A clustering method using the strength of citation

Tatsuki Saito

Journal of Information Science 1990 16: 175
DOI: 10.1177/016555159001600305

A clustering method using the strength of citation

Tatsuki Saito
Department of Precision Engineering, Faculty of Engineering, Hokkaido University, North 13 West 8, Kita-ku, Sapporo, 060 Japan

Received 20 September 1989
Revised 11 December 1989

A new method for modelling and clustering a relational graph produced from the citation relation among scientific articles is discussed. One article corresponds to a point and a citation link corresponds to a directed walk in the graph. This graph is a direct-citation graph and a total-citation graph is derived from it. There exist two types of directed citation graph; i.e., citing directed-graph and cited directed-graph. The former is considered in this paper. These graphs are represented in the form of similarity matrices which are asymmetric. The characteristics of these graphs are analyzed by clustering. For this study, a research database was designed and produced to acquire bibliographic information and obtain relations between articles in it. A modelling methodology of its relational structure is described in this article. Combinatorial clustering methods have been examined and a clustering method for an asymmetric similarity matrix is also proposed.

1 Introduction

Various kinds of important information exist in scientific articles. Relations between articles are useful for the study of the properties of research. While marketing bibliographic databases bring us bibliographic information comprehensively, they sometimes contain inaccurate information [1]. The research database named "ANGEL" used to process detailed information, especially on citation relations, has been produced for the reason that direct acquisition from original articles is necessary [2,3]. There have been some studies on how to develop a methodology enabling improved retrieval efficiency by using citation relations [2–15] and to evaluate journals or evaluate the dynamic influence of scientific activity [2–6,16–21]. The aim of this study is not only to do this, but also to develop a method to analyze the structure of relations between articles and clarify the properties of the research. A methodology for modelling the relational structure among scientific articles is treated and clustering methods, which include a present technique, are examined in this article. Ordinarily, the graph used for modelling is a non-directed graph by reason of its ease of processing [22,23]. However, scientific article information has directionality in the form of the relation between a citing article and a cited article. There are two types of citing directed-graph and cited directed-graph. Though the former is described in this article, when a direct-citation relation matrix is generated, it is necessary to exchange "cite" for "cited by" if a cited relation graph is required. We discuss representation by a directed-valued graph which constructs a more precise model in order to deal with the strength of the relation. That is, this is a discussion about the validity of a method to represent the relations among scientific articles by asymmetric similarity matrices. Cluster analyses have been applied because the processing model is represented in the form of a similarity matrix in its final form. In cases where a processing object is expressed by a symmetric matrix, combinatorial cluster analyses are suitable [24–30]. However, in the case of an object expressed by an asymmetric matrix, it is necessary to transform the matrix to a symmetric matrix. Then the problem of distortion of the original structure appears. This fact has been confirmed by our experiment. A new cluster analysis method which makes it possible to employ an asymmetric similarity matrix is proposed in order to avoid such distortion.

2 Representation of citation relation

2.1 Representation of relational strength by a valued graph

The graph is made up of the non-empty finite set P that has a point, and the specified set L

Journal of Information Science 16 (1990) 175–181

that has unordered q pairs belonging to P. Pair $l = (i, j)$ of point i and point j belongs to L, and is called a line of the graph. A graph that has p points and q lines is called a (p, q) graph or graph G = (P; L). It can be used to model structural objects; that is, the graph is a relational graph in which a point corresponds to an objective article and a line corresponds to a relation between two articles. There are two types of graphs, non-directed and directed. A non-directed graph is thought to be a special directed graph that always accompanies a directed line of reverse direction. In a directed graph, the finite point-set P, which is not empty, and the specified set L, which has the ordered pairs of two different points, are dealt with simultaneously. In the author relation graph, articles by the same author are connected by lines. Though the author relation has no direction, the citation relation has direction. Therefore, the citation relation is represented by a directed graph. While a directed graph can express the presence or absence of a relation between articles, it cannot express the strength of the relation. Consequently, a valued graph is introduced to enable the expression of the strength of the relation. A valued graph (P; L; Y) is a graph in which a line of set L is accompanied by the value r, where r is mapped onto the real number set Y. To express $r(i, j)$, r is called a value of a line. The production of the valued graph and its relational similarity matrix are described in Section 2.2. While the value is 0 or 1 in a direct-citation relation graph, it assumes various values in a total-citation relation graph as shown in the next procedure.

2.2 *Modelling procedure*

The modelling is processed as follows. To begin with, considering the binary citation relation between article i and article j, it is represented as a relational graph which is expressed by point i, point j and line (i, j). Then, after generating a directed graph that corresponds to citation relations between articles, the graph is represented as a similarity matrix. The generation of a direct-citation relation graph and the production of its similarity matrix are described below.

Direct-citation relation matrix

The direct-citation relation matrix is defined as $A = [a_{ij}]$. This is a so-called adjacency matrix, where

$a_{ij} = 1$; when article i cites article j.

$= 0$; otherwise.

If a cited relation graph is needed, it is necessary to exchange "cites" for "is cited by."

Total-citation relation matrix

Considering the total-citation relation that involves indirect citation, the total-citation relation matrix is defined as $S = [s_{ij}]$, where

$$S_{ij} = \sum_{k=1}^{n} w_k \cdot {}_k a_{ij}.$$

Here, k implies the length of a directed walk, and ${}_k a_{ij}$ means the number of directed walks whose length between article i and article j is k. The upper boundary n does not exceed $max(k)$ (the maximum length of the directed walk), and w_k is a weight that takes a value such as $1/k$ or $1/k^2$. The ${}_k a_{ij}$ can be obtained by the kth power of matrix A by resetting the diagonal elements to 0. An efficient algorithm that calculates only the non-zero elements of matrix A has been developed.

3 Cluster analysis of the similarity matrix

3.1 *Characteristics of the citation relation matrix*

Methods of clustering that analyze the relational graph represented by the similarity matrix are discussed here. Hierarchical techniques of the combinatorial cluster analysis are discussed in Section 3.2 and the present method is described in Section 3.3. The matrix to be analyzed is given by

$$R = [r_{ij}], \quad r_{ij} = (s_{ij})^p,$$

where p is an enhancing exponent.

This matrix has the following characteristics. First, each element of the matrix has a positive real quantity. Element s_{ij} is similarity as a correlation-like measure and element r_{ij} has a positive real quantity. Second, the matrix is asymmetric. The similarity matrix of the citation relation is asymmetric because there is a time sequence in the citation relation between a citing article and a

cited article. In consequence, matrix R becomes asymmetric. When the methods of the combinatorial cluster analyses discussed in Section 3.2 are applied, it is necessary be change it to a symmetric matrix. However, this is not desirable due to the occurrence of the distortion of the original matrix. We propose a new clustering method for the asymmetric matrix in Section 3.3.

3.2 Application of combinatorial clustering methods

Combinatorial cluster analysis is also called the method of hierarchical cluster analysis for the reason that it constructs a tree. All computer programs using the following combinatorial methods have been implemented and applied for the analysis of the relational graph model. It should be noted that the original similarity matrix of the relational graph model is asymmetric but that it must be changed to a symmetric matrix by the method described below because the combinatorial methods are applicable only to symmetric matrices. Seven hierarchical clustering methods have been examined. The nearest neighbour method is known as single linkage because clusters are joined at each stage by the single shortest or strongest link between individuals. The furthest neighbour method is also called the complete-linkage method because all individuals in a cluster are linked to each other by some max–min similarity. The median method adopts the middle value of the nearest neighbour value and the furthest neighbour value. The method of average linkage within the new group is not influenced by extreme values for defining clusters so it cannot make any statements about the minimum or maximum similarity within a cluster. Average linkage between a merged group, also 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 difference between the latter and the former is whether the sums of within-group similarities are ignored or not. The centroid method uses both the mean value of similarities and the number of individuals for the merger. The minimum variance method (the Ward method), used to find at each stage those two clusters whose merger gives the minimum increase in the total within-group error-sum of squares, is generally

When these combinatorial methods were applied, the asymmetric similarity matrix R was changed to a symmetric matrix by $r_{ij} = (max(s_{ij}, s_{ji}))^p$, $(i > j)$.

3.3 Cluster analysis method for the asymmetric matrix

Though combinatorial clustering methods are versatile, the matrix must be symmetrized when applied to an asymmetric similarity matrix. 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 the asymmetric similarity matrix. It is called "total-relationship cluster analysis" and classifies objects by the total relationship between articles. To use this method, let $od(i)$ denote the outdegree at article i, which is obtained by $\sum_{j=1}^{t} a_{ij}$, where t is the total number of articles.

If combinatorial analysis of the total-citation relation is required, let z_{ij} be as $z_{ij} = r_{ij}$, otherwise, let a matrix which is represented by

$$Z = [z_{ij}], \quad z_{ij} = \sum_{k=1}^{t} w_k \cdot {}_k y_{ij}, \quad y_{ij} = \frac{r_{ij}}{(od(i))^e},$$

be introduced, where w_k is a weight (for example, $w_k = 1$, $1/k$ or $1/k^2$). The quantity ${}_k y_{ij}$ is an (i, j) element which is a powered matrix $[y_{ij}]$ k times and e is an enhancing exponent. The quantity z_{ij} is construed as the influence of article i on article j.

Total-relationship cluster analysis (TRCA)

Let value $z_{\cdot j}$ designate the total relationship between article j and all other articles. It is introduced by

$$z_{\cdot j} = \sum_{i=1}^{t} z_{ij}.$$

To find the best u seeds through $z_{\cdot j}$ if the number of goal clusters is u, let them be denoted as $j_s (j_s = j_1, j_2, \ldots, j_u)$, where u is the number of goal clusters. Article i is clustered into article j_0, where $z_{ij_0} \geqq z_{ij} (j_0 \neq j)$ and $j, j_0 \in j_s$). The procedure of total-relationship cluster analysis is as follows.

Procedure 1: Find the best u seeds through $z_{\cdot j}$,

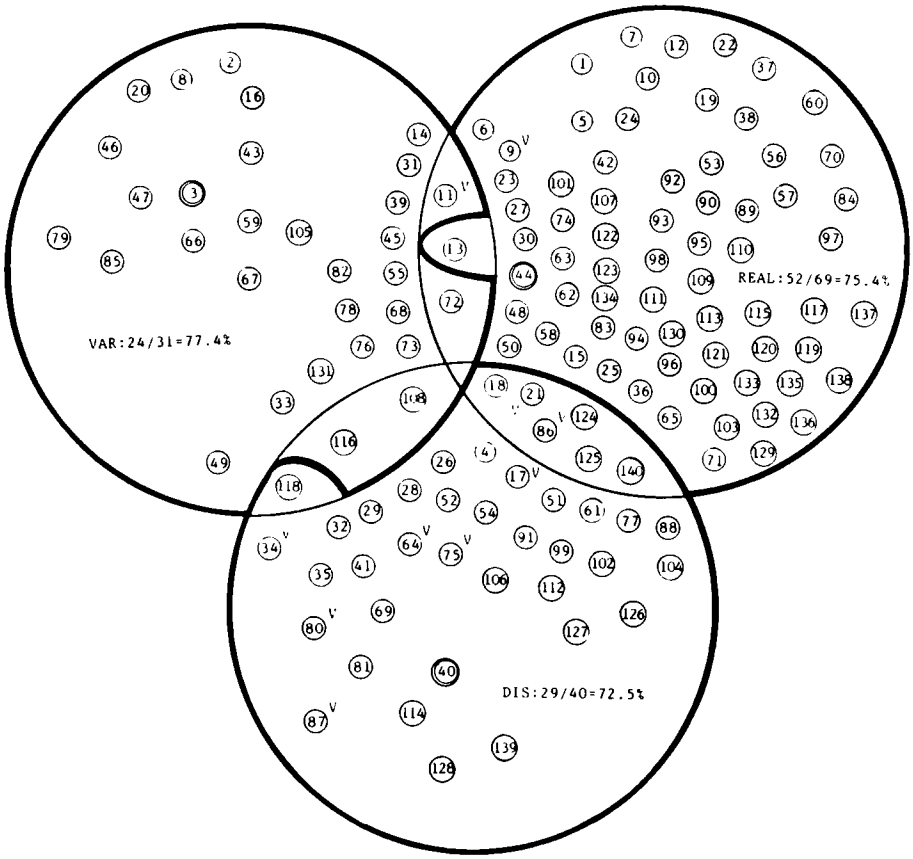
Procedure 2: Article i is clustered into article j, where z_{ij} has the maximum value on the same i-row in Z ($j = j_0$, $j_0 \in j_s$).

One of the important properties of total-relationship cluster analysis is that the number of goal clusters can be controlled and no hierarchical structure arises between articles, except at the core level.

4 Exploratory result

In order to test the present method, scientific articles for two different research fields were investigated. One group included 231 articles concerning CAD/CAM and the other 140 articles addressed the two or three bodies problem in nuclear physics. All of them had a citation relation between articles in each field. The former contained mainly articles about computational geometry and several articles relevant to artificial intelligence (AI). In the following comparison with manual classification by experts, the conformance percentage is the separation rate between the AI-cluster and the proper CAD/CAM cluster. Because its research field is obviously different from the research field of CAD/CAM it was hoped that this would suppress the occurrence of error due to the analyst's subjectivity. On the other hand, the latter contained three groups which use a different approach or methodology. That is, it is the first group of articles using the direct solution (DIS), the second group of articles using the variation solution (VAR) and the third group of articles focusing on the realistic interaction (REAL). The clustering methods used were the seven combinatorial methods and the present method. Computer programs by Anderberg were utilized for combinatorial clustering [31].

The average linkage within the new group method was the best among hierarchical methods. Other combinatorial methods are not useful be-



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.