal to z_{ki}, article i is clustered to cluster k, otherwise article k is clustered to cluster i. However, if there exists z_{lm} which is greater than $z_{ik}\,(m \neq k)$, then article i is joined to cluster m.

The property of this clustering method is to construct a hierarchical structure and has a tendency to make small cluster.

4. Illustrations

Scientific articles for two different research fields have been investigated in order to test the present method. One is 231 articles for CAD/CAM. The other is 140 articles of two or three bodies problem for nuclear physics. All of them have citation relation articles in each field. The former contains mainly the articles for computational geometry and several articles relevant to AI (Artificial Intelligence). In the following comparison with

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.